Master equation models of macromolecular dynamics from atomistic simulation

by

John D. Chodera

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

BIOPHYSICS

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Approved by Committee in Charge:

_____

Chair Date

_____

Date

_____

Date

Deposited in the Library, University of California, San Francisco

_____

Date University Librarian

Degree Conferred:

Master equation models of macromolecular dynamics from atomistic simulation

Dedicated to the memory of Peter A. Kollman,

whose unbounded enthusiasm for the field of biomolecular simulation was an inspiration.

Acknowledgments

No scientist works in a vacuum. The scientific output of each individual is shaped to no small degree by those who have touched their lives deeply. This dissertation is no exception. I am indebted to so many people for their contributions to this work through direct and indirect means that it is not possible to name them all here, but I will try to touch on a few whose omission would certainly be a great injustice.

I thank the following people, not in order of importance, but rather mostly in order of appearance: My mother, who taught me to read; my father, who first showed me that it is possible to understand how the world around us works; my constant childhood friend and mentor Delfo Tomeoni, who instilled within me the joy of finding things out; Jr-Gang Cheng, who somehow recognized the seeds of an actual scientist within me while I was still a green-around-the-ears undergraduate, and taught me how to hold a pipetman; Jerry Solomon and David Liney, who introduced me to protein folding, the problem that would consume much of my waking thoughts for the next ten years, and to the computational tools that could be used to understand it; Peggy Gabriel, who made all those undergraduate years a lot more fun; Peter Kollman, whose booming voice echoing through the hallways as he excitedly talked about free energy calculations I can still recall today; Lillian Chong, my first rotation advisor, for "showing me the ropes" with AMBER and molecular dynamics simulation, whom I was delighted to have the chance to work with again at IBM; Ken Dill, whose gentle guidance has made my graduate career a wonderful experience, and whose clarity of insight helped me develop a much deeper understanding of statistical mechanics; the Dill lab crew, especially Banu Ozkan, whose work inspired my course in this thesis, Larry Schweitzer for keeping everything running so smoothly, and my classmates, Justin Bradford, Vincent Voelz, and Byoung-Chul Lee; Bill Swope, whose scientific enthusiasm, attention to detail, and thoroughness I have come to greatly appreciate, and whom, along with Jed Pitera, I had the privilege to work with on body of work this dissertation encompasses; Libusha Kelly for being my constant scientific and ethical sounding board, someone I can always count on to put me in my place when I screw up, and without whose constant encouragement this dissertation would have never been finished; Nina Singhal and Vijay Pande, whom it has been my pleasure to work with throughout much of this dissertation; Matt Jacobson, for always keeping his office door open, which let to numerous enjoyable and enlightening scientific discussions; the Brass at UCSF kids, especially Alvin Mok, Brian Fife, Ann Wehman, Nick Endres, Terry Lang, and Ian Harwood, for ensuring there was at least *one* night a week I wasn't working; the Biophysics/BMI/CCB kids, especially Terry Lang,

Michael Kim, and Michael Reese, for making science a whole lot of fun; and finally, but perhaps most importantly, Julie Ransom, the benevolent shepherd of the Biophysics program, without whose tireless efforts I would surely be destitute, lying in a ditch somewhere in Tijuana, clutching a cheap bottle of bourbon.

. . .

My thesis committee, consisting of Ken Dill (chair), Matt Jacobson, and Vijay Pande, deserves special thanks for reviewing this dissertation. The text of Chapter 2 largely contains material to appear in *Journal of Chemical Theory and Computation*. Chapter 3 contains material to appear in *Multiscale Modeling and Simulation*. Chapter 4 contains material to be submitted to the *Journal of Physical Chemistry B*. In these three chapters, co-authors made the following contributions: Bill Swope aided in the development of the theory and the interpretation of data and helped write the manuscript, and Jed Pitera participated in discussions and provided code upon which the parallel tempering simulation code is based. Chaok Seok also contributed to development of some of the early theoretical development for Chapter 2, and implemented some early test code. Ken Dill supervised the research that produced these chapters. Chapter 5 contains material to be submitted to the *Journal of Chemical Physics*. There, Nina Singhal and I cowrote the manuscript with the assistance of Bill Swope, and Nina was responsible for the majority of the application of the algorithm to the various test systems and interpretation of the results. Both Ken Dill and Vijay Pande supervised the research that produced this chapter.

Abstract

Master equation models of macromolecular dynamics from atomistic simulation

by

John D. Chodera

Doctor of Philosophy in Biophysics

University of California, San Francisco

Professor Ken A. Dill, Chair

Abstract goes here.

_____

Professor Ken A. Dill
Dissertation Committee Chair

Contents

List of Figures **x**

List of Tables **xiv**

1 Introduction **1**

2 The weighted histogram analysis method **7**
2.1 Introduction 8
2.2 Independent Canonical Simulations 10
2.2.1 Motivation and Definitions. 11
2.2.2 Obtaining an Estimate of the Density of States from Each Simulation. . . . 12
2.2.3 Optimal Estimator from Independent Observations and Associated Uncertainties. 14
2.2.4 Statistical Uncertainty in the Estimator for Correlated Time Series Data. . . 15
2.2.5 Optimal Estimate of the Density of States. 16
2.2.6 Estimating an Observable at the Temperature of Interest. 19
2.2.7 Statistical Uncertainty of the Estimator for the Expectation. 21
2.3 Simulated and Parallel Tempering 23
2.3.1 Simulated Tempering. 23
2.3.2 Parallel Tempering or Independent Simulated Tempering Simulations. . . . 26
2.4 Applications 28
2.4.1 One-Dimensional Model Potential. 28
2.4.2 Alanine Dipeptide in Implicit and Explicit Solvent. 30
2.5 Practical Considerations 33
2.5.1 Choice of Bin Width and Number of Bins. 34
2.5.2 Computing Integrated Correlation Times. 34
2.5.3 Neglect of Bin Statistical Inefficiencies g_{kn}. 35
2.5.4 The Statistical Inefficiency for the Cross-Correlation Term, $g_{k,wA;w}$. 37
2.6 Conclusion 37
2.7 Acknowledgments 38
2.8 Relation for Statistical Inefficiencies 38
2.9 Uncertainty Estimates for Confidence Curves 39

3 A master equation can describe peptide dynamics **43**
3.1 Introduction . . . 44
3.2 Theory . . . 46
3.2.1 Conformational dynamics as a Markov process. . . . 46
3.2.2 Construction of the Markov chain model from simulation. . . . 47
3.3 Application to terminally-blocked alanine peptide . . . 49
3.3.1 System setup and equilibration. . . . 49
3.3.2 Parallel tempering. . . . 50
3.3.3 State decomposition. . . . 51
3.3.4 Construction of Markov chain model from short trajectories. . . . 51
3.3.5 Comparison with long trajectories. . . . 52
3.3.6 Long-time dynamics from the Markov chain model. . . . 55
3.4 Discussion . . . 55
3.5 Acknowledgements . . . 57

4 Validating master equation models of macromolecular dynamics **58**
4.1 Introduction . . . 59
4.2 Master equation and Markov models . . . 62
4.2.1 The discrete-state master equation. . . . 62
4.2.2 Construction from molecular dynamics simulation. . . . 64
4.3 Terminally-blocked alanine peptide in explicit solvent as a model system . . . 66
4.3.1 System setup and equilibration. . . . 66
4.3.2 Parallel tempering simulation. . . . 66
4.3.3 State decomposition. . . . 67
4.3.4 Shooting simulations. . . . 68
4.3.5 Uncertainty estimation. . . . 70
4.3.6 Estimation of transition probabilities. . . . 71
4.4 Tests for Markovianity. . . . 72
4.4.1 The implied rate matrix. . . . 73
4.4.2 Eigenvalues of the implied rate matrix. . . . 74
4.4.3 Second-order transition probabilities. . . . 77
4.4.4 Chapman-Kolmogorov equation. . . . 78
4.4.5 Discrete lifetime distribution. . . . 80
4.5 Discussion and Conclusions . . . 82
4.6 Acknowledgements . . . 83

5 An automatic state decomposition algorithm **84**
5.1 Introduction . . . 85
5.2 Theory . . . 88
5.2.1 Markov chain and master equation models of conformational dynamics. . . . 88
5.2.2 Construction from simulation data given a state partitioning. . . . 91
5.2.3 Requirements for a useful Markov model. . . . 92
5.2.4 Validation of Markov models. . . . 93
5.3 The automatic state decomposition algorithm . . . 95
5.3.1 Practical considerations for an automatic state decomposition algorithm. . . . 95

5.3.2 Sketch of the method. 96
5.3.3 Implementation. 98
5.3.4 Validation. 101
5.4 Applications 101
5.4.1 Alanine dipeptide. 101
5.4.2 The F_s helical peptide. 104
5.4.3 The trpzip2 β-peptide. 109
5.5 Discussion 110
5.6 Supporting Information 113
5.7 Acknowledgments 114

6 Conclusion **122**

Bibliography **124**

List of Figures

2.1 **Confidence curves for Metropolis Monte Carlo simulations on the 1D model potential.** The fraction of statistically independent blocks for which the true uncertainty (the deviation of the estimated expectation over the block from the mean of the block estimates) is less than a multiplier of the predicted 1σ uncertainty (here plotted as the independent variable). The solid curve shows the fraction expected to fall within the interval for the normal distribution. Ideally, the curves would coincide. The results are shown for (MMC) a single Metropolis Monte Carlo simulation at $\beta = 4$; (4MMC) a set of four independent canonical simulations spanning the range β = 1 – 4; (ST) a simulated tempering simulation spanning β = 1 – 4; (PT) a parallel tempering simulation with four replicas spanning β = 1 – 4. Uncertainties, with 95% confidence intervals shown here as vertical bars, were computed as described in Appendix 2.9. 30

2.2 **Terminally-blocked alanine peptide with (ϕ, ψ) torsions labeled.** 31

2.3 **Potential of mean force in ψ for implicit and explicit solvent parallel tempering simulations.** Left: implicit solvent; right: explicit solvent. Upper panels: The potential of mean force in the ψ torsion angle at 300 K. The solid line shows the PMF estimated from the entire simulation, while the filled circles show the estimated PMF uncertainty using the method described in the text for a single 2 ns/replica block. Lower panels: The computed uncertainties for the same 2 ns block (left bars) along with the average uncertainty expected for a simulation 2 ns/replica in length, estimated from the standard deviation of the PMFs computed from all nonoverlapping blocks of length 2 ns in the full simulation. All uncertainties are shown as one standard deviation. 41

2.4 **Confidence curves for implicit and explicit solvent parallel tempering simulations.** As in Figure 2.1, the fraction of statistically independent 2 ns blocks for which the true uncertainty is less than a multiplier of the predicted 1σ uncertainty is shown. The observable used is an indicator function for the α_{R} configuration. Left: implicit solvent (statistics over 50 blocks); right: explicit solvent (statistics over 10 blocks). 42

3.1 **Potential of mean force and state boundaries.** Left: The terminally-blocked alanine peptide with (ϕ, ψ) torsions labeled. Right: The potential of mean force in the (ϕ, ψ) torsions at 302 K estimated from the parallel tempering simulation, truncated at $10k_BT$ (white regions), with reference scale (far right) labeled in units of k_BT. Boundaries defining the six manually-identified states are superimposed and the states labeled. 49

3.2 **Transition matrix elements as a function of lag time estimated from 10 ps shooting trajectories.** Each plot, labeled above by the state from which the trajectories originated, shows state-to-state transition probabilities as a function of the lag time τ estimated from a set of 1000 trajectories 10 ps in length originating from an equilibrium distribution within each state. Vertical bars depict 95% confidence intervals. Equilibrium state probabilities obtained from the parallel tempering simulations are shown as solid horizontal lines in the corresponding color. 52

3.3 **Temporal evolution of state populations from Markov chains constructed at different lag times compared with long simulations.** Evolution of state probabilities from an ensemble prepared at equilibrium within each state for Markov model estimated from the set of 10 ps shooting trajectories (solid lines) superimposed on fractional population of each state as a function of time for ensemble of 100 ps trajectories initiated from each state (points). Vertical bars depict 95% confidence intervals in state populations estimated from the long trajectories. 54

3.4 **An artificial trajectory generated from the transition matrix constructed from a lag time of 10 ps.** 55

4.1 **Terminally-blocked alanine and potential of mean force at 302 K.** Top: The terminally-blocked alanine peptide with (ϕ, ψ) torsions labeled. Bottom: The potential of mean force in the (ϕ, ψ) torsions at 302 K estimated from the parallel tempering simulation, with "good" (left) and "poor" (right) state decompositions labeled, and colorbar graded in units of k_BT and truncated at $10k_BT$ (white). . . . 69

4.2 **Implied transition rates as a function of lag time.** For each state, the implied transition rates to all other states are shown. Top: Rate matrix elements implied by the observed transition matrix as a function of lag time for the "good" state decomposition. Bottom: Rate matrix elements implied by the observed transition matrix as a function of lag time for the "poor" state decomposition. 73

4.3 **Implied timescales of the rate matrix as a function of lag time.** Implied timescales and associated uncertainties are shown for good decomposition (top), poor decomposition (middle upper), good lumping (middle lower), and poor lumping (lower). Timescales are colored by order from longest (blue) to shortest (purple). The black line denotes $\tau_k = \tau$, such that processes whose timescales fall below this line occur on times shorter than the lag time. 75

4.4 **Second-order transition probabilities compared with products of first-order transition probabilities as a function of lag time.** Observed second-order transition probabilities $T_{k|ji}(\tau)$ are shown as solid lines with the envelope indicating a 68% confidence interval, while first-order transition probabilities are shown as points with error bars. Because there are $6^3 = 216$ second-order transition probabilities, and many of them are not well estimated by the limited dataset, only the 36 $T_{i|ji}(\tau)$ transition probabilities are shown. Each plot shows 6 transition probabilities originating from a single state i, occupying a state j at time τ, and returning to i at time 2τ. 77

4.5 **Test of the Chapman-Kolmorov equation.** Observed transition probabilities $T_{ji}(2\tau)$ are shown as points with error bars, while predicted transition probabilities from $([\boldsymbol{T}(\tau)]^2)_{ji}$ are shown as lines enveloping a 68% confidence interval. 79

4.6 **Observed and geometric discrete lifetime probability distributions.** The logarithm of the discrete lifetime probability mass function $P(L;\tau)$ for each state, which indicate the probability of observing the system to remain within a given state for exactly a a number of consecutive observation intervals L, is shown as a function of $L\tau$ along with the characteristic lifetimes obtained from a linear fit to the log probability over the interval $L\tau \in [1.5, 10]$ ps, evaluated with sampling interval τ_{sample} = 0.1 ps. The number of intervals L for which the system remains in one state has been multiplied by the sampling interval τ so that the x-axis appears in units of time. Top: good states; middle top: poor states; middle bottom: good lumping; bottom: poor lumping. 81

5.1 **Flowchart of the automatic state decomposition algorithm.** 115

5.2 **Potential of mean force and manual state decomposition for alanine dipeptide.** Left: The terminally-blocked alanine peptide with ϕ, ψ, and ω backbone torsions labeled. Right: The potential of mean force in the (ϕ, ψ) torsions at 400 K estimated from the parallel tempering simulation, truncated at 10 k_BT (white regions), with reference scale (far right) labeled in units of k_BT. Boundaries defining the six states manually identified in Ref. [25] from examining the 300 K PMF are superimposed, and the states labeled. 116

5.3 **Comparison of manual and automatic state decompositions for alanine dipeptide.** The left panels depict state partitionings, and the right panels the associated timescales (in picoseconds) as a function of lag time with uncertainties shown, as estimated from the procedure described in Section 5.3.4. Top two panels: Manual "good" or "gold standard" state decomposition from Ref. [25] and manual "poor" state decomposition, where the state boundaries are grossly distorted so as to include internal kinetic barriers within the states. Bottom two panels: Two nearly-equivalent partitionings obtained from the automatic state decomposition algorithm. 117

5.4 **Stability and recovery of optimal state decomposition for alanine dipeptide.** Top: Ten cycles of automatic state decomposition applied to a "good" manual partitioning (left) to yield an automatic partitioning (right). Bottom: Ten cycles of automatic state decomposition applied to a "poor" manual partitioning (left) to yield an automatic partitioning (right). 118

5.5 **Implied timescales of the F_s peptide as a function of lag time for 20-state automatic state decomposition.** The five longest timescales are shown. Circles represent the maximum likelihood estimate, and vertical bars depict 68% symmetric confidence intervals about the mean. Note the timescales associated with two processes appear to cross, but are here colored and uncertainties are estimated with bootstrapping by ordering them by rank. This may cause the uncertainties depicted here to be an underestimate of the true uncertainties of each process. 118

5.6 **Reproduction of observed state population evolution by Markov model for the F_s peptide.** The time evolution of the Markov model constructed from the 5 ns lag time transition matrix is shown by the filled circles with flat error bars, which denote the 68% confidence interval from realizations of a bootstrap sample of 40 transition matrices computed from a 5 ns lag time. Vertical bars without flat ends denote the 68% asymmetric confidence interval for the probability of finding the system in the 20 macrostates a given time after initial preparation in a specific state. The system was originally prepared in state 2 (top, red), 13 (middle, yellow), or 19 (bottom, purple). The most populous states are colored green (state 1), red (state 2), and blue (state 3). 119

5.7 **Comparison of some trpzip2 macrostates found by automatic state decomposition with misregistered hydrogen bonding states identified in a previous study.** Left: The five hydrogen bonding patterns enumerated in Pitera *et al.* [143] that occurred in sufficient numbers in the subsampled trpzip2 dataset used here, with representative conformational ensembles. Right: A selection of macrostates discovered by automatic state decomposition that the contain the largest numbers of hydrogen bonding pattern states. The backbone is depicted in alpha carbon trace, and tryptophan sidechains are shown in light blue (Trp2), orange (Trp4), magenta (Trp9), and teal (Trp11). A complete set of macrostates obtained from the 40-state decomposition of the trpzip2 dataset is available as Supplementary Information. . 120

5.8 **Implied timescales of trpzip2 as a function of lag time for 40-state automatic state decomposition.** The five longest timescales are show. Vertical bars depict 68% confidence intervals. 121

5.9 **Automatic state decomposition applied to trpzip2 to produce 40 macrostates.** . 121

List of Tables

3.1 **State definitions for the manual decomposition of (ϕ, ψ)-space into metastable states and populations at 302 K.** 50
3.2 **Transition matrices[a] at several lag times estimated from set of 10 ps trajectories.** 53

5.1 **Macrostates from a 20-state state decomposition of the $\mathbf{F}_s$ helical peptide.** The backbone is depicted in alpha carbon trace, and arginine sidechains are shown in blue (Arg10), magenta (Arg15), and green (Arg20) for clarity. 105

Chapter 1

Introduction

Perspective

Conformational dynamics plays an integral role in the function of biological macromolecules. Proteins, after translation by the ribosome, rapidly fold to well-defined native topologies that bring disparate chemical moieties together into a particular geometry, conferring the ability to perform chemical catalysis [170]. Errant misfolding can lead to aggregation and the formation of amyloid plaques, a phenomenon associated with diseases such as Alzheimer's, Parkinson's, and Creutzfeldt-Jacob ("Mad Cow") Diseases [43]. Once folded, excursions to partially unfolded states can expose proteins to proteolysis; to avoid this, some organisms appear to have evolved secreted proteases that are kinetically stable to maintain competitive advantage in harsh extracellular environments [89]. Conformational changes of folded proteins can be critical for protein [82] or substrate binding [188] and catalytic function [15, 50, 200]. Together with order-disorder transitions [46], ordered sequences of conformational changes are integral to the transformation of chemical energy into mechanical work by motor proteins [42, 91] or the reverse by ATP synthases [132]. Binding or post-translational modification events far from the active site can modulate the activity of proteins through allosteric effects, which involve poorly understood structural or dynamical changes transmitted through largely unknown mechanical or electrostatic pathways [21, 65, 114]. Slow prolyl *cis-trans* isomerization dynamics can be used as a molecular timer [48], which can be modulated by phosphorylation events for signaling purposes [198]. Ion channel gating, an inherently stochastic kinetic process involving conformational switching between multiple conductance states [30, 115], can be modulated by transmembrane voltage [14, 29], ligand binding at allosteric sites [33], temperature [181], mechanical pressure [142], and phosphorylation [104], and undergo time-dependent

inactivation [38]. The conformational plasticity of RNA no doubt contributes to its many characterized biological roles in information transmission [133, 149], binding [76], and catalysis [93, 203]. In each of these cases, conformational dynamics plays an integral role in macromolecular folding, assembly, function, and regulation, and a detailed description of this dynamical behavior is likely critical to achieving an understanding of the corresponding biological phenomena.

Structural biology has proven to be a valuable tool for the study of these machines. Structural studies, especially X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, allow the determination of experimentally-derived structural models which can aid the formulation of hypotheses about function, mechanism, and disease. However, these methods have generally been limited to producing *static* pictures of macromolecules. While attempts have been made to relate crystallographic Debye-Waller factors ("B-factors") to biologically relevant conformational fluctuations [6, 96], this interpretation is complicated by the fact that the majority of these crystals are cryogenically cooled before the collection of scattering data to temperatures well below the glass transition temperature [195] over times sufficiently long to allow for significant structural rearrangement [97]. Additionally, the presence of crystallographic and non-crystallographic neighbors, salts, and cosolvents, and binding partners in a crystal lattice suggests the relevance of information on dynamics or heterogeneity obtained from these data should be treated as suspect unless proven otherwise by extensive comparison with experimental data under more biological conditions. In principle, NMR provides information on both conformational heterogeneity and dynamics in solution, but the difficulty of interpreting this data, coupled with the small number of experimental observations per residue, has made extraction of anything but average structures (a surprisingly robust problem [176]) difficult, though work continues on improving this situation [107, 146]. Standard NMR refinement protocols [130] produce ensembles of structures, but the resulting ensemble may not represent conformational heterogeneity; in fact, assumptions that the data comes from a single conformation ensure this cannot be the case. As a result, the ensemble is composed of *virtual* structures that may not exist in solution with high probability [88]. These structures simply represent minima of the target function, a sum of a least-squares error function with experiment and a molecular mechanics forcefield. Additionally, the conformational diversity observed in these ensembles cannot easily be determined to be due to conformational averaging in solution or a lack of sufficient experimental restraints, but is usually an underestimate of the true diversity that is possible for an ensemble that satisfies the NMR restraints on average [166]. Despite these problems, these methods have provided structural models which have been tremendously useful for the generation of models and experimentally testable hypotheses, rapidly accelerating our ability to understand

biological function and mechanism.

In order to more directly probe conformational heterogeneity and dynamics, a number of other biophysical techniques have been developed. Certain NMR experiments can provide information on picosecond and nanosecond timescales [86], but the amount of information that can be extracted and its interpretation has proven even more difficult that the extraction of information about conformational heterogeneity. Förster resonance energy transfer (FRET), in which two fluorescent probe molecules with overlapping emission and absorbance spectra are covalently attached to different groups of the molecule, allows for determination of the interprobe distance from the observed fluorescence emission. Other spectroscopic assays that do not require covalent modification, such as UV circular dichroism (CD), Fourier transform infrared spectroscopy (FTIR), and tryptophan fluorescence, also provide sensitive probes of different aspects of molecular structure and environment, though the interpretation of these spectra is often difficult and the information that can be extracted limited. Recent advances, such as two-dimensional infrared spectroscopy (2DIR) [94] can provide more information at the expense of sacrificed time resolution, allowing some individual chemical moieties to be resolved.

These spectroscopic probes can be employed to study either equilibrium thermodynamics over a range of conditions (such as temperature or denaturant concentration) or kinetics, in which relaxation from nonequilibrium initial conditions or equilibrium fluctuations are observed. In *ensemble* experiments, in which a solution of many macromolecules is monitored at equilibrium or after a rapid perturbation (such as rapid heating of the solvent with a short laser pulse [75]), high time resolution is possible because signal is collected from many molecules. However, due to the presence of large numbers of molecules, these experiments can only provide information about the average spectroscopic signal over the ensemble — information on heterogeneity within the ensemble is lost. As a result, much effort has recently been focused on the development of *single molecule* experiments, which can provide information about the heterogeneity of both equilibrium distributions and individual microscopic trajectories. To obtain equilibrium distributions, it is sufficient to work with solutions that are sufficiently dilute such that it is unlikely that more than one molecule is in the region under spectroscopic surveillance at any one time. To observe dynamic trajectories, however, these molecules must be prevented from diffusing away from the observation area. This is typically done through the use of covalently attached or noncovalently-bound molecular linkers tethered to the glass slide, or by encapsulation in immobilized vesicles [145]. However, to gather sufficient numbers of photons to give a reasonable signal-to-noise ratio, time resolution must be sacrificed, leading to experimental time resolution of milliseconds for single-molecule experiments,

rather than the nanosecond resolution achievable by ensemble experiments.

Because of these limitations, there is still an unmet desire to observe the dynamics of individual molecules with high time resolution and in atomic detail. Clearly, further advances in engineering and ongoing development of new methods will continue to push the boundaries of what is experimentally observable. However, fundamental limitations, such as the tradeoff between time resolution and information about heterogeneity in ensemble and single-molecule experiments above, and the ability to observe only spectroscopically active changes mean there is only so much information that can be expected from experimental methods.

Additionally, there appears to be a lack of consensus regarding the fundamental physical nature of the experimental observations and how to interpret them, or even how to summarize them in terms of sufficient statistics. The temporal signal from ensemble kinetics experiments, for example, has been variously fit to exponentials [180], sums of exponentials [126, 168, 179, 199], and so-called *stretched exponentials* [112]. The presence of a *burst phase* means that there is immediate and unexplained loss of spectroscopic signal in the *dead time* of the experiment, immediately after the perturbation (e.g. stopped flow mixing or laser temperature-jump [75]). A statistical mechanical framework which would permit explanation of all of these observed phenomena and at least provide a physical functional form of the resulting observations, and ideally a connection with the actual microscopic dynamics, would be beneficial.

With the advent of the modern microcomputer, a new kind of experiment became possible, in which the detailed atomic motions of the macromolecule and its environment were simulated given a suitable model for the interatomic forces. These molecular dynamics simulations promised the ability to model molecular processes in atomic detail and high time resolution, providing the microscopic detail missing from single-molecule or ensemble experiments mentioned above. However, gathering insight into biological processes from these trajectories faces several challenges. On contemporary workstations, molecular dynamics simulations with explicit representations of the solvent environment can reach simulation times generally limited to tens of nanoseconds — far shorter than even the fastest characterized folding protein times of microseconds. While current-generation supercomputers, with several moths of massively parallel computation, can reach timescales of up to 10 μs for small proteins [73], a *single* long trajectory in which only one event of interest occurs (e.g. a protein folding event) does not give much information about an inherently stochastic and heterogeneous process. By contrast, distributed computing projects [139, 158] offer the ability to produce many thousands of short trajectories, but there exists the danger that the mechanism by which the event of interest occurs (e.g. protein folding) may be biased, in that the

mechanism at short times may differ from the mechanism at long times, where the bulk of events might occur [59, 116].

The forcefields commonly employed in molecular dynamics simulations are also known to be vast simplifications of the actual physical interactions, neglecting some contributions, such as polarizability, completely. Comparison with experiment to assess the severity of these omissions, however, is frustrated by the fact that ensemble experiments provide information about averages over many macromolecules while simulations typically consider only single molecules. On the other hand, comparison with single molecule experiments is difficult due to the low time resolution of the experimental data and the difficulty of generating sufficiently long simulation trajectories.

As a result, the accuracy with which current-generation forcefields can model biomolecular kinetic processes is still largely unknown. What is needed is some way to bridge the *timescale gap* between short atomistic simulations of single molecules and long experimental observations of ensembles of molecules. Ideally, this could be done through the construction of a statistical model that contains information about the heterogeneity by which the dynamical processes of interest may occur. This model would have to be constructed from relatively short simulations of single molecules, and yet describe the stochastic dynamics of either a single molecule or a (noninteracting) ensemble of molecules over much longer times.

Fortunately, there is good evidence that the fundamental physical nature of intramolecular interactions makes it is possible to construct simple stochastic models of macromolecular dynamics. Pioneering work by Christof Schütte, Huisinga, and coworkers at the Zuse Institute of Berlin [41, 60, 60, 66, 67, 83, 152, 154] (and later Weber and Kube [98, 194]), as well as independent work by Shalloway and coworkers [27, 156, 186, 187], and Berry and coworkers [8, 39, 40, 106], proposed that macromolecules might exhibit behavior suggestive of long-lived *metastable conformational states*. The dynamics of a system with strongly metastable states is characterized by long waiting times *within* these states, punctuated by infrequent stochastic transitions *between* states.

The existence of metastable states is a simple consequence of the presence of a separation of timescales between *fast intrastate motion* and *slow interstate motion*. It is widely believed that the nature of the energy landscape of biomacromolecules is hierarchical [4, 7, 9, 105, 106]. Indeed, proteins are known to exhibit a wide dynamic range of timescales, from femtosecond bond vibration to nanosecond helix formation to microsecond or greater folding times. The hierarchical nature of the energy landscape presents an intriguing possibility: If there are many gaps in the spectrum of timescales (as would be expected from a hierarchical landscape), rather than a continuum (which would have to have a continuous and relatively flat distribution of barrier heights), then it should be

possible to construct *many* models with different numbers of metastable states capable of describing conformational dynamics, each with a different spatial and temporal resolution. These models need only be as detailed as necessary for describing the phenomena of interest, simplifying the process of interpreting experiments, understanding dynamics, and extracting chemical insight.

The resulting stochastic model, produced by coarse-graining conformation space into metastable states, is a discrete-state, continuous-time *master equation model*, in which transitions between states are described by first-order kinetics governed by a *rate matrix*. As will be discussed at length, the simplicity of the model comes at the cost of incurring a coarse-graining in time; temporal resolution is lost because conformational dynamics occurring on timescales comparable with motion *within* states is omitted in the model. Despite this, the master equation model possesses numerous benefits. The entire statistical dynamics over times longer than some intrinsic *internal equilibration time* is available, allowing the production of single-molecule trajectories or ensemble evolution experiments, as well as allowing direct comparison with nonequilibrium relaxation kinetics experiments, the computation of unobservable properties like P_{fold} [44] that aid in the understanding of mechanism [103], and the summarization of primary events in kinetics processes such as folding, most notably seen in the work of Banu Ozkan and Ken Dill [136, 137].

Synopsis

This thesis is organized as follows. Chapter 2 contains a variant of the weighted histogram analysis method that can treat simulated and parallel tempering simulations, and estimate the statistical uncertainty in equilibrium averages computed from these simulations. This was used extensively as a tool in subsequent work. Chapter 3 contains a manuscript which is to appear in the journal *Multiscale Modeling & Simulation* that introduces the Markov chain or master equation model and illustrates that a model constructed from short trajectories can describe the long-time statistical dynamics of terminally-blocked alanine in explicit solvent. Chapter 4 contains a manuscript to be submitted to the *Journal of Physical Chemistry B* that is concerned with how these models, once constructed short trajectories, can be validated to determine the timescale for emergence of Markovian behavior. Finally, Chapter 5 describes progress toward algorithms for the discovery of metastable states without prior knowledge of the relevant degrees of freedom, the final piece of the puzzle necessary for the construction of these models for biomolecules.

Chapter 2

The weighted histogram analysis method

The material in this chapter was submitted to the Journal of Chemical Theory and Computation, and has been accepted to appear as:

Use of the Weighted Histogram Analysis Method for the Analysis of Simulated and Parallel Tempering Simulations

John D. Chodera[1] [†], **William C. Swope**[‡], **Jed W. Pitera**[‡], **Chaok Seok**[§], **and Ken A. Dill**[¶]

[†] *Graduate Group in Biophysics and* [¶] *Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, CA 94143*

[‡] *IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120*

[§] *Department of Chemistry, College of Natural Sciences, Seoul National University, Gwanak-gu, Shillim-dong, san 56-1 Seoul 151-747, Republic of Korea*

Abstract

The growing adoption of generalized-ensemble algorithms [134] for biomolecular simulation has resulted in a resurgence in the use of the weighted histogram analysis method (WHAM) [100] to make use of all data generated by these simulations. Unfortunately, the original presentation of WHAM by Kumar *et al.* [100] is not directly applicable to data generated by these methods. WHAM was originally formulated to combine data from independent samplings of the *canonical* ensemble, whereas many generalized-ensemble algorithms sample from mixtures of canonical ensembles at different temperatures. Sorting configurations generated from a parallel tempering simulation by temperature obscures the temporal correlation in the data and results in an improper

treatment of the statistical uncertainties used in constructing the estimate of the density of states. Here we present variants of WHAM, derived with the same set of assumptions, that can be directly applied to several generalized ensemble algorithms, including simulated tempering [120], parallel tempering (better known as replica-exchange among temperatures) [171], and replica-exchange simulated tempering [120]. We present methods that explicitly capture the considerable temporal correlation in sequentially generated configurations using autocorrelation analysis. This allows estimation of the statistical uncertainty in WHAM estimates of expectations for the canonical ensemble. We test the method with a one-dimensional model system, and then apply it to the estimation of potentials of mean force from parallel tempering simulations of the alanine dipeptide in both implicit and explicit solvent.

2.1 Introduction

The difficulty of computing equilibrium averages for complex systems such as solvated biopolymers by Monte Carlo or molecular dynamics simulation is well-known. Numerous minima and large free-energy barriers tend to slow exploration in phase space and trap the simulation in metastable regions of configuration space. This hampers the ability of the system both to equilibrate (reach the thermodynamically relevant region of phase space) and to sample sufficiently for estimates of ensemble averages to converge (reduce the statistical uncertainty in the estimate to an acceptable level) in finite computer time.

The emergence of a new class of simulation algorithms, termed *generalized-ensemble* algorithms [134], has helped to mitigate these problems. In a generalized-ensemble simulation, the probability distribution from which conformations are sampled is altered from a canonical distribution to one that will induce a broader sampling of the potential energy. Proper application should in principle allow the system to overcome energetic barriers and sample configuration space more thoroughly, at the expense of spending more time in high-energy regions that may be irrelevant at the temperature of interest. The particular method by which sampling is enhanced depends on the algorithm. In the *multicanonical algorithm* (MUCA) [12, 13, 55, 78, 128], conformations are sampled with a probability proportional to an approximation of the inverse potential energy density of states in an attempt to produce a random walk in the potential energy. In *simulated tempering* (ST) [110, 117, 120], a random walk between canonical ensembles at different temperatures is used to produce a random walk in energy, but an estimate of the free energy as a function of temperature is needed as input to ensure equal visitation of all

temperatures. *Parallel tempering* (PT), a special case of the replica-exchange method (REM) [77, 171], eliminates the need to know these free energies *a priori* by coupling temperature changes between pairs of a pool of simulated tempering simulations conducted in parallel. Several other algorithms and combinations thereof have also been proposed [120–122, 172].

In several of these algorithms, such as simulated tempering and parallel tempering, each replica generates configurations from a *mixed-canonical* distribution (a term coined in [61]) — that is, a number of configurations are generated from the canonical distribution at each of several temperatures. To compute expectations over the canonical ensemble at a single temperature, either the configurations from all replicas that visit the temperature of interest must be collected and the remainder discarded (as in [150]) or else a reweighting scheme must be used to properly weight the data generated at other temperatures. Fortunately, the weighted histogram analysis method (WHAM) [100], an extension of the single- and multiple-histogram methods introduced by Ferrenberg and Swendsen [57, 58], allows configurations generated from independent canonical simulations at different temperatures to be reweighted to compute expectations from the canonical ensemble at any temperature of interest. Okamoto and coworkers have applied this method to both replica-exchange simulated tempering (REST) [120] and parallel tempering [171] methods by reordering sampled configurations into pseudotrajectories, grouping configurations generated at a particular temperature together regardless of which replica they came from. Unfortunately, this permutation obscures the correlation among the stored configurations, causing the apparent correlation times for each pseudotrajectory to appear artificially shorter than the true correlation times within the independent replica trajectories. The permutation also introduces correlation *between* the pseudotrajectories, which is problematic because WHAM as presented in [100] is constructed to operate on *independent* canonical trajectories. Additionally, it is difficult to estimate the statistical uncertainty in the resulting estimate of the expectation from these pseudotrajectories, since standard autocorrelation analysis techniques [64, 87, 127, 173] can no longer be applied.

Recently, Gallicchio *et al.* [68] have described a new method for computing expectations and uncertainties from canonical simulations at different temperatures based on Bayesian inference. While Bayesian approaches are usually superior to those based on first-order Taylor expansion methods for the propagation of uncertainties (of the sort we describe in this work), they are less suitable for treating highly correlated measurements where the functional form of the correlation is essentially unknown.

Here, we derive variants of WHAM that operate on replica trajectories that are not reordered or collected by temperature. It should be noted that even if simulation data has been stored to disk

sorted by temperature, it can be permuted back to the original replica trajectories to perform the proposed analyses if information about the replica-to-temperature mapping or swapping was stored. Our presentation takes a careful approach to the correlation times involved, and we show under which conditions the almost universally omitted statistical inefficiency term that appears in all formulations of WHAM-like methods can be properly neglected. Finally, we show how the statistical uncertainty in the estimator for the configuration space average for some observable can be estimated by considering the effect of temporal correlation. The method is simple and inexpensive enough to employ in all cases where WHAM is used, and we hope all researchers using WHAM will report these statistical uncertainties in future to assess both the significance and the degree of reproducibility of results from simulations.

This paper is organized as follows: In Section 2.2, we present a derivation of the Kumar *et al.* WHAM for independent simulations sampling from the canonical ensemble. Careful attention is paid to the proper treatment of time correlation in estimating the statistical uncertainty in the histograms and the resulting estimator for the expectation, and a novel way of obtaining estimates for multiple observables is presented. In Section 2.3, we derive an analogue of the method for treating simulated and parallel tempering simulations while properly capturing the correlations among sequential configurations. In Section 2.4, we validate our uncertainty estimates in a one-dimensional model system and demonstrate an application for biomolecular systems by estimating the potential of mean force and corresponding uncertainties from parallel tempering simulations of alanine dipeptide in implicit and explicit solvent. An illustrative efficient implementation of the method in Fortran 95 for use in the analysis of simulated and parallel tempering simulations can be found in the Supplementary Material, and a version that can be compiled and run can be downloaded online[2].

2.2 Independent Canonical Simulations

In this section, we review the derivation of WHAM for computing expectations from multiple independent simulations in the canonical ensemble. Conducting independent simulations at the same or different temperatures can reduce statistical uncertainty while obtaining perfect parallelism (after the initial time to reach equilibrium has been discarded). Some of these simulations might be conducted at a higher temperature than the temperature of interest to promote

[2] An implementation of the method for the analysis of simulated parallel tempering simulations can be found at `http://www.dillgroup.ucsf.edu/~jchodera/code/wham`.

greater sampling across barriers, for example. Sometimes, the expectation value of one or more observables is desired over a range of temperatures. Additionally, simulations started from different initial conditions can be used as a check of equilibration and convergence [72]. Below, we follow roughly the same approach as Kumar *et al.* [100] in deriving the WHAM equations, though our notation differs substantially and we include a more detailed treatment of statistical uncertainty. Additionally, we arrive at a novel way of computing expectations of multiple observables and avoid the use of many-dimensional histograms. While the method presented in [100] has the full generality of treating simulations conducted with arbitrary biasing potentials, we focus on the case of independent canonical simulations at different temperatures, since variations on this approach will allow us to consider simulated and parallel tempering simulations in Section 2.3. (For an informative treatment of the case of a multiple biasing potentials at a single temperature, as in the case of umbrella sampling, see [165].)

2.2.1 Motivation and Definitions.

Suppose we have an observable A that is only a function of the Cartesian coordinates of the system $\mathbf{q}$, and we wish to estimate the expectation of A over the canonical ensemble at some temperature of interest T. Instead of this temperature T, we will generally refer to its corresponding inverse temperature $\beta = (k_B T)^{-1}$, where k_B is the Boltzmann constant. We denote the expectation of A over the canonical ensemble at inverse temperature β by $\langle A \rangle_\beta$, which can be written as

$$\langle A \rangle_\beta = \frac{\int d\mathbf{q}\, e^{-\beta U(\mathbf{q})}\, A(\mathbf{q})}{\int d\mathbf{q}\, e^{-\beta U(\mathbf{q})}} \tag{2.1}$$

where $U(\mathbf{q})$ is the potential energy function of the system.

Further suppose we have carried out K independent simulations that sample from the canonical ensemble (using such techniques as Metropolis Monte Carlo or thermally controlled molecular dynamics) at corresponding inverse temperatures $\beta_1, \beta_2, \ldots, \beta_K$, some or all of which may be different from the temperature of interest. We denote the coordinates and potential energies sampled at a fixed time interval Δt from simulation k by the time series $\{\mathbf{q}_{kn}, U_{kn}\}_{n=1}^{N_k}$, where $U_{kn} = U(\mathbf{q}_{kn})$ and N_k is the number of configurations collected from simulation k.

We first consider the probability density function from which the configurations are generated in simulation k. For a simulation sampling from the canonical distribution, the probability of generating a configuration with potential energy in the interval dU about U at inverse temperature

β is given by

$$p(U|\beta)\,dU \quad = \quad [Z(\beta)]^{-1}\,\Omega(U)\,dU\,e^{-\beta U} \tag{2.2}$$

with the normalizing constant $Z(\beta)$, often referred to as the *configurational partition function*, chosen to ensure that $p(U)$ integrates to unity. The quantity $\Omega(U)$ is the *potential energy density of states*, and $\Omega(U)\,dU$ represents the volume of configuration space with potential energy in the interval dU around U.

While the Boltzmann factor $e^{-\beta_k U}$ and normalization constant $Z(\beta_k)$ differ for each simulation k, the density of states $\Omega(U)$ is independent of temperature. Since the Boltzmann factor is a known function and the configurational partition function is simply a normalizing constant, knowledge of the density of states allows the potential energy probability density to be computed at *any* temperature. If the average of the observable A over all configurations with potential energy U is known, these can be combined to give the expectation at a desired inverse temperature β

$$\langle A \rangle_\beta \quad = \quad \frac{\int dU\,\Omega(U)\,e^{-\beta U}\,A(U)}{\int dU\,\Omega(U)\,e^{-\beta U}} \tag{2.3}$$

where $A(U)$ is defined as the average of A over all configurations with potential energy U

$$A(U') \quad \equiv \quad \frac{\int d\mathbf{q}\,\delta(U(\mathbf{q}) - U')\,A(\mathbf{q})}{\int d\mathbf{q}\,\delta(U(\mathbf{q}) - U')}. \tag{2.4}$$

It is easily seen that substituting this expression into Eq. 2.3 recovers the configuration space average in Eq. 2.1.

Our aim is to obtain the best estimate of the density of states and the expectation of the observable by combining information from several simulations. Since each simulation samples an energy range determined by its temperature, our final estimate of the density of states will be more accurate if we account for the different uncertainties in the estimate obtained from each simulation. We will therefore need a separate estimate of the density of states and its corresponding uncertainty from *each* simulation.

2.2.2 Obtaining an Estimate of the Density of States from Each Simulation.

To obtain an estimate of the density of states from each simulation, we first need a way of mathematically expressing the form of the observed probability density function $p(U)$. While it may be possible to assume a particular functional form for this density, this would generally be inexact. A better approach is to use a *nonparametric density estimator* (see, for example, [159] for

an overview) that makes no prior assumptions as to the true functional form of $p(U)$. Kumar *et al.* [100], as Ferrenberg and Swendsen [57,58] earlier, chose a histogram-based estimator, in which the range of sampled energies is discretized into a set of nonoverlapping bins of equal width. While there are a number of more sophisticated smooth nonparametric estimators [196], the histogram estimator is simpler and more efficient to apply.

Accordingly, we construct an estimate of the probability density function $p(U|\beta)$ on a set of M points, labeled U_m, that span the sampled potential energy range and are spaced ΔU apart. We denote the estimate of $p(U|\beta)$ at U_m by $p_m(\beta)$, and the corresponding estimate of the density of states $\Omega(U_m)$ by Ω_m.

$$p_m(\beta) \equiv p(U_m|\beta) = [Z(\beta)]^{-1}\Omega_m e^{-\beta U_m}. \tag{2.5}$$

The normalization factor $Z(\beta)$ can then be approximated by a discretized integration

$$\begin{aligned} Z(\beta) &= \int dU\, \Omega(U)\, e^{-\beta U} \\ &\approx \sum_{m=1}^{M} \Delta U\, \Omega_m\, e^{-\beta U_m} \end{aligned} \tag{2.6}$$

where we have made the assumption that the integrand, $\Omega(U)\, e^{-\beta U}$, does not change significantly over the bin width ΔU. As the density of states $\Omega(U)$ increases with U and the Boltzmann factor $e^{-\beta U}$ decreases, their product is expected to vary less rapidly than either term individually.

We define $\psi_m(U)$ as the indicator or characteristic function for the energy bin of width ΔU centered about U_m

$$\psi_m(U) = \begin{cases} 1 & \text{if } U \in [U_m - \Delta U/2, U_m + \Delta U/2) \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

and the time series defined by this indicator function as $\{\psi_{mkn}\}_{n=1}^{N_k}$, where $\psi_{mkn} \equiv \psi_m(U_{kn})$. We denote the count of configurations from simulation k that fall in energy bin m — the "histogram" from which the weighted histogram analysis method derives its name — by H_{mk}, and see that it can be computed by

$$H_{mk} = \sum_{n=1}^{N_k} \psi_{mkn}. \tag{2.8}$$

We will also use the total number of configurations over all simulations that fall in energy bin m, which we term H_m:

$$H_m = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \psi_{mkn}. \tag{2.9}$$

Note that through out our discussion, pairs of variables that only differ by the number of written subscripts, such as H_m and H_{mk}, represent similar quantities related in this way.
We can estimate $p_m(\beta_k)$ by the number of configurations sampled from the simulation at temperature β_k with energies that fall in the bin centered about U_m:

$$p_m(\beta_k) \approx \frac{1}{\Delta U} \cdot \frac{H_{mk}}{N_k}. \tag{2.10}$$

Equating this with the definition of p_m from Eq. 2.5 and rearranging terms, we can obtain an estimate of Ω_{mk}, the density of states at energy U_m from simulation k, which we will denote by $\hat{\Omega}_{mk}$:

$$\begin{aligned} \hat{\Omega}_{mk} &= \frac{1}{\Delta U} \cdot \frac{H_{mk}}{N_k} \cdot \frac{Z(\beta_k)}{e^{-\beta_k U_m}} \\ &= \frac{H_{mk}}{N_k \, \Delta U \, \exp[f_k - \beta_k U_m]}. \end{aligned} \tag{2.11}$$

In the last step, we have replaced the partition function $Z(\beta_k)$ by an exponentiated dimensionless free energy $f_k \equiv -\ln Z(\beta_k)$. Each independent simulation k contributes an estimate of the density of states $\hat{\Omega}_{mk}$ for energy bin m. Each of these estimates in turn carries a statistical uncertainty $\delta^2\hat{\Omega}_{mk}$, determined primarily by the number of uncorrelated samples of the energy bin. (Expressions for $\delta^2\hat{\Omega}_{mk}$ will be derived later in Section 2.2.5.) We will combine these individual estimates $\hat{\Omega}_{mk}$ to produce a single optimal estimator $\hat{\Omega}_m$ in a such way that the statistical uncertainty in the resulting estimate is minimized, giving more weight to the $\hat{\Omega}_{mk}$ with smaller uncertainties. To do this, we must first briefly review the maximum-likelihood method for combining independent measurements with associated uncertainties into an optimal estimate, and also consider the uncertainty in a mean computed from a set of correlated observations.

2.2.3 Optimal Estimator from Independent Observations and Associated Uncertainties.

Suppose we have K independent observations or measurements of some random variable X denoted $x_1, \ldots, x_K$, each with corresponding squared uncertainty $\delta^2 x_k$, defined by

$$\delta^2 x_k \equiv \langle (x_k - \langle x_k \rangle)^2 \rangle = \langle x_k^2 \rangle - \langle x_k \rangle^2 \tag{2.12}$$

where $\langle \cdot \rangle$ here denotes the expectation over repeated measurements or experimental trials. We can then write $\hat{X}$, the optimal estimator for $\langle X \rangle$ in the sense of minimizing $\delta^2\hat{X}$, by a weighted sum of

the individual estimates

$$\hat{X} = \frac{\sum_{k=1}^{K} [\delta^2 x_k]^{-1} x_k}{\sum_{k=1}^{K} [\delta^2 x_k]^{-1}}. \tag{2.13}$$

Note that observations with smaller uncertainties get greater weight, and if all the uncertainties are equal, the weight is simply $1/K$, as would be expected.
The uncertainty in the resulting estimate is simply given by

$$\delta^2 \hat{X} = \left\{ \sum_{k=1}^{K} [\delta^2 x_k]^{-1} \right\}^{-1} \tag{2.14}$$

These are standard formulas that come from maximum likelihood considerations [31].

2.2.4 Statistical Uncertainty in the Estimator for Correlated Time Series Data.

We briefly review the estimation of statistical uncertainty for a time series of correlated measurements. (See Müller-Krumbhaar *et al.* [127] for an early exposition of this method as applied to the analysis of Monte Carlo simulations of spin systems, Swope *et al.* [173] for the analysis of molecular dynamics simulations, or Janke [87] for a recent general illustration.) Suppose we have a time series of correlated sequential observations of the random variable X denoted $\{x_n\}_{n=1}^{N}$ that come from a stationary, time-reversible stochastic process. Our estimate for the expectation of X is given by the time average

$$\hat{X} = \frac{1}{N} \sum_{n=1}^{N} x_n \tag{2.15}$$

but the statistical uncertainty is more complicated than in the independent observation case

$$\begin{aligned} \delta^2 \hat{X} &\equiv \langle (\hat{X} - \langle \hat{X} \rangle)^2 \rangle = \langle \hat{X}^2 \rangle - \langle \hat{X} \rangle^2 \\ &= \frac{1}{N^2} \sum_{n,n'=1}^{N} \left[\langle x_n x_{n'} \rangle - \langle x_n \rangle \langle x_{n'} \rangle \right] \\ &= \frac{1}{N^2} \sum_{n=1}^{N} \left[\langle x_n^2 \rangle - \langle x_n \rangle^2 \right] \\ &+ \frac{1}{N^2} \sum_{n \neq n'=1}^{N} \left[\langle x_n x_{n'} \rangle - \langle x_n \rangle \langle x_{n'} \rangle \right]. \end{aligned} \tag{2.16}$$

In the last step, we have split the sum into two sums — a term capturing the variance in the observations, and a remaining term capturing the correlation between observations. Using the properties of stationarity and time-reversibility, we can further manipulate this to obtain

$$\begin{aligned} \delta^2 \hat{X} &= \frac{1}{N}\left[\langle x_n^2 \rangle - \langle x_n \rangle^2\right] \\ &+ \frac{2}{N}\sum_{t=1}^{N-1}\left(\frac{N-t}{N}\right)\left[\langle x_n x_{n+t} \rangle - \langle x_n \rangle \langle x_{n+t} \rangle\right] \\ &\equiv \frac{\sigma_x^2}{N}(1+2\tau) = \frac{\sigma_x^2}{g^{-1}N} \end{aligned} \tag{2.17}$$

where the variance σ_x^2, statistical inefficiency g, and integrated autocorrelation time τ (in units of the sampling interval) are given by

$$\sigma_x^2 \equiv \langle x_n^2 \rangle - \langle x_n \rangle^2 \tag{2.18}$$

$$\tau \equiv \sum_{t=1}^{N-1}\left(1 - \frac{t}{N}\right) C_t \tag{2.19}$$

$$g \equiv 1 + 2\tau \tag{2.20}$$

with the discrete-time normalized fluctuation autocorrelation function C_t defined as

$$C_t \equiv \frac{\langle x_n x_{n+t} \rangle - \langle x_n \rangle^2}{\langle x_n^2 \rangle - \langle x_n \rangle^2}. \tag{2.21}$$

The quantity $g \equiv (1 + 2\tau) \geq 1$ can be thought of as a *statistical inefficiency*, in that $g^{-1}N$ gives the effective number of *uncorrelated* configurations contained in the time series. The statistical inefficiency will depend on the time interval at which configurations are collected for analysis; longer intervals will reduce the statistical inefficiency, which will approach unity as the sampling interval exceeds the correlation time. Practically, we use our best estimates for the variance σ_x^2 and autocorrelation function C_t to compute an estimate of the statistical uncertainty $\delta^2 \hat{X}$.

2.2.5 Optimal Estimate of the Density of States.

We now construct an optimal estimator of the density of states Ω_m from the individual estimates obtained from the K independent canonical simulations. From the results of Section 2.2.3, we can

write this estimator and its corresponding uncertainty as

$$\hat{\Omega}_m = \frac{\sum_{k=1}^{K} [\delta^2 \hat{\Omega}_{mk}]^{-1} \hat{\Omega}_{mk}}{\sum_{k=1}^{K} [\delta^2 \hat{\Omega}_{mk}]^{-1}} \tag{2.22}$$

$$\delta^2 \hat{\Omega}_m = \left\{ \sum_{k=1}^{K} [\delta^2 \hat{\Omega}_{mk}]^{-1} \right\}^{-1}. \tag{2.23}$$

The results of Section 2.2.4 show us how to write the $\delta^2 \hat{\Omega}_{mk}$, the uncertainty in our estimate of the density of states for energy bin m from simulation k. In Eq. 2.11 above, we see that this uncertainty comes only from $\delta^2 H_{mk}$, the uncertainty in the histogram count for the energy bin, since all other terms are known with certainty:

$$\delta^2 \hat{\Omega}_{mk} = \frac{\delta^2 H_{mk}}{\{N_k \, \Delta U \, \exp[f_k - \beta_k U_m]\}^2}. \tag{2.24}$$

H_{mk}, the histogram count from simulation k, can be written as a time average of the indicator function ψ_m over the correlated configurations collected from the simulation:

$$H_{mk} = N_k \cdot \frac{1}{N_k} \sum_{n=1}^{N_k} \psi_{mkn}. \tag{2.25}$$

We can use the result of Section 2.2.4 above to obtain an expression for $\delta^2 H_{mk}$, the uncertainty in the histogram count:

$$\begin{aligned} \delta^2 H_{mk} &= N_k^2 \cdot \frac{\sigma_{mk}^2}{N_k} g_{mk} \\ &= g_{mk} \, N_k \, \left(\langle \psi_{mk}^2 \rangle - \langle \psi_{mk} \rangle^2 \right) \\ &= g_{mk} \, N_k \langle \psi_{mk} \rangle \, (1 - \langle \psi_{mk} \rangle) \\ &= g_{mk} \, \langle H_{mk} \rangle \, \left(1 - \frac{\langle H_{mk} \rangle}{N_k} \right) \end{aligned} \tag{2.26}$$

where, because $\psi_m(U)$ is an indicator function (eq 2.7), $[\psi_m(U)]^2 = \psi_m(U)$. If the histograms are sparsely populated, a reasonable assumption if there are a sufficient number of histogram bins spanning the energy range sampled by each simulation, then $\langle H_{mk} \rangle / N_k \ll 1$, and we can further simplify this to

$$\delta^2 H_{mk} \approx g_{mk} \, \langle H_{mk} \rangle. \tag{2.27}$$

The statistical inefficiency g_{mk} here reflects the number of configurations required for an uncorrelated sampling of the energy bin. This will, in general, depend on the bin index, bin width,

and temperature. This dependence was omitted in the original Kumar *et al.* presentation [100]. At higher temperatures, the correlation time, and hence the statistical inefficiency, is expected to be smaller as the simulation can move through configuration space more easily. The structure of the energy landscape may cause the simulation to be stuck in certain regions of configuration space for different times, hence the dependence on energy bin index is also potentially important.
The expectation $\langle H_{mk} \rangle$ should be replaced by our best estimate of the histogram count for energy bin m at temperature β_k, which could be obtained from our yet-to-be-determined optimal estimate of the density of states $\hat{\Omega}_m$:

$$\begin{aligned} \langle H_{mk} \rangle &= N_k \, p_m(\beta_k) \, \Delta U \\ &\approx N_k \, \Delta U \, \hat{\Omega}_m \, \exp[f_k - \beta_k U_m] \end{aligned} \tag{2.28}$$

Substituting this expression back into Eqs. 2.27 and 2.24, we obtain

$$\begin{aligned} \delta^2 \hat{\Omega}_{mk} &= \frac{g_{mk} \, N_k \, \Delta U \, \hat{\Omega}_m \, \exp[f_k - \beta_k U_m]}{\{N_k \, \Delta U \, \exp[f_k - \beta_k U_m]\}^2} \\ &= \frac{\hat{\Omega}_m}{g_{mk}^{-1} \, N_k \, \Delta U \, \exp[f_k - \beta_k U_m]}. \end{aligned} \tag{2.29}$$

Using Eq. 2.29 for the uncertainty in the density of states associated with simulation k, and Eq. 2.22 for the best estimate of the density of states, along with Eq. 2.11, we obtain

$$\hat{\Omega}_m = \frac{\sum_{k=1}^{K} g_{mk}^{-1} \, H_{mk}}{\sum_{k=1}^{K} g_{mk}^{-1} \, N_k \, \Delta U \, \exp[f_k - \beta_k U_m]} \tag{2.30}$$

Everything in the above expression can be easily evaluated, except for the f_k, which depend on the $\hat{\Omega}_m$ through

$$f_k = -\ln \sum_{m=1}^{M} \hat{\Omega}_m \, \Delta U \, e^{-\beta_k U_m}. \tag{2.31}$$

The f_k may therefore be solved for self consistency by iteration of Eqs. 2.30 and 2.31 starting from an arbitrary choice, such as $f_k = 0$.
The statistical uncertainty in $\hat{\Omega}_m$ is given by Eq. 2.14:

$$\delta^2 \hat{\Omega}_m = \frac{\hat{\Omega}_m}{\sum_{k=1}^{K} g_{mk}^{-1} \, N_k \, \Delta U \, \exp[f_k - \beta_k U_m]}. \tag{2.32}$$

We note that the relative uncertainty in this estimate is given by

$$\frac{\delta^2\hat{\Omega}_m}{\hat{\Omega}_m^2} = \left[\sum_{k=1}^{K} g_{mk}^{-1} H_{mk}\right]^{-1} \tag{2.33}$$

which is approximately equal to H_m^{-1}, the inverse of the total number of configurations from all simulations in energy bin m, if all g_{mk} are unity. This is reasonable, since the uncertainty in our estimate for Ω_m should diminish as more independent samples are collected in energy bin m.

2.2.6 Estimating an Observable at the Temperature of Interest.

Using the estimate of the density of states obtained above, we can obtain an estimate for the expectation of any configuration function $A(\mathbf{q})$ at an arbitrary temperature by writing analogous equations to Eqs. 2.3 and 2.4 where we have discretized the energy U:

$$\langle A \rangle_\beta \approx \frac{\sum_{m=1}^{M} \hat{\Omega}_m \, \Delta U \, e^{-\beta U_m} \, A_m}{\sum_{m=1}^{M} \hat{\Omega}_m \, \Delta U \, e^{-\beta U_m}} \tag{2.34}$$

where

$$A_m = \frac{\int d\mathbf{q} \, A(\mathbf{q}) \, \psi_m(U(\mathbf{q}))}{\int d\mathbf{q} \, \psi_m(U(\mathbf{q}))}. \tag{2.35}$$

A_m, the mean of observable A over all configurations with potential energies consistent with energy bin m, can be best approximated by pooling configurations from *all* K simulations that have energies in bin m:

$$\hat{A}_m = H_m^{-1} \sum_{k=1}^{K} \sum_{n=1}^{N_k} \psi_{mkn} \, A_{kn} \tag{2.36}$$

where $H_m = \sum_{k=1}^{K} \sum_{n=1}^{N_k} \psi_{mkn}$ is the total count of configurations in energy bin m from all simulations. Substituting this expression for A_m into Eq. 2.34 above produces an estimator $\hat{A}(\beta)$

for $\langle A \rangle_\beta$:

$$\begin{aligned}
\hat{A}(\beta) &= \frac{\sum_{m=1}^{M} \hat{\Omega}_m \, \Delta U \, e^{-\beta U_m} \, A_m}{\sum_{m=1}^{M} \hat{\Omega}_m \, \Delta U \, e^{-\beta U_m}} \\
&= \frac{\sum_{m=1}^{M} \hat{\Omega}_m \, e^{-\beta U_m} \left[H_m^{-1} \sum_{k=1}^{K} \sum_{n=1}^{N_k} \psi_{mkn} \, A_{kn} \right]}{\sum_{m=1}^{M} \hat{\Omega}_m \, e^{-\beta U_m} \left[H_m^{-1} \sum_{k=1}^{K} \sum_{n=1}^{N_k} \psi_{mkn} \right]} \\
&= \frac{\sum_{k=1}^{K} \sum_{n=1}^{N_k} w_{kn}(\beta) \, A_{kn}}{\sum_{k=1}^{K} \sum_{n=1}^{N_k} w_{kn}(\beta)}
\end{aligned} \tag{2.37}$$

where we have defined the per-configuration weights $w_{kn}(\beta)$ by

$$w_{kn}(\beta) = \sum_{m=1}^{M} \psi_{mkn} \, H_m^{-1} \, \hat{\Omega}_m \, e^{-\beta U_m} \tag{2.38}$$

where only one term of the sum will contribute — the bin m containing the energy U_{kn} — due to the presence of the indicator function ψ_{mkn}. H_m is the total count of configurations from all simulations with energy in bin m. Note that we only need to compute the weight w_{kn} up to a constant of proportionality because this constant drops out in the normalized sum in Eq. 2.37. This relationship is significant in that we now have an expression for the canonical expectation of observable A in terms of a weighted sum over *all* of the data. These weights are determined by the temperature of interest from the WHAM equations, and are simple functions of the count of configurations with energies falling in a particular energy bin. The weights $w_{kn}(\beta)$ can be computed once for the temperature of interest and then used to calculate expectations of many observables.

It should be noted that our estimate of $\langle A \rangle_\beta$ will only be reasonable if the inverse temperature of interest β lies near or within the range of inverse temperatures sampled by the canonical simulations — the uncertainty in the estimate will increase as the temperature of interest deviates from the sampled range of temperatures (see [56, 129] for an examination of this issue).

2.2.7 Statistical Uncertainty of the Estimator for the Expectation.

If the observable of interest has a long correlation time compared to fluctuations in the potential energy (e.g., if the observable is a function of the large scale molecular conformation), then it is possible that the density of states Ω_m and dimensionless free energies $\{f_l\}$ may be sufficiently well-converged that they are not dominant contributors to the uncertainty in the estimate of the observable of interest. Instead, the long time-correlation in the observable means that there are many fewer effectively independent observations of the observable than stored configurations. We may then use the following procedure.

We can rewrite the expectation $\langle A \rangle_\beta$ as a ratio of two random quantities X and Y

$$\langle A \rangle_\beta \quad = \quad \frac{X}{Y} \tag{2.39}$$

where

$$\hat{X} \equiv \sum_{k=1}^{K} \sum_{n=1}^{N_k} w_{kn}(\beta) A_{kn} \quad ; \quad \hat{Y} \equiv \sum_{k=1}^{K} \sum_{n=1}^{N_k} w_{kn}(\beta). \tag{2.40}$$

Applying standard error propagation techniques for a function of random variables (see, *e.g.* [177]), which amounts to a first-order Taylor series expansion of $\hat{A}$ about $\langle X \rangle / \langle Y \rangle$, we can estimate the uncertainty in $\hat{A}$ as

$$\delta^2 \hat{A} \quad = \quad \left[\frac{\hat{X}}{\hat{Y}} \right]^2 \left[\frac{\delta^2 \hat{X}}{\hat{X}^2} + \frac{\delta^2 \hat{Y}}{\hat{Y}^2} - 2 \frac{\delta \hat{X} \delta \hat{Y}}{\hat{X}\,\hat{Y}} \right]. \tag{2.41}$$

Here, the cross-term $\delta \hat{X} \delta \hat{Y} \equiv \langle (\hat{X} - \langle \hat{X} \rangle)(\hat{Y} - \langle \hat{Y} \rangle) \rangle$ is nonzero only if the random variables X and Y are correlated, in which case the term involving it in the equation above serves to reduce the uncertainty in the estimate of the ratio $\hat{A}$.

Recognizing that X and Y include contributions from K statistically independent simulations, we can collect these terms and write

$$\begin{aligned} X \equiv \sum_{k=1}^{K} N_k \, X_k \quad ; \quad Y \equiv \sum_{k=1}^{K} N_k \, Y_k \\ X_k \equiv \frac{1}{N_k} \sum_{n=1}^{N_k} w_{kn} \, A_{kn} \quad ; \quad Y_k \equiv \frac{1}{N_k} \sum_{n=1}^{N_k} w_{kn} \end{aligned} \tag{2.42}$$

where the argument β has been omitted for notational convenience. Because the K individual

simulations are *independent*, the uncertainties required in Eq. 2.41 are given by

$$\delta^2 X = \sum_{k=1}^{K} N_k^2 \, \delta^2 X_k \quad ; \quad \delta^2 Y = \sum_{k=1}^{K} N_k^2 \, \delta^2 Y_k$$
$$\delta X \delta Y = \sum_{k=1}^{K} N_k^2 \, \delta X_k \delta Y_k. \tag{2.43}$$

These uncertainties involve the correlated data of simulation k and can be estimated by standard correlation analysis methods [87, 173] or by block transformation methods [64], though the latter method requires some modification to estimate the uncertainty cross-term $\delta X_k \delta Y_k$.
To compute the uncertainties by correlation analysis methods as in Section 2.2.4, we first define new observables $x_{kn} = w_{kn} A_{kn}$ and $y_{kn} = w_{kn}$, and compute the uncertainties

$$\delta^2 X_k = \frac{\sigma^2_{k,x;x}}{g^{-1}_{k,x;x} N_k} \quad ; \quad \delta^2 Y_k = \frac{\sigma^2_{k,y;y}}{g^{-1}_{k,y;y} N_k}$$
$$\delta X_k \delta Y_k = \frac{\sigma^2_{k,x;y}}{g^{-1}_{k,x;y} N_k}. \tag{2.44}$$

These uncertainties involve (co)variances of the type $\sigma^2_{k,x;y}$, estimated for each replica by

$$\hat{\sigma}^2_{k,x;y} = \frac{1}{N_k - 1} \sum_{n=1}^{N_k} (x_{kn} - \hat{X}_k)(y_{kn} - \hat{Y}_k). \tag{2.45}$$

The statistical inefficiencies of the form $g_{k,x;y}$ are computed by

$$g_{k,x;y} \equiv 1 + 2\tau_{k,x;y} \tag{2.46}$$
$$\tau_{k,x;y} \equiv \sum_{t=1}^{N_k - 1} \left(1 - \frac{t}{N_k}\right) C_{kt,x;y} \tag{2.47}$$

with the correlation function for simulation k computed by taking advantage of stationarity and time-reversibility:

$$C_{kt,x;y} \approx \frac{1}{2\hat{\sigma}^2_{k,x;y}} \frac{1}{(N_k - t)} \times \sum_{n=1}^{N_k - t} \Big[(x_{kn} - \hat{X}_k)(y_{kn+t} - \hat{Y}_k) + (y_{kn} - \hat{Y}_k)(x_{kn+t} - \hat{X}_k) \Big]. \tag{2.48}$$

See Section 2.5.2 for a discussion on efficiently computing the integrated correlation time τ from $\hat{C}_{kt;x,y}$.

2.3 Simulated and Parallel Tempering

2.3.1 Simulated Tempering.

In a simulated tempering simulation [117, 120], a single simulation is conducted in which configurations are sampled from a *mixed-canonical* ensemble [61]. In practice, a simulation algorithm that samples from the canonical ensemble is used to generate configurations, and at regular intervals attempts are made to change the temperature among a discrete set of choices $\beta_1, \ldots, \beta_L$. The probability of accepting a proposed temperature change is given by the Metropolis-like criterion

$$P(\beta_l \rightarrow \beta_{l'}) = \min\{1, \exp[-(\beta_{l'} - \beta_l)U + (a_{l'} - a_l)]\} \tag{2.49}$$

where the constants $a_l,\ l = 1, \ldots, L$ are specified beforehand and chosen, often by tedious exploratory simulations, to attempt to achieve near-equal visitation of each temperature and, hopefully, potential energy. The optimal choice of $\{a_l\}$ is given by the dimensionless free energies $\{f_k\}$ in Eq. above, and proposed temperature changes are usually between neighboring temperatures because the exchange probability diminishes with increased temperature separation. Use of the above criterion for accepting or rejecting proposed temperature changes ensures that, if the configurations were originally distributed from the equilibrium distribution at the old temperature, they are also distributed from the canonical distribution at the new temperature. As a result of this procedure, the system spends a fraction of time in each of a number of different temperatures. Since we know the number of times each temperature was visited, we can write the probability density for energy bin m as a weighted sum of the canonical probability density functions at these different temperatures:

$$p_m = \sum_{l=1}^{L} \frac{N_l}{N} \cdot \frac{\Omega_m\, e^{-\beta_l U_m}}{Z(\beta_l)} \tag{2.50}$$

where N_l/N is the fraction of configurations generated at inverse temperature β_l over the course of the simulation. As above, we introduce the Helmholtz free energy $f_l \equiv -\ln Z(\beta_l)$, which allows us to write

$$\begin{aligned} p_m &= \sum_{l=1}^{L} \frac{N_l}{N}\, \Omega_m\, \exp[f_l - \beta_l U_m] \\ &= \Omega_m \sum_{l=1}^{L} (N_l/N)\, \exp[f_l - \beta_l U_m]. \end{aligned} \tag{2.51}$$

We can approximate p_m as before using our histogram count, H_m, the number of configurations with potential energy in the bin centered about U_m:

$$p_m \approx \frac{1}{\Delta U} \cdot \frac{H_m}{N}. \tag{2.52}$$

Rearranging and including our definition of f_l, we obtain the coupled set of equations for estimating the density of states

$$\hat{\Omega}_m = \frac{H_m}{\sum_{l=1}^{L} N_l \, \Delta U \, \exp[f_l - \beta_l U_m]} \tag{2.53}$$

$$f_l = -\ln \sum_{m=1}^{M} \hat{\Omega}_m \, \Delta U \, e^{-\beta_l U_m} \tag{2.54}$$

These equations are similar to Eqs. 2.30 and 2.31 for the canonical ensemble WHAM if the configurations are grouped by the temperature at which they were generated, but lacking statistical inefficiency terms since we are not combining data from multiple simulations.
The uncertainty in $\hat{\Omega}_m$ is then given by

$$\begin{aligned} \delta^2 \hat{\Omega}_m &= \frac{\delta^2 H_m}{\left\{ \sum_{l=1}^{L} N_l \, \Delta U \, \exp[f_l - \beta_l U_m] \right\}^2} \\ &= \frac{g_m \langle H_m \rangle}{\left\{ \sum_{l=1}^{L} N_l \, \Delta U \, \exp[f_l - \beta_l U_m] \right\}^2} \end{aligned} \tag{2.55}$$

where, as in Eq. 2.27, we assume the histograms are sparsely populated and introduce the statistical inefficiency g_m to estimate the histogram uncertainty.
Th estimate for the expectation of the total histogram count in energy bin m is given by the sampling probability

$$\begin{aligned} \langle H_m \rangle &= N \, \Delta U \, p_m \\ &\approx N \, \Delta U \, \hat{\Omega}_m \sum_{l=1}^{L} (N_l/N) \, \exp[f_l - \beta_l U_m] \end{aligned} \tag{2.56}$$

which gives the final estimate for the uncertainty as

$$\delta^2 \hat{\Omega}_m = \frac{\hat{\Omega}_m}{\sum_{l=1}^{L} g_m^{-1} \, N_l \, \Delta U \, \exp[f_l - \beta_l U_m]}. \tag{2.57}$$

Following the approach in Section 2.2.6, we can again write the estimator in the form of a weighted sum over configurations:

$$\hat{A}(\beta) = \frac{\sum_{n=1}^{N} w_n(\beta)\, A_n}{\sum_{n=1}^{N} w_n(\beta)} \tag{2.58}$$

$$w_n(\beta) = \sum_{m=1}^{M} \psi_{mn}\, H_m^{-1}\, \hat{\Omega}_m e^{-\beta U_m} \tag{2.59}$$

where again, only one term contributes to the sum in the expression for the weight w_n. The statistical uncertainty in this estimate, as in Section 2.2.7, can be computed by Eq. 2.41, where X and Y are now given by

$$X \equiv \frac{1}{N} \sum_{n=1}^{N} w_n\, A_n \quad ; \quad Y \equiv \frac{1}{N} \sum_{n=1}^{N} w_n \tag{2.60}$$

These uncertainties are simply computed as in Eqs. 2.44 – 2.48, without the subscript k as there is only one simulation instead of many. The quantities $\hat{X}$ and $\hat{Y}$ no longer correspond to canonical averages, since they are the expectations over the simulated tempering trajectory which spends a different amount of time at each of the L temperatures — it is a mixed canonical average. The sample mean over the trajectory provides the best estimator for these quantities.

Here, the statistical inefficiency g_m appearing in Eq. 2.57 and the inefficiencies required in applying Eqs. 2.44 – 2.48 are computed from the correlation functions computed over the simulated tempering trajectory, which includes unphysical jumps in temperature. It is worth noting that expressions for $\langle A \rangle_\beta$ given in formulations by Okamoto and coworkers (*e.g.* Eq. 24 of [121]) instead contain a statistical inefficiency for each *temperature*. In principle, one could account for a temperature-dependent statistical inefficiency, since one might expect correlation times to be different at each temperature, but in practice, the limited number of configurations sampled between temperature changes is likely too short to allow temperature-dependent correlation times to be computed. Additionally, a temperature-dependent treatment does not account for the correlation between configurations sampled before and after a temperature swap. The derivation presented here assumes the statistical inefficiency g_m depends only on the energy bin m, which causes these factors to cancel out of our estimator for $\langle A \rangle_\beta$ in Eqs. 2.58 and 2.59.

2.3.2 Parallel Tempering or Independent Simulated Tempering Simulations.

In a parallel tempering (or replica-exchange among temperatures) simulation [77, 171], it was recognized that the constants a_k needed in the simulated tempering simulation to ensure equal sampling of temperatures could be eliminated if multiple simulated tempering simulations were conducted in parallel and the temperature changes of two simulations were coupled together into a temperature swap between the replicas. In practice, a number K of *replicas* are simulated independently at inverse temperatures $\beta_1, \ldots, \beta_K$ using some simulation method that samples from the canonical distribution. At given intervals, an attempt is made to exchange the temperatures of two replicas i and j, with the exchange accepted with probability

$$\begin{aligned} P_{\text{exch}} &= \min\{1, \exp[-(\beta_j - \beta_i)U_i + (a_j - a_i)]\} \\ &\times \min\{1, \exp[-(\beta_i - \beta_j)U_j + (a_i - a_j)]\} \\ &= \min\{1,\ \exp[-(\beta_j - \beta_i)(U_i - U_j)]\} \end{aligned} \tag{2.61}$$

where β_i is the current inverse temperature of replica i and U_i the corresponding potential energy. Because of this exchange procedure, each replica executes a more or less random walk in temperature, eliminating the need to perform exploratory simulations to determine the parameters $\{a_i\}_{i=1}^K$ required for simulated tempering. Each replica simulation is nearly independent, as the correlation between configurations of different replicas introduced by the exchange of temperatures is minimal. The dominant contribution to statistical uncertainties will almost certainly be due to the variance and temporal correlation in the value of the observable of interest within each replica, which reduces the effective number of independent samples. We can therefore analyze a parallel tempering simulation as a set of *independent* simulated tempering simulations, each with a number L of accessible temperatures, with L equal to the number of replicas K. Below, we derive an analogue of the Kumar *et al.* WHAM procedure for the treatment of K independent simulated tempering simulations (replicas) each capable of visiting L temperatures, allowing this method to also treat simulations generated by procedures such as REST [120]. We make use of the sampling distribution for simulated tempering described above and properly account for the correlation within each replica, eliminating the need to artificially reorder configurations from parallel tempering simulations by temperature.

We can use the simulated tempering Eqs. 2.53 and 2.57 above to write the estimator and

uncertainty for the density of states obtained from each replica k as

$$\hat{\Omega}_{mk} = \frac{H_{mk}}{\sum_{l=1}^{L} N_{kl}\, \Delta U\, \exp[f_l - \beta_l U_m]} \tag{2.62}$$

$$\delta^2 \hat{\Omega}_{mk} = \frac{\hat{\Omega}_m}{g_{mk}^{-1} \sum_{l=1}^{L} N_{kl}\, \Delta U\, \exp[f_l - \beta_l U_m]} \tag{2.63}$$

where we have added the index k to denote the *replica* from which the data is generated. H_{mk} therefore denotes the number of configurations sampled with potential energy in energy bin m from replica k, and N_{kl} the number at temperature β_l from replica k. g_{mk} is the statistical inefficiency computed from replica k for energy bin m.

Again using the optimal combination rule of Eq. 2.13, we obtain the optimal estimate for the density of states

$$\hat{\Omega}_m = \frac{\sum_{k=1}^{K} g_{mk}^{-1} H_{mk}}{\sum_{k=1}^{K} g_{mk}^{-1} \sum_{l=1}^{L} N_{kl}\, \Delta U\, \exp[f_l - \beta_l U_m]} \tag{2.64}$$

and the statistical uncertainty from Eq. 2.14:

$$\begin{aligned} \delta^2 \hat{\Omega}_m &= \left\{ \sum_{k=1}^{K} \left[\frac{g_{mk}^{-1} \sum_{l=1}^{L} N_{kl}\, \Delta U\, \exp[f_l - \beta_l U_m]}{\hat{\Omega}_m} \right] \right\}^{-1} \\ &= \frac{\hat{\Omega}_m}{\sum_{k=1}^{K} g_{mk}^{-1} \sum_{l=1}^{L} N_{kl}\, \Delta U\, \exp[f_l - \beta_l U_m]} \end{aligned} \tag{2.65}$$

We can rewrite Eq. 2.64 as

$$\hat{\Omega}_m = \frac{H_m^{\text{eff}}}{\sum_{l=1}^{L} N_{ml}^{\text{eff}}\, \Delta U\, \exp[f_l - \beta_l U_m]} \tag{2.66}$$

where $H_m^{\text{eff}} \equiv \sum_{k=1}^{K} g_{mk}^{-1} H_{mk}$ is the effective number of independent samples in energy bin m from all replicas, and $N_{ml}^{\text{eff}} \equiv \sum_{k=1}^{K} g_{mk}^{-1} N_{kl}$ is an effective number of independent samples at temperature β_l from all replicas.

To compute the estimator of the expectation for an observable A, we apply the same technique in Section 2.2.6 above and write the expectation as a weighted sum over configurations

$$\hat{A}(\beta) = \frac{\sum_{k=1}^{K}\sum_{n=1}^{N_k} w_{kn}(\beta)\, A_{kn}}{\sum_{k=1}^{K}\sum_{n=1}^{N_k} w_{kn}(\beta)} \tag{2.67}$$

where the weights are given by

$$w_{kn}(\beta) \equiv \sum_{m=1}^{M} \psi_{mkn}\, H_m^{-1}\, \hat{\Omega}_m e^{-\beta U_m}. \tag{2.68}$$

As in Eq. 2.38 and 2.59, the sum over m reduces to a single term, the one with energy bin index appropriate for configuration n of replica k. A_{kn} is the value of the observable A for configuration n of replica k and $H_m = \sum_{k=1}^{K}\sum_{n=1}^{N} \psi_{mkn}$, the total number of configurations from all replicas with potential energy in bin m.

Again, if the observable of interest has a correlation time that is long compared to fluctuations in the potential energy, we may compute the dominant contribution to the statistical uncertainty $\delta^2 A(\beta)$ by Eqs. 2.41 – 2.48, with the important distinction that k now indexes the *replicas*, rather than the temperatures. The correlation times are, as in the simulated tempering case, computed over the nonphysical replica trajectories; because the replicas perform random walks in temperature, these times are likely to be shorter than the correlation time for this observable computed from a canonical simulation at the lowest temperature. These replica correlation times properly capture the correlation between successive snapshots generated by a sampling method like Metropolis Monte Carlo or molecular dynamics, and their use in estimating the uncertainty is the primary novel result of this paper. Collecting configurations from all replicas into pseudotrajectories of constant temperature, as suggested in previous attempts to apply the method to parallel tempering simulations [171], would give correlation times that are erroneously short and make the incorrect assumption that these pseudotrajectories are statistically independent.

2.4 Applications

2.4.1 One-Dimensional Model Potential.

To validate the methods described above for estimating expectations and corresponding uncertainties, we consider a one-dimensional model potential where canonical expectations can be

computed directly and a large quantity of simulation data can be obtained in order to verify our uncertainty formulae. We use an asymmetric double well potential, given by

$$U(q) \quad = \quad (q-1)^2(q+1)^2 + 0.1\,q. \tag{2.69}$$

All simulations utilize the Metropolis Monte Carlo method [119] with the trial displacement Δq uniformly distributed on the interval $[-0.2, +0.2]$ to generate a series of configurations which are sampled every 10 move attempts, resulting in highly correlated data. In the following simulations, we estimate the expectation $\langle q \rangle_{\beta^*}$ at $\beta^* = 4$, where the integrated correlation time of q is rather long — approximately 130 samples. The initial conformation was chosen uniformly on the interval $[-1.8, +1.8]$ and the first 10^5 steps discarded to equilibration.

Four types of simulations were performed: a standard canonical Metropolis Monte Carlo (MMC) simulation at $\beta = \beta^* = 4$, as described above; a set of four independent canonical (4MMC) simulations with inverse temperatures β exponentially spaced in the range 1–4 ($\beta \approx \{4, 2.52, 1.59, 1\}$); a simulated tempering simulation (ST) with the same four possible temperatures and analytically-computed optimal weights; and a parallel tempering (PT) simulation with replicas at the same four temperatures.

All simulations were conducted for 5×10^7 steps each (per replica, if multiple replicas are used), generating 5×10^6 samples (per replica). The data were then divided into 500 sequential blocks of 10^4 configurations (per replica) each, whose expectations were verified to be statistically independent by computing the correlation between expectations in neighboring blocks. The standard deviation of the set of expectations computed from each block is indicative of the statistical uncertainty in simulations of a single block length — 10^4 samples (per replica) — and the difference between the mean of these estimates and the expectation computed from the potential directly is indicative of the bias. Expectations and uncertainties for each block were computed using the code appearing in Listing 1 of the Supplementary Material.

To assess the performance of the uncertainty estimate for each block, we compute the fraction of blocks for which the true magnitude of the deviation from the mean of the block expectations is smaller than a multiplicative constant σ times the estimated uncertainties, for $\sigma \in [0.1, 3]$. This fraction is related to a confidence interval if compared to the error function Gaussian integral (Figure 2.1). For example, for our uncertainty estimates to be meaningful, we expect the difference between the true mean and our estimate to be within one standard deviation σ approximately 66% of the time. It is readily apparent from the figure that the computed uncertainty estimate computed for each block is in fact quite good. Additionally, the bias is small — less than 10% of the

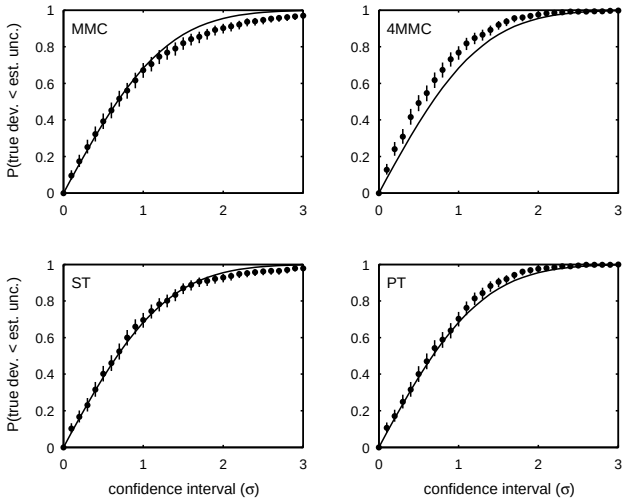


Figure 2.1: **Confidence curves for Metropolis Monte Carlo simulations on the 1D model potential.** The fraction of statistically independent blocks for which the true uncertainty (the deviation of the estimated expectation over the block from the mean of the block estimates) is less than a multiplier of the predicted 1σ uncertainty (here plotted as the independent variable). The solid curve shows the fraction expected to fall within the interval for the normal distribution. Ideally, the curves would coincide. The results are shown for (MMC) a single Metropolis Monte Carlo simulation at $\beta = 4$; (4MMC) a set of four independent canonical simulations spanning the range $\beta = 1 - 4$; (ST) a simulated tempering simulation spanning $\beta = 1 - 4$; (PT) a parallel tempering simulation with four replicas spanning $\beta = 1 - 4$. Uncertainties, with 95% confidence intervals shown here as vertical bars, were computed as described in Appendix 2.9.

magnitude of the statistical uncertainty in the cases studied (data not shown).

2.4.2 Alanine Dipeptide in Implicit and Explicit Solvent.

To illustrate the utility and verify the correctness of the procedures described above for simulations of biological interest, we demonstrate their use in the analysis of parallel tempering simulations of alanine dipeptide in implicit and explicit solvent. A similar strategy to the 1D model system described above was adopted, with a long simulation partitioned into short blocks (here, 2 ns/replica per block) whose expectations were verified to be statistically independent by the same procedure described above.

Using the `LEaP` program from the AMBER7 molecular mechanics package [18], a terminally-blocked alanine peptide (sequence ACE-ALA-NME, see Figure 2.2) was generated in the extended conformation. For the explicit solvent system, the peptide was solvated with 431

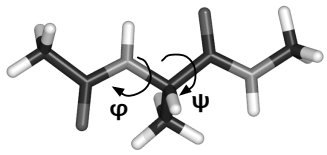


Figure 2.2: **Terminally-blocked alanine peptide with (ϕ, ψ) torsions labeled.**

TIP3P water molecules [90] in a truncated octahedral simulation box whose dimensions were chosen to ensure a minimum distance to the box boundaries from the initial extended peptide configuration of 7Å. Peptide force field parameters were taken from the `parm96` parameter set [95]. For the implicit solvent simulation, the Generalized Born method of Tsui and Case [183] (corresponding to the flag `igb=1`) was employed with radii from AMBER6, along with a surface area penalty term of the default 5 cal mol^{-1} Å^{-2}. Covalent bonds to hydrogen were constrained with SHAKE using a tolerance of 10^{-8} Å [148]. Long-range electrostatics for the explicit solvent simulation were treated by the particle-mesh Ewald (PME) method [34] with default settings. Each system was first subjected to 50 steps of steepest descent energy minimization, followed by 1000 steps of conjugate gradient optimization. To equilibrate the explicit solvent system to the appropriate volume, a 100 ps molecular dynamics simulation was performed with the temperature adjusted to 300 K and the pressure to 1 atm by the Berendsen weak-coupling algorithm [10] with temperature and pressure relaxation time constants of 1 ps and 0.2 ps, respectively. The simulation box was fixed at the final size obtained from this equilibration step, with a volume of 13 232 Å^3, in all subsequent simulations.

A parallel tempering (or replica-exchange among temperatures) molecular dynamics simulation [171] was conducted using a parallel `Perl` wrapper for the `sander` program[3]. Replica temperatures were exponentially distributed over the range 273-600K, with 10 replicas required for the implicit solvent simulation (yielding an exchange acceptance probability between neighboring temperatures of approximately 75%) and 40 replicas for the explicit solvent simulation (yielding an acceptance probability of approximately 50%). All momenta were reassigned from the Maxwell-Boltzmann distribution at the appropriate replica temperature after

[3] A copy of this Perl wrapper to perform replica-exchange simulations using AMBER7 and AMBER8 can be obtained from `http://www.dillgroup.ucsf.edu/~jchodera/code/rex`.

each exchange attempt. Between exchanges, constant-energy, constant-volume molecular dynamics was carried out for the explicit solvent simulation, while the implicit solvent simulation utilized Langevin dynamics with a friction coefficient of 95 ps^{-1} to mimic the viscosity of water. All dynamics utilized a 2 fs timestep. The algorithm used to select pairs of replicas for temperature exchange attempts starts from the highest-temperature replica and attempts to swap the configuration for the next-lowest temperature replica using a Metropolis-like criteria, and proceeds down the temperatures in this manner. On the next iteration, swapping attempts start from the lowest temperature and proceed upward, and this alternation in direction is continued in subsequent pairs of iterations.

Starting all replicas from the minimized or volume-equilibrated configuration described above, 100 iterations were conducted with 1 ps between exchange attempts to equilibrate the replicas to their respective temperatures. This equilibration run was followed by a production run with 20 ps between exchange attempts, giving a total of 100 ns/replica for the implicit solvent production run and 20 ns/replica for the explicit solvent run. Solute configurations and potential energies were saved from the production run every 1 ps. Expectations and uncertainties were again estimated using Listing 1 appearing in the Supplementary Material.

Over 2 ns blocks of simulation time (containing 2000 configurations/replica in each block), we computed the probability of the peptide occupying the α_R conformation at 300K, with α_R here defined as $-105 \leq \phi < 0$ and $-124 \leq \psi < 28$. This corresponds to configurations that would be classified as right-handed alpha helical. To validate the uncertainty estimates, confidence curves of the type description in Section 2.4.1 were computed and are shown in Figure 2.4. Though the confidence intervals are larger because the data contain fewer independent blocks, the uncertainty estimates are still good indicators of the expected deviation from the true expectation.

The potential of mean force (PMF) for the ψ torsion angle at 300K was also computed, and is shown in Figure 2.3. The computed PMF and uncertainty for a representative block is depicted in the top panel, along with the PMF computed using the entire trajectory. The deviations of the block PMFs from the whole-simulation estimate fall within the 1σ uncertainty bars to the expected degree. In the lower panel, the uncertainties computed from the representative block are compared to the standard deviation of the PMF computed from all blocks, which should be indicative (to within an order of magnitude) of the uncertainty expected from a simulation of the block length. These too compare favorably.

It is important to note that our neglect of the uncertainty in the dimensionless free energies, $\{f_l\}$, is only reasonable if the correlation time of the observable of interest is much longer than that of the

potential energy. When this condition is satisfied, the dominant contribution to the uncertainty in the computed expectation of the observable is due to the small number of effectively independent samples of this observable present in the simulation data. To demonstrate that this is the case for systems of interest, we have assessed the relative contribution of the neglected uncertainty in the $\{f_l\}$ to the uncertainty of the estimated probability of the α_R conformation of the alanine dipeptide system considered here. The resulting contribution is 10 times smaller than the uncertainty due to the time correlation treated above for the explicit solvent system and 100 times smaller for the implicit solvent system. [footnote{The impact of the uncertainty in the $\{f_l\}$ on the uncertainty in the estimated observable was computed in the following manner: We first computed estimates of the $\{f_l\}$ over all uncorrelated 2 ns/replica blocks of simulation data to form a pool of dimensionless free energies that represent the typical uncertainty in a simulation of this length. Next, for each 2 ns/replica block, we computed the standard deviation in the estimated α_R probability when all $\{f_l\}$ in this pool were substituted into the WHAM equations. The mean of this standard deviation over all blocks then provides an estimate of the magnitude of the impact of typical uncertainties in the $\{f_l\}$ on the observable of interest.}] However, if the observable has a correlation time comparable to that of the potential energy (e.g., if the expectation of the potential energy itself is of interest) then the uncertainty due to imperfect knowledge of the $\{f_l\}$ can be comparable to the uncertainty due to the correlation in the observable. In these cases where correlation times are comparable, an algorithm that combines our approach with the T-WHAM method of Gallicchio, et al. [68], which explicitly treats the uncertainty in the $\{f_l\}$ when the potential energy samples are uncorrelated, may provide a superior estimate of the uncertainty in the estimate of the observable.
We further note that pathological cases may arise where simulations at neighboring temperatures may have poor energy overlap, resulting in large uncertainties in some of the $\{f_l\}$. Fortunately, these cases are easily detected by examination of the exchange acceptance rates between neighboring temperatures, where they will be conspicuously low, and detectable early in the simulation. Such cases are easily remedied by adjusting the temperature spacing or by the addition of more replicas at intermediate temperatures.

2.5 Practical Considerations

Several issues of great importance to successful implementation of the algorithm have received little discussion in the literature.

2.5.1 Choice of Bin Width and Number of Bins.

There is a bias-variance tradeoff in the choice of energy histogram width. As the energy bin width increases, the uncertainty in our histogram estimator for $p_m(\beta)$, the probability density for energy bin m, decreases. At the same time, one expects the resulting estimate of the density of states Ω_m to become increasingly biased, especially considering the dependence of $p(U)$ on the rapidly-varying exponential Boltzmann factor $e^{-\beta U}$. Because of this, a reasonable assumption might be that the bin width ΔU should be chosen such that $\Delta U \ll k_B T$. However, if the bin size is too small, the uncertainty in the estimate for the $p_m(\beta)$ will be large. One possibility might be to use a data-based choice of histogram bin width, as in Wand [192], which uses concepts from non-parametric density estimation in attempting to minimize the mean integrated square error (MISE) to the true probability density.

For the alanine dipeptide simulations described in Section 2.4.2 above, we find that the estimated probability of occupying the α_{R} region of conformation space is largely insensitive to the number of bins used to discretize the sampled potential energy range. In fact, the variation in the computed expectation is well within the statistical uncertainty over the range of 50 to 5000 bins (corresponding to a range of bin widths of 0.5 $k_B T$ to 50 $k_B T$).

2.5.2 Computing Integrated Correlation Times.

Estimating the correlation time τ, defined above in Eqs. 2.19 and 2.21, can be difficult when one is confronted with noisy correlation functions. While ensuring trajectories are many times longer than the longest correlation times is necessary for an accurate estimate, even if this is achieved, performing the straightforward sum over the entirety of the correlation function C_t as in Eq. 2.19 is almost always a poor choice, as the uncertainty in the computed correlation function grows approximately linearly with the lag time t [204]. Even for trajectories many times longer than the correlation length, this sum will be dominated by contributions from the noisy tail, likely resulting in large errors or even negative values for the computed correlation time τ. Janke proposes a self-consistent approach where the summation is performed only out to lag times of 6τ, after which the correlation function is assumed to be negligible [87]. Evertz contends that this approach produces incorrect results [53], instead proposing an exponential fit to the tail of the correlation function and use of this fit to evalute the summand when the correlation function is dominated by noise. Neither solution is both stable and straightforward to apply, so we instead truncate the sum when the normalized fluctuation correlation function C_t first crosses zero, since it is likely

unphysical for the correlation function to be negative for most observables[4]. The zero crossing is an indication that the statistical uncertainty dominates the signal, and that the remainder of the correlation function should be considered indistinguishable from zero.

For most systems and observables, the correlation function will decay rapidly at first and then slowly, approximately exponentially for large t. To avoid the expense of computing C_t at each value of t while still obtaining reasonably accurate integrated correlation times for observables with very different decay timescales, we use an adaptive integration scheme in which the correlation function C_t is computed only at times $t_i = 1 + i(i-1)/2$, where $i = 1, 2, 3, \ldots$. In computing the correlation time τ, the sum in Eq. 2.19 is now performed only over the t_i terms, with each term weighted by $t_{i+1} - t_i$, with $t_1 = 1$. This approach ensures high time resolution at small t when C_t is likely to be rapidly changing but avoids the expense of computing C_t at every t in the slowly-decaying tails. We find the accuracy of this approach to be acceptable — differences typically amount to at most ten percent.

2.5.3 Neglect of Bin Statistical Inefficiencies g_{kn}.

It is often assumed or stated without justification that the energy bin statistical inefficiencies g_{mk} appearing in Eqs. 2.30 and 2.64, representing the number of snapshots required for a statistically independent sampling of the energy bin, are all equal or equal to unity [100, 134, 171]. All g_{mk} will be equal to unity only if the $\{\psi_{mkn}\}_{n=1}^{N_k}$ are uncorrelated. To test this assumption, we have computed the statistical inefficiencies for the systems mentioned in Section 2.4 above. To our knowledge, this is the first time a test of this claim has been reported in the literature. Indeed, for the explicit solvent system studied in Section 2.4.2 above, we find large differences in the statistical inefficiencies for the same replica but different energy bins, sometimes differing up to two orders of magnitude. Similarly, for the same energy bin, the statistical inefficiencies from different replicas can differ by up to two orders of magnitude.

In the limit that our parallel tempering simulation is very long — long enough for each replica to execute an unrestricted random walk through all temperatures and explore all relevant regions of configuration space — each replica can be considered to be equivalent. In this case, the statistical inefficiencies should be independent of replica index k, and we can write g_m as the statistical

[4] Velocity correlation functions, where there is often a clear negative peak at short times, are an obvious exception.

inefficiency for energy bin m. Applied to Eqs. 2.64 and 2.65, this yields a new set of expressions:

$$\hat{\Omega}_m = \frac{\sum_{k=1}^{K} H_{mk}}{\sum_{k=1}^{K}\sum_{l=1}^{L} N_{kl}\,\Delta U\,\exp[f_l - \beta_l U_m]} \tag{2.70}$$

$$\delta^2\hat{\Omega}_m = \frac{\hat{\Omega}_m}{g_m^{-1}\sum_{k=1}^{K}\sum_{l=1}^{L} N_{kl}\,\Delta U\,\exp[f_l - \beta_l U_m]} \tag{2.71}$$

The quantity $\sum_{k=1}^{K} N_{kl}$, simply represents the total number of configurations stored from any replica at temperature β_l. For a parallel tempering simulation, where each temperature must be populated by exactly one replica at all times, this is simply N, the total number of configurations stored per replica. Additionally, $H_m \equiv \sum_{k=1}^{K} H_{mk}$ is the total number of configurations over all replicas (and hence temperatures) with energy in bin m. This gives

$$\hat{\Omega}_m = \frac{H_m}{N\sum_{l=1}^{L}\exp[f_l - \beta_l U_m]} \tag{2.72}$$

which is identical to the WHAM result for independent canonical simulations for the case where all g_{mk} are identical or unity. If there is no correlation in the data — that is, all configurations are independent — it does not matter whether we apply this analysis to the original data or collect up the configurations by temperature and apply WHAM equations for independent canonical simulations. This expression is in fact identical to the one used in many published works that have previously attempted to use the weighted histogram analysis method for the analysis of parallel tempering simulations, such as [134, 171].

Under these same assumptions, the correlation functions for any observable A should also be identical for each replica. One can therefore average estimates of the unnormalized correlation functions $\langle A_n A_{n+t}\rangle$ over the replicas and use optimal estimates of the mean and variance computed over all of the replicas to obtain an optimal estimate of the statistical inefficiencies and uncertainties. An implementation illustrating this procedure is provided in the Supplementary Material as Listing 2.

Note, however, that the above assumptions of equivalence cannot be made in cases where the replicas are clearly *inequivalent*, such as in a simulated tempering replica-exchange (STREM) simulation [120, 121]. In that case, the expression above will only be recovered if the time between samples is so long that all the g_{mk} are unity.

2.5.4 The Statistical Inefficiency for the Cross-Correlation Term, $g_{k,wA;w}$.

In computing the statistical inefficiency for the cross-correlation term, uncertainties in the computed integrated autocorrelation times due to insufficient data or approximations may cause the cross-correlation term to dominate and the estimate of the square uncertainty of $\hat{A}$ (eq 2.41) to be negative. Clearly, this should not be allowed to occur, as the squared-uncertainty should be a strictly positive quantity.
The statistical inefficiencies g should obey the following relation (derived in Appendix 2.8):

$$g_{x;y} \quad \leq \quad (g_x \, g_y)^{1/2} \, \frac{\sigma_x \sigma_y}{|\sigma^2_{x;y}|} \tag{2.73}$$

This is often violated when the correlation function is noisy, and can lead to negative estimates of the squared uncertainty when the cross-correlation term dominates. In these cases, we find it best to simply limit $g_{x;y}$ to its maximum allowed value computed from the right-hand side of Eq. 2.73. Since the autocorrelation times are usually shorter than the cross-correlation time, it is believed that these estimates will be better than the integrated cross-correlation time.

2.6 Conclusion

We have presented an extension of the weighted histogram analysis method (WHAM) for the analysis of one or more independent simulated tempering or parallel tempering simulations. The method provides not only estimators of canonical expectations but also estimators for the statistical uncertainties in the resulting estimates. We hope that, with the availability of the provided example code, workers using these simulation techniques will provide uncertainty estimates so that the statistical significance of results obtained from them can be assessed. We have shown that the estimator for the expectation has small bias and produces excellent uncertainty estimates for both a 1D model system and a solvated biomolecular system in implcit and explicit solvent.
While other workers had attempted to apply WHAM to simulated or parallel tempering data in the past [120, 171], the key advance here is the consideration of the correlated nature of the configurations sampled by each replica as it performs a pseudorandom walk in temperature, allowing a proper assessment of the true number of independent samples present in the data. This produces correct optimal estimators and makes possible the estimation of statistical uncertainties. This method can be extended to the analysis of other generalized-ensemble simulations, such as the multicanonical method (MUCA) [12, 13, 55, 78, 128], by consideration of replica correlation

times as the system samples various energy levels biased by the estimate of the density of states. Still, it is important to point out that, while we consider the contribution from time-correlation of the observable to the uncertainty estimate, we currently neglect the contribution of the uncertainty in the per-configuration weights (which originates from the uncertainty in the density of states) to the estimate of the expectation — we assume it is negligible and await a more complete treatment of the uncertainty in cases where it is not.

2.7 Acknowledgments

JDC was supported by an Howard Hughes Medical Institute and an IBM predoctoral fellowship. WS acknowledges support from NSF MRSEC Center on Polymer Interfaces and Macromolecular Assemblies DMR – 0213618, and KD the support of NIH grant GM34993. The authors would like to thank Libusha Kelly, Bosco K. Ho, and M. Scott Shell (University of California, San Francisco) for critical reading of this manuscript, as well as the anonymous referee who raised an excellent point regarding our neglect of the uncertainty in the dimensionless free energies. This manuscript was strengthened as a result of their input.

2.8 Relation for Statistical Inefficiencies

Consider a random process where we make a series of N time-correlated measurements of two (possibly correlated) observables X and Y, resulting in the time series $\{x_n, y_n\}_{n=1}^{N}$. We estimate the quantity $Z = \langle X \rangle / \langle Y \rangle$ from our sample means, and wish to compute the uncertainty in our estimate, defined as

$$\delta^2 Z \equiv \sigma_Z^2 = \langle (Z - \langle Z \rangle)^2 \rangle = \langle Z^2 \rangle - \langle Z \rangle^2. \tag{2.74}$$

To first order about $\langle X \rangle / \langle Y \rangle$, this uncertainty is given by

$$\sigma_Z^2 = \left[\frac{\langle X \rangle}{\langle Y \rangle} \right]^2 \left[\frac{\sigma_X^2}{\langle X \rangle^2} + \frac{\sigma_Y^2}{\langle Y \rangle^2} - 2 \frac{\sigma_{X;Y}^2}{\langle X \rangle \langle Y \rangle} \right] \tag{2.75}$$

where $\sigma_{X;Y}^2$ denotes the (not necessarily positive) covariance of the expectations of X and Y. The Schwartz inequality requires that this covariance obey the relation

$$|\sigma_{X;Y}^2| \leq \sigma_X \sigma_Y \tag{2.76}$$

(see, for example, [19]). Given this, we note

$$\left| \frac{\sigma^2_{X;Y}}{\sigma_X \sigma_Y} \right| \quad \leq \quad 1 \tag{2.77}$$

Correlation analysis gives estimators for these quantities as

$$\begin{aligned} \sigma^2_X &= \sigma^2_x/(g_x^{-1}N) \\ \sigma^2_Y &= \sigma^2_y/(g_y^{-1}N) \\ \sigma^2_{X;Y} &= \sigma^2_{x;y}/(g_{x;y}^{-1}N) \end{aligned} \tag{2.78}$$

where σ^2_x denotes the sample variance of the observations $\{x_n\}_{n=1}^N$ and g denotes the statistical inefficiency obtained from the autocorrelation time, *i.e.*

$$g_x = 1 + 2\tau_x \geq 1. \tag{2.79}$$

Combining Eqs. 2.77 and 2.78 gives

$$\frac{|\sigma^2_{x;y}|}{\sigma_x\, \sigma_y} \cdot \frac{g_{x;y}}{(g_x\, g_y)^{1/2}} \quad \leq \quad 1 \tag{2.80}$$

where we have moved the statistical inefficiencies and variances out of the absolute value, as they are always positive. Finally, we obtain an upper bound for $g_{x;y}$:

$$g_{x;y} \quad \leq \quad (g_x\, g_y)^{1/2}\, \frac{\sigma_x \sigma_y}{|\sigma^2_{x;y}|} \tag{2.81}$$

In using numerical methods to estimate the statistical inefficiencies g from finite trajectories, this inequality may not hold, sometimes leading to negative squared uncertainties. Limiting the estimated $g_{x;y}$ by capping it at this value will prevent this from occurring.

2.9 Uncertainty Estimates for Confidence Curves

To estimate the uncertainties in Figures 2.1 and 2.4, a Bayesian inference scheme was used. Since the expectations computed from each block are independent, the number of blocks n that fall inside the given scaled deviation σ is described by a binomial distribution with parameter θ, true (unknown) probability that the blocks fall within the given deviation σ:

$$P(n|\theta) \quad = \quad \frac{N!}{n!(N-n)!}\theta^n(1-\theta)^{N-n} \tag{2.82}$$

We can write the posterior distribution for the probability p given the observed number of blocks within the given deviation n using Bayes' rule

$$p(\theta|n) \quad \propto \quad p(n|\theta)\, p(\theta) \tag{2.83}$$

where $p(\theta)$ is the prior distribution for the parameter θ. If we choose the prior $p(\theta)$ to be a Beta distribution with hyperparameters α and β, given by

$$p(\theta|\alpha,\beta) \quad = \quad [B(\alpha,\beta)]^{-1} x^{\alpha-1}(1-x)^{\beta-1} \tag{2.84}$$

where $B(\alpha,\beta)$ is the beta function, then the posterior $p(\theta|n)$ will also be a Beta distribution with parameters $n+\alpha$ and $(N-n)+\beta$, as the Beta distribution is a conjugate prior to the Binomial distribution. We take the hyperparameters α and β to be unity to make the prior distribution uniform, resulting in a posterior $\theta \sim \mathrm{Beta}(n+1, N-n+1)$. A 95% central confidence interval, corresponding to the location where the cumulative distribution function for the Beta distribution reaches the values of 0.025 and 0.0975, was plotted.

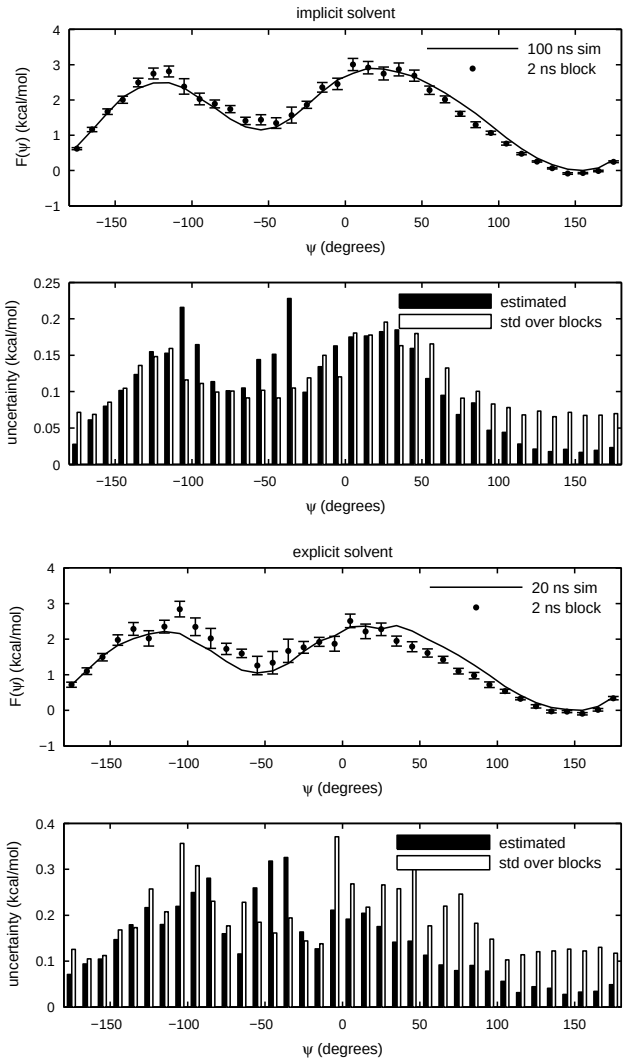


Figure 2.3: **Potential of mean force in ψ for implicit and explicit solvent parallel tempering simulations.** Left: implicit solvent; right: explicit solvent. Upper panels: The potential of mean force in the ψ torsion angle at 300 K. The solid line shows the PMF estimated from the entire simulation, while the filled circles show the estimated PMF uncertainty using the method described in the text for a single 2 ns/replica block. Lower panels: The computed uncertainties for the same 2 ns block (left bars) along with the average uncertainty expected for a simulation 2 ns/replica in length, estimated from the standard deviation of the PMFs computed from all nonoverlapping blocks of length 2 ns in the full simulation. All uncertainties are shown as one standard deviation.

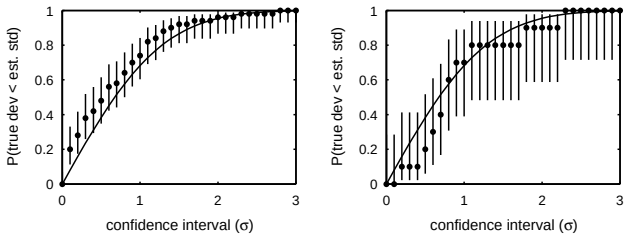


Figure 2.4: **Confidence curves for implicit and explicit solvent parallel tempering simulations.** As in Figure 2.1, the fraction of statistically independent 2 ns blocks for which the true uncertainty is less than a multiplier of the predicted 1σ uncertainty is shown. The observable used is an indicator function for the α_{R} configuration. Left: implicit solvent (statistics over 50 blocks); right: explicit solvent (statistics over 10 blocks).

Chapter 3

A master equation can describe peptide dynamics

The material in this chapter was submitted to *Multiscale Modeling and Simulation*, and has been accepted to appear in a Special Issue on Biological Modeling as

Long-time protein folding dynamics from short-time molecular dynamics simulations

John D. Chodera[†], William C. Swope[‡], Jed W. Pitera[‡], and Ken A. Dill[§]
[†] *Graduate Group in Biophysics and* [§] *Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, CA 94143*
[‡] *IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120*

Abstract

Protein folding involves physical timescales — microseconds to seconds — that are too long to be studied directly by straightforward molecular dynamics simulation, where the fundamental timestep is constrained to femtoseconds. Here, we show how the long-time statistical dynamics of a simple solvated biomolecular system can be well-described by a discrete-state Markov chain model constructed from trajectories that are an order of magnitude shorter than the longest relaxation times of the system. This suggests that such models, appropriately constructed from short molecular dynamics simulations, may have utility in the study of long-time conformational dynamics.

3.1 Introduction

Proteins can fold to well-defined native topologies[1] with surprising determinism. Many small, single domain proteins can fold rapidly, reversibly, cooperatively, and without the aid of other molecular machinery. In response to an environmental perturbation such as the introduction or removal of denaturant or a rapid change in solvent temperature, these fast-folding proteins exhibit nearly exponential relaxation kinetics with observed time constants on the order of microseconds. Other proteins exhibit slow and complex kinetics, suggesting the presence of one or more kinetic intermediates. A detailed understanding of this process has been the focus of much of modern biophysics. Ultimately, knowledge of the general mechanistic features by which proteins fold and aggregate is critical for understanding a variety of folding and misfolding diseases, elucidating principles necessary for effective protein design, and developing the basic tools needed for other related technological applications of complex molecular structures.

A description of the mechanism by which a particular protein folds must by necessity be a statistical one. While the initial microscopic state[2] and dynamical trajectory may differ for each molecule in an experiment, many proteins refold to their native (folded) topologies upon the restoration of native conditions with the certainty of macroscopic law [3]. A proper statistical description would summarize the salient features and relative probabilities of relevant folding routes in a way that is meaningful to the physical chemist. This manner of model has been difficult to extract from experiments. Despite the high time resolution possible with optical spectroscopy, the majority of these experiments rely on the observation of an *ensemble* of molecules to obtain sufficient signal, resulting in the ability to observe only (possibly time-dependent) ensemble averages, rather than the behavior of any single molecule. While observations of single molecules are now possible with fluorescence techniques, atomic force microscopy, or optical traps, high temporal resolution is sacrificed to achieve sufficient signal for reliable measurement. In contrast, computer simulation promises the ability to produce information with both atomic detail and high temporal resolution.

In practice, however, the presence of fast vibrational motion constrains the fundamental integration timestep to femtoseconds in order to ensure stability, limiting practical straightforward molecular dynamics simulations of atomically-detailed representations of solvated proteins to tens of

[1] By *native topology* or *native structure*, we refer to the *ensemble* of configurations sharing a coarse overall structure, or fold, with the experimental structure.

[2] By *microscopic state*, we refer to the set of generalized coordinates and momenta that completely determine the microscopic state of the system, such as the phase space point.

nanoseconds. As even the fastest folding proteins exhibit relaxation timescales of several microseconds [99], this leads to a *timescale gap* of at least two decades in time. Using supercomputers such as Blue Gene [69] and software specialized for molecular dynamics simulations on these computer systems [62, 73], one can produce atomistic simulations of protein molecules with explicit representation of surrounding solvent on several microsecond timescales. However, the number of trajectories that need be generated to provide an adequate *statistical* characterization of the folding mechanism of even a single protein makes such an endeavor extremely challenging. Distributed computing projects like Folding@Home [140] regularly collect tens of thousands of trajectories tens of nanoseconds in length, but extracting insight about microsecond timescale dynamics from these large datasets can be difficult [59, 116, 138].

Kinetics models may provide the necessary link between short simulations of a single molecule and long experimental observations of ensembles of molecules. If time evolution of a protein system is characterized by long waiting times within metastable states punctuated by infrequent transitions between these states, interstate dynamics may appear stochastic and memoryless on some short timescale. In this case, long trajectories may be modeled as a Markov chain realized on a discrete state space of a (hopefully small) number of states. While this model could not describe dynamical behavior at very short timescales, which is dominated by molecular motion *within* a metastable state, it could nevertheless faithfully describe long-timescale transitions *between* states. This approach would have numerous advantages. It is precisely these slow transitions involving major structural rearrangements that are of primary interest; elimination of high-frequency detail is often desirable in aiding interpretation of trajectories. To generate a statistical description of folding dynamics, instead of generating many simulations each long enough to contain complete folding events, we need only generate simulations long enough to characterize transition rates between pairs of conformational substates. Construction would therefore be amenable to parallelization on loosely-coupled grids of computer systems. The resulting kinetic model could then be used to compute the stochastic temporal evolution of either a single molecule or an ensemble of molecules, allowing direct comparison to data from both kinds of kinetics experiments, or to answer statistical questions about folding pathways and mechanism that are currently experimentally inaccessible.

This proposition is not entirely novel. Several groups have constructed stochastic kinetic models from states defined by local potential energy minima of small peptides, using transition state theory to estimate interstate transition rates [8, 101, 105, 124, 125]. Unfortunately, the number of minima grows rapidly with increasing system size, making the procedure prohibitively expensive for larger proteins or systems containing explicit solvent molecules. Other work [2, 35, 151, 162, 167, 175]

has focused on the construction of discrete- or continuous-time Markovian models to describe dynamics between a small number of states. These models, however, have yet to demonstrate that they can adequately describe the dynamics on timescales much longer than the trajectories from which the models were constructed.

Here, we present a proof of principle for how the dynamics of a solvated biomolecular system can be described using information from short simulations. This is illustrated using terminally-blocked alanine, a system small enough that its dynamical behavior can also be thoroughly characterized by straightforward molecular dynamics simulation. First, a parallel tempering simulation is conducted to explore the thermally relevant regions of configuration space. Next, a set of metastable states, corresponding to regions of configuration space with low probabilities of escape, are identified. Due to the simplicity of the system chosen in this work, these states can readily be identified by hand. This removes the complication of choice of state decomposition, which we shall not consider here. Finally, a number of short trajectories are initiated from each state (possibly in parallel) and a Markov chain model is constructed from the number of interstate transitions observed in these trajectories.

This paper is organized as follows: In Section 3.2, the Markov chain model and its method of construction are described. In Section 3.3, the method is applied to terminally-blocked alanine in explicit solvent. A discussion of the significance of this result, as well as problems remaining to be solved before the method can be applied to larger biomolecules, follows in Section 3.4.

3.2 Theory

3.2.1 Conformational dynamics as a Markov process.

Consider the dynamics of a macromolecular system in equilibrium at some temperature of interest, where we have decomposed all of configuration space into a set of M disjoint states. If we observe a trajectory of this system at times $t = 0, \tau, 2\tau, \ldots, n\tau$, where τ denotes the observation interval, we can represent the trajectory in terms of the state the system occupies at each of these discrete times, $s_0, s_1, s_2, \ldots, s_n$. The sequence of states produced by such a trajectory is a *discrete-time stochastic process*. If this process is a Markov chain, it must satisfy the *Markov property*, whereby the probability of observing the system in state s_n at the next time point, $n\tau$, given the state history $s_0, s_1, s_2, \ldots, s_{n-1}$ is independent of all but the current state s_{n-1}. For a stationary process which

has no explicit dependence on time, this property is given by

$$P(s_n|s_0, s_1, s_2, \ldots, s_{n-1}) \;=\; P(s_n|s_{n-1}). \tag{3.1}$$

As there are a finite number of states, this process can be entirely characterized by an $M \times M$ *transition matrix* $\boldsymbol{T}(\tau)$ dependent only on lag time τ. The element $T_{ji}(\tau)$ denotes the probability of observing the system in state j at time τ given that it was initially in state i at time 0:

$$T_{ji}(\tau) \;\equiv\; P(j|i). \tag{3.2}$$

If we do not know the precise initial state of the system at time 0 but only the probability the system started in each state, or if we observe an ensemble of many non-interacting systems in an experiment, we can instead consider the probability of finding one particular molecule in each state i at time $n\tau$ as components of the vector of state probabilities $\boldsymbol{p}(n\tau)$. If the initial probability vector is given by $\boldsymbol{p}(0)$, we can write the probability vector at some later time $n\tau$ as

$$\boldsymbol{p}(n\tau) \;=\; \boldsymbol{T}(n\tau)\boldsymbol{p}(0) \;=\; [\boldsymbol{T}(\tau)]^n\boldsymbol{p}(0). \tag{3.3}$$

This property is termed the *Chapman-Kolmogorov* equation.
Once the transition matrix is known, the entire statistical dynamics of realizations of this process, corresponding to trajectories of the macromolecule under equilibrium conditions, could be extracted from it. Macroscopically observable properties, such as the time evolution of spectroscopically observable quantities for a noninteracting ensemble of molecules, can be computed:

$$\begin{aligned} A(n\tau) &= \boldsymbol{a}^{\mathrm{T}}\boldsymbol{p}(n\tau) \\ &= \boldsymbol{a}^{\mathrm{T}}[\boldsymbol{T}(\tau)]^n\boldsymbol{p}(0) \end{aligned} \tag{3.4}$$

where $\boldsymbol{a}$ denotes the vector containing the phase averages of the observable A over each state. In addition, microscopic quantities such as state lifetimes [174], mean first-passage times [162], hidden intermediates [137], and P_{fold} values (transmission coefficients) [103] can be obtained.

3.2.2 Construction of the Markov chain model from simulation.

For a system in which the dynamics are Newtonian but the initial configurations come from a canonical distribution, Swope et al. [174] show that the transition probability $T_{ji}(\tau)$ can be written

as

$$\begin{aligned} T_{ji}(\tau) &\equiv \frac{\langle \chi_j(\boldsymbol{z}(\tau))\chi_i(\boldsymbol{z}(0))\rangle}{\langle \chi_i(\boldsymbol{z}(0))\rangle} \\ &= \frac{\int d\boldsymbol{z}(0)\, e^{-\beta H(\boldsymbol{z}(0))}\, \chi_j(\boldsymbol{z}(\tau))\, \chi_i(\boldsymbol{z}(0))}{\int d\boldsymbol{z}(0)\, e^{-\beta H(\boldsymbol{z}(0))}\, \chi_i(\boldsymbol{z}(0))} \\ &= \int d\boldsymbol{z}(0)\, p_i(\boldsymbol{z}(0))\, \chi_j(\boldsymbol{z}(\tau)) \end{aligned} \tag{3.5}$$

where $\boldsymbol{z}$ denotes a point in phase space, $\chi_i(\boldsymbol{z})$ denotes the indicator function for state i, $\beta = (k_B T)^{-1}$ is the inverse temperature, $H(\boldsymbol{z})$ is the Hamiltonian, and $p_i(\boldsymbol{z})$ denotes the canonical distribution restricted to state i:

$$p_i(\boldsymbol{z}) = \frac{e^{-\beta H(\boldsymbol{z})}\, \chi_i(\boldsymbol{z})}{\int d\boldsymbol{z}\, e^{-\beta H(\boldsymbol{z})}\, \chi_i(\boldsymbol{z})}. \tag{3.6}$$

This result simply states the obvious: the transition matrix element $T_{ji}(\tau)$ can be estimated in a straightforward (though potentially inefficient) manner by initiating a number of simulations from configurations selected from a canonical distribution within state i, evolving the dynamics for a time τ, and determining the fraction of trajectories that terminate in state j:

$$T_{ji}(\tau) \approx \frac{N_{ji}(\tau)}{\sum_{j'=1}^{M} N_{j'i}(\tau)}. \tag{3.7}$$

Here, $N_{ji}(\tau)$ denotes the number of trajectories initiated from state i that terminate in state j at time τ. This procedure corresponds to the method proposed earlier by Swope et al. in the special case that the *selection cells* from which sets of simulations are initiated are coincident with the states [174].

We do not expect dynamics of a macromolecule in solution to resemble a Markov process for all observation intervals τ, as ballistic motion dominates on very short times and sufficient time must be allowed for collisions with the solvent and decorrelation of the trajectory within a metastable state. Imperfect definitions of the metastable states may also lead to non-Markovian behavior on short times [174]. At sufficiently long intervals τ, however, we may see that dynamics resembles a Markov process. While it is impractical to test the condition of complete history independence (Eq. 3.1), we can simply check the (weaker) condition imposed by the Chapman-Kolmogorov equation: For transition matrices constructed for a given τ, we check whether Eq. 3.3 holds for several lag times $n = 2, 3, 4, \ldots$ to within statistical uncertainty. If so, the Markovian model can be assumed to be a reasonable model of dynamics.

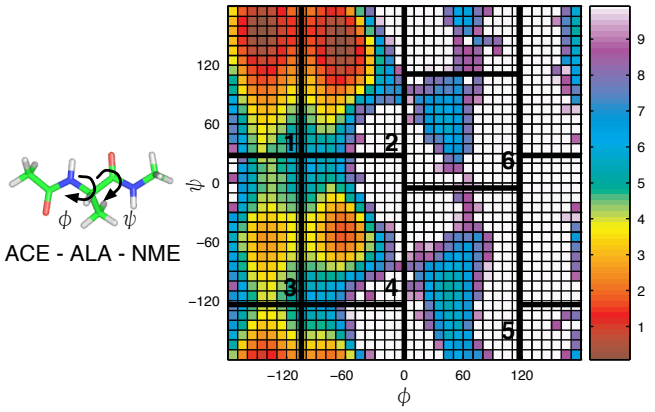


Figure 3.1: **Potential of mean force and state boundaries.** Left: The terminally-blocked alanine peptide with (ϕ, ψ) torsions labeled. Right: The potential of mean force in the (ϕ, ψ) torsions at 302 K estimated from the parallel tempering simulation, truncated at $10k_BT$ (white regions), with reference scale (far right) labeled in units of k_BT. Boundaries defining the six manually-identified states are superimposed and the states labeled.

3.3 Application to terminally-blocked alanine peptide

3.3.1 System setup and equilibration.

Using the `LEaP` program from the AMBER7 molecular mechanics package [18], a terminally-blocked alanine peptide (sequence ACE-ALA-NME, see Figure 3.1) was generated in the extended conformation, with peptide force field parameters taken from the AMBER `parm96` parameter set [95]. The system was subsequently solvated with 431 TIP3P water molecules [90] in a truncated octahedral simulation box with dimensions chosen to ensure all box boundaries were at least 7 Å from any atom of the extended peptide. All minimization and molecular dynamics simulations were conducted using the `sander` program from the AMBER7 package. Default nonbonded cutoffs were used, bonds to hydrogen were constrained with SHAKE using a tolerance of 10^{-8} [148], and long-range electrostatics were treated by the particle-mesh Ewald (PME) method [34] with the default settings.

The system was first subjected to 50 steps of steepest descent energy minimization, followed by 1000 steps of conjugate gradient optimization. To equilibrate the explicit solvent system to the appropriate volume, a 100 ps molecular dynamics simulation was performed with the temperature adjusted to 300 K and the pressure to 1 atm by the Berendsen weak-coupling algorithm [10] with temperature and pressure relaxation time constants of 1 ps and 0.2 ps, respectively. The simulation

Table 3.1: **State definitions for the manual decomposition of (ϕ, ψ)-space into metastable states and populations at 302 K.**

		state definitions		
state	label[a]	ϕ	ψ	P_{eq}[b]
1	C_5	$[117, -105)$	$[28, -124)$	.4787 (.0613)
2	P_{II}	$[-105, 0)$	$[28, -124)$	.4159 (.0486)
3	α_P	$[117, -105)$	$[-124, 28)$	.0425 (.0038)
4	α_R	$[-105, 0)$	$[-124, 28)$	.0588 (.0079)
5	C_7^{ax}	$[0, 117)$	$[111, -5)$	.0030 (.0015)
6	α_L	$[0, 117)$	$[-5, 111)$	.0011 (.0004)

[a] Corresponding state labels from [147]. [b] Equilibrium probabilities at 302 K estimated from the replica exchange simulation by WHAM, with corresponding uncertainties representing one standard deviation shown in parenthesis.

box was fixed at the final size obtained from this equilibration step, with a volume of 13 232 Å^3, in all subsequent simulations.

3.3.2 Parallel tempering.

In order to broadly explore the configuration space of the peptide and ensure that all important conformational substates were located, a parallel tempering (or replica-exchange among temperatures) molecular dynamics simulation [171] was conducted using a parallel `Perl` wrapper for the `sander` program[3]. Forty replicas were used, with replica temperatures exponentially distributed over the range 273–600 K, yielding an average exchange acceptance probability of about 50%. All momenta were reassigned from the Maxwell-Boltzmann distribution at the appropriate replica temperature after each exchange attempt, and constant-energy, constant-volume molecular dynamics with a 2 fs timestep was performed between exchange attempts. The algorithm used to select pairs of replicas for temperature exchange attempts starts from the highest-temperature replica and attempts to swap the configuration for the next-lowest temperature replica using the Metropolis-like criteria, and proceeds down the temperatures in this manner. On the next iteration, swapping attempts start from the lowest temperature and proceed upward, and this alternation in direction is continued in subsequent pairs of iterations.

Starting all replicas from the volume-equilibrated configuration described above, 100 iterations were conducted with 1 ps between exchange attempts to equilibrate the replicas to their respective temperatures. This equilibration run was followed by a production run of 500 iterations with 20 ps

[3] A copy of this Perl wrapper to perform replica-exchange simulations using AMBER7 and AMBER8 can be obtained from `http://www.dillgroup.ucsf.edu/~jchodera/code/rex`.

between exchange attempts, a total of 10 ns/replica. Solute configurations and potential energies from the production run were written to disk every 0.1 ps, while full-system restart files were recorded every 1 ps for the purpose of starting new simulations from these configurations, as described in Section 3.3.4.

3.3.3 State decomposition.

The slow degrees of freedom for terminally-blocked alanine peptide (neglecting those involving solvent motion) can be captured by the two backbone torsion angles labeled ϕ and ψ (see Figure 3.1) [17, 111]. To this end, the potential of mean force at 302 K was computed from the parallel tempering data using the weighted histogram analysis method (WHAM) [26, 100] and is shown in Figure 3.1. Six free energy basins are readily visible, and rectangular regions around these basins were chosen for the decomposition of all of configuration space into a set of six states. State definitions are listed in Table 3.1 and plotted as thick dividing lines in Figure 3.1.

3.3.4 Construction of Markov chain model from short trajectories.

To construct a Markov chain model of dynamics once the states were identified, the interstate transition probabilities were computed using the procedure described in Section 3.2.2. A set of 1000 energy-conserving trajectories 10 ps in length were generated from a canonical distribution of initial conditions within each state. This initial distribution was generated by selecting initial configurations from all replicas of the replica exchange simulation with a probability proportional to their weight in the canonical ensemble at 302K, as determined by WHAM, and assigning initial momenta from the Maxwell-Boltzmann distribution. For each lag time τ, the state populations, which correspond to an estimate of the transition probability $T_{ji}(\tau)$, were then obtained from Eq. 3.7. A bootstrap procedure [49], in which 200 replicates of 1000 trajectories from each state were chosen with replacement from the set of trajectories emanating from each state, was used to estimate the uncertainty in the observed transition probabilities.

The observed transition probabilities out of each state as a function of τ are shown in Figure 3.2, along with the corresponding equilibrium probabilities of each state determined from the replica-exchange simulation. None of the state populations reach their equilibrium values within 10 ps, indicating the slowest relaxation timescales are much longer, perhaps substantially so for trajectories originating from states 5 and 6. Transition matrices at several lag times — 0.1 ps, 1 ps, 6 ps, and 10 ps — are shown in Table 3.2.

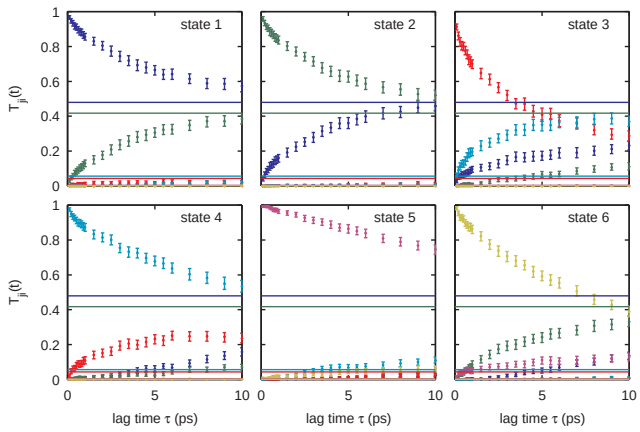


Figure 3.2: **Transition matrix elements as a function of lag time estimated from 10 ps shooting trajectories.** Each plot, labeled above by the state from which the trajectories originated, shows state-to-state transition probabilities as a function of the lag time τ estimated from a set of 1000 trajectories 10 ps in length originating from an equilibrium distribution within each state. Vertical bars depict 95% confidence intervals. Equilibrium state probabilities obtained from the parallel tempering simulations are shown as solid horizontal lines in the corresponding color.

3.3.5 Comparison with long trajectories.

To determine the accuracy with which transition matrices constructed from different lag times from short (10 ps) simulations are able to reproduce the statistical dynamics over long times (approximately 100 ps), state populations for an ensemble of trajectories emanating from each state were computed from the model and compared to the observed time evolution of a separate set of long trajectories. For this comparison, 1000 trajectories 100 ps in length were initiated from each state, using the same protocol in Section 3.3.4. Figure 3.3 shows the time evolution of state populations from these trajectories (along with corresponding uncertainties) as a function of time. Superimposed are state populations computed by Eq. 3.3 from the transition matrices constructed for different lag times τ from the short simulations described in Section 3.3.4. These are connected by straight line segments solely to guide the eye; the model cannot make predictions for the populations at times that are not integral multiples of the lag time τ.

The transition probabilities are poorly reproduced in the model constructed with a lag time of only 0.1 ps. Apparently, this time is so short that the system does not behave in a Markovian manner on this timescale. At a lag time of 1 ps, the agreement between the model and long simulations is clearly better, though there are still visible systematic deviations. By a lag time of 6 ps, the

Table 3.2: **Transition matrices[a] at several lag times estimated from set of 10 ps trajectories.**

$$\boldsymbol{T}(0.1\text{ ps}) = \begin{bmatrix} .967 & .041 & .029 & & & .002 \\ .030 & .959 & & .003 & & .001 \\ .003 & & .912 & .022 & & \\ & & .059 & .975 & & \\ & & & & .993 & .015 \\ & & & & .007 & .982 \end{bmatrix}$$

$$\boldsymbol{T}(1\text{ ps}) = \begin{bmatrix} .856 & .161 & .096 & .011 & .002 & .004 \\ .130 & .835 & .008 & .007 & & .086 \\ .014 & .002 & .701 & .109 & .001 & \\ & .002 & .195 & .873 & .014 & \\ & & & & .966 & .047 \\ & & & & .017 & .863 \end{bmatrix}$$

$$\boldsymbol{T}(6\text{ ps}) = \begin{bmatrix} .642 & .400 & .190 & .068 & .010 & .069 \\ .324 & .586 & .069 & .043 & .011 & .268 \\ .023 & .009 & .373 & .251 & .017 & .002 \\ .011 & .005 & .367 & .637 & .075 & .004 \\ & & .001 & .001 & .839 & .104 \\ & & & & .048 & .553 \end{bmatrix}$$

$$\boldsymbol{T}(10\text{ ps}) = \begin{bmatrix} .573 & .459 & .232 & .157 & .022 & .138 \\ .385 & .520 & .110 & .072 & .033 & .333 \\ .018 & .013 & .286 & .235 & .030 & .005 \\ .022 & .008 & .371 & .535 & .111 & .009 \\ & & .001 & .001 & .745 & .127 \\ .002 & & & & .059 & .388 \end{bmatrix}$$

[a] Blank entries denote estimated transition probabilities of zero.

agreement is excellent. The model constructed from a lag time of 10 ps also shows excellent agreement, but by this time, the temporal resolution has started to become rather poor. Information about the system is only known for times that are integral multiples of 10 ps. One can imagine that the most useful model would be constructed from the shortest lag time at which dynamics is Markovian, as this model has the highest temporal resolution while still correctly describing long-time dynamics

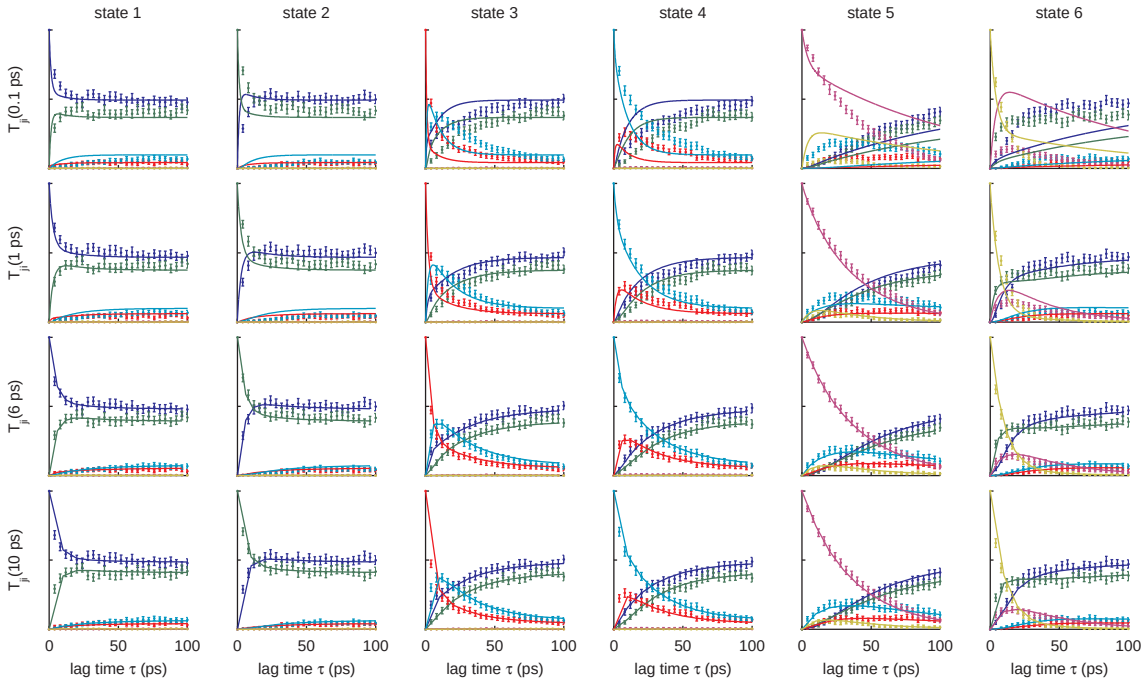


Figure 3.3: **Temporal evolution of state populations from Markov chains constructed at different lag times compared with long simulations.** Evolution of state probabilities from an ensemble prepared at equilibrium within each state for Markov model estimated from the set of 10 ps shooting trajectories (solid lines) superimposed on fractional population of each state as a function of time for ensemble of 100 ps trajectories initiated from each state (points). Vertical bars depict 95% confidence intervals in state populations estimated from the long trajectories.

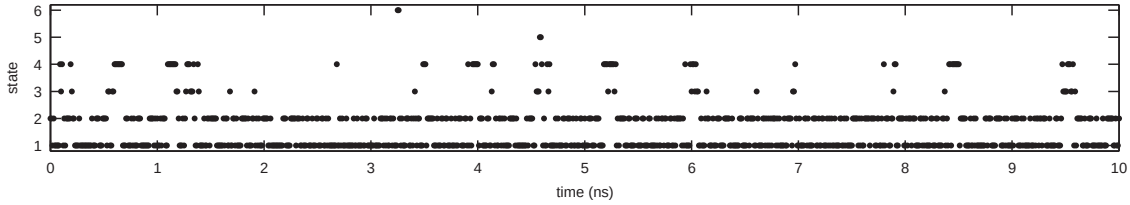


Figure 3.4: **An artificial trajectory generated from the transition matrix constructed from a lag time of 10 ps.**

3.3.6 Long-time dynamics from the Markov chain model.

As an illustration of the utility of the Markov chain model, Figure 3.4 depicts an *artificial* trajectory generated by realization of the Markov process, 10 ns in length, three orders of magnitude longer than the short trajectories used to construct the model. While statistical properties of the dynamics can also be extracted in other ways, such as through an eigenvalue decomposition, it may be useful to generate artificial trajectories and analyze them directly. Note the infrequent sampling of states 5 and 6, states with very small equilibrium probabilities, and the long dwell times in the region formed by stable states 1 and 2.

3.4 Discussion

We have demonstrated that a Markov model constructed from simulations roughly one order of magnitude shorter than the slowest relaxation time in the system is sufficient to capture the long-time dynamics of a simple biomolecular system, terminally-blocked alanine peptide in explicit solvent. Instead of generating large numbers of long trajectories to statistically characterize dynamics, we require only a sufficient number of trajectories to estimate transition probabilities between pairs of states. In addition, these trajectories need only be long enough for interstate dynamics to appear Markovian. Once so constructed, the model can be used to answer various questions of interest regarding the long-time statistical dynamics without the need to perform additional simulations.

While it is impossible to predict what the minimum trajectory length required for Markovian behavior will be for other, larger systems, it is important to recall that most proteins fold on the millisecond to second timescale. Even fast folding proteins can require tens of microseconds to fold [99]. To bring the treatment of these systems within the realm of feasibility, the Markov time

would need to remain sufficiently short to allow the collection of a significant number of trajectories despite the presence of relaxation times many orders of magnitude longer. No statement can yet be made about the number of states necessary to model more complex systems, or whether this number might make this approach prohibitively expensive.

To determine the lag time to construct the transition matrix so that the Markov chain is an accurate description of long-time dynamics, it was necessary to compare to an additional set of long trajectories. This, of course, defeats the utility of a model constructed from short trajectories. The question of how best to validate a Markov chain model constructed from short trajectories without additional long-time information is a topic of active research. In the worst case, the transition matrix constructed from a lag time of 0.1 ps clearly disagree at long times — this disagreement is clearly an indicator of disagreement at long times. Other methods, such as tests of eigenvalue behavior [174] or direct tests of Markovity [141], may provide alternatives.

In this work, we have avoided the issue of how best to define the number and location of states used for construction of the Markov model. Ideally, these states will be significantly *metastable* so that the system rapidly loses memory of its previous location after entering a state, before making a transition to another state. For the system considered, the slow degrees of freedom were known beforehand, so the potential of mean force in these coordinates revealed a useful set of states. In more complex systems, the coordinates in which dynamics is slow will be much more difficult to discern; some automatic method for the identification of metastable states is necessary, which is the subject of work soon to be reported [160].

Here, we employed the most straightforward approach to estimating interstate transition probabilities, whereby a large number of short trajectories are initiated from equilibrium within each state. While this approach is amenable to distributed or grid computing, the metastable nature of well-chosen states will result in many of these trajectories simply remaining in their state of origin, rather than contributing to estimates of the off-diagonal elements of the transition matrix. It is precisely these off-diagonal elements that are critical in determining which trajectories through state space are most likely. Algorithms employing importance sampling techniques in *trajectory* space — such as transition path sampling [36], transition interface sampling [189], and the string method [47] — may provide an efficient way to compute these interstate transition probabilities.

3.5 Acknowledgements

JDC was supported by Howard Hughes Medical Institute and IBM predoctoral fellowships. WCS acknowledges support from NSF MRSEC Center on Polymer Interfaces and Macromolecular Assemblies DMR – 0213618, and KAD the support of NIH grant GM34993. JDC gratefully acknowledges Libusha Kelly (UCSF) for critical reading of this manuscript and Nina Singhal (Stanford) for stimulating discussion and insightful criticism.

Chapter 4

Validating master equation models of macromolecular dynamics

The material in this chapter is being prepared for submission to *The Journal of Physical Chemistry B* as

Describing protein folding kinetics by molecular dynamics simulations. 3. Validation of state space decomposition, with application to terminally-blocked alanine in explicit solvent

John D. Chodera[1], **William C. Swope**[2], **Jed W. Pitera**[2], **and Ken A. Dill**[3,1]

[1] *Graduate Group in Biophysics and* [3] *Department of Pharmaceutical Chemistry, University of California at San Francisco, San Francisco, CA 94143*

[2] *IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120*

Abstract

Despite recent interest in the construction of simple stochastic models to describe the conformational dynamics of peptides and proteins, little work has focused on verifying whether these models provide an accurate description of dynamics. Previously, we demonstrated that the statistical dynamics of a small solvated peptide (terminally-blocked alanine) over long times could be described by a discrete-state master equation model constructed from short molecular dynamics simulations in explicit solvent [25]. The accuracy of the model was verified by comparison with a large number of trajectories long enough to reach equilibrium, a method which is impractical for systems with timescales that exceed typical simulation lengths by orders of magnitude. Here, we

focus on how to determine whether a model constructed from short trajectories will accurately reproduce dynamics over long times without the need to conduct additional simulations. We examine a number of tests of Markovianity to assess their ability determine whether models constructed from these short trajectories will be able to reproduce dynamics over long times, and if so, on what timescale the dynamics appears Markovian. We use the same solvated peptide system as in a previous work [25], and analyze both good state decompositions (where dynamics is expected to be Markovian on short timescales) and poor decompositions (where dynamics is only Markovian on longer timescales, possibly longer than the short trajectories used to construct the model).
Keywords: master equation model; Markov chain model; molecular dynamics; peptide dynamics; protein folding; transition matrix; rate matrix

4.1 Introduction

Since the first crystallographic studies revealed the intricate three-dimensional structures of proteins, an understanding of the specific mechanisms and general principles governing the process by which they fold has been sought by experimentalists and theorists alike. Today, fundamental questions, such as whether folding pathways are microscopically homogeneous or heterogeneous, or whether non-native traps or long-lived intermediates exist, remain unresolved. The ability to characterize folding pathways in detail for even a few proteins would help resolve these questions. Understanding of folding will also provide insight into the mechanism of protein misfolding diseases, how native structures are encoded by sequence, and how proteins with novel folds and functions might be engineered.

Molecular dynamics simulations allow us to study the folding process in atomic spatial and temporal detail. However, characterization of protein folding pathways by straightforward molecular dynamics simulation is extremely challenging; proteins fold on the order of microseconds to seconds, while typical molecular simulations reach timescales of only tens of nanoseconds. In addition, simulations typically model a *single* solvated protein molecule, while experiments generally observe the average behavior of a large *ensemble* of molecules. While massively parallel supercomputers such as Blue Gene [69] can reach timescales of a few microseconds with great feats of engineering [62, 73], a useful characterization of a protein folding mechanism would still require a *statistical* description of events. Such a description could be at a coarse-grained level that fits our level of physical interest, such as “helix A forms before helix B”

(which we shall term a *pathway* or *route*), with each pathway carrying an associated statistical weight. If there is more than one pathway, a very large number of folding trajectories would be required to statistically characterize the relative probabilities of different pathways. Even if there is only a single folding pathway, our molecular dynamics trajectories would need to be long enough to experience a folding event starting from a suitable unfolded state, and a sufficient number of trajectories would need to be collected to establish that there is only a single pathway. Distributed computing projects now allow for the collection of thousands of trajectories tens of nanoseconds in length, such that a small number of them can contain folding events [158, 201]. Unfortunately, the possibility remains that folding events observed in short trajectories may be statistically different from the bulk of folding events, which occur over much longer timescales [59, 116].

In order to statistically characterize probable folding routes or verify that only single routes dominate, it is necessary to construct a model from which these probabilities can be computed using simulation data that can be obtained with reasonable computational effort. Master equation or Markov chain models [25, 162, 174, 190] present such a solution. Master equations are Markov models of dynamics either in the low dimensional manifold of the slow degrees of freedom or, as we consider here, a discrete state space where the states correspond to approximations of the metastable conformational substates accessible to the molecule. These models can describe dynamics either in discrete time (here termed a *Markov chain model*) or in continuous time (*master equation model*); in either case it is understood that there is a coarse-graining of the time coordinate such that only processes occurring slower than a given timescale are adequately described by the given model. Once the model is constructed and the timescale for Markovian behavior determined, it can be used to compute the stochastic temporal evolution of either a single macromolecule or a population of noninteracting macromolecules, allowing direct comparison of simulated and experimental observables for both single-molecule or ensemble kinetics experiments. In addition, useful properties difficult to access experimentally, such as state lifetimes [174], relaxation from experimentally inaccessible prepared states [25], mean first-passage times [162], the existence of hidden intermediates [137], and P_{fold} values [103] or committor distributions [71], can easily be obtained.

Such models are a natural choice for representing protein dynamics, in which processes with intrinsic timescales ranging from femtoseconds (*e.g.* bond vibrations) to microseconds or milliseconds (folding) can be identified. If there is a clear *separation of timescales* between fast, uninteresting dynamics and slow, large-scale conformational changes, then these models can provide an excellent description of the slow dynamics of interest. Several groups have

demonstrated that the hierarchical nature of the energy landscape results in a natural separation of timescales in peptide simulations with implicit models of solvent *in vacuo* [105, 125], a property that may also hold for solvated systems.

Construction of master equation models for realistic molecular mechanics systems has, however, proved problematic. Techniques based on enumerating all minima and computing interstate transition rates with transition state theory have proved insightful for simple systems *in vacuo* or with an implicit model of solvent [8, 32, 52, 101, 105, 124, 125], but as the number of minima grows exponentially with system size, this procedure rapidly becomes untenable for larger proteins or systems containing explicit solvent. Due to these limitations, much work has focused on the construction of discrete- or continuous-time Markov models to describe dynamics among a small number of states which may each contain many minima within large regions of configuration space [2, 35, 51, 74, 141, 151, 161, 162, 164, 167, 175].

Here, we follow a method for the construction of a discrete-state master equation model based on calculation of state-to-state time-correlation functions proposed by Swope *et al.* [174]. A previous investigation applying this method to a model system (terminally-blocked alanine peptide in explicit solvent) demonstrated that a 6-state Markov chain model with a time resolution of 6 ps constructed from short (10 ps) trajectories was able to describe dynamics over timescales of at least 100 ps [25]. In that work, accurate reproduction of statistical dynamics was verified by comparison with a separate set of long (100 ps) trajectories. The goal, however, is to construct and validate models of dynamics to allow long timescales to be reached without the need to generate long trajectories. A method is needed to determine the timescale at which the model will be Markovian and decide between multiple state decompositions given only the set of short trajectories used to construct the model.

Here, we focus on how to determine whether a particular choice of state definitions can lead to a master equation model capable of adequately describing interstate dynamics and for what coarse-grained timescale this model will be accurate. Since the optimal state definition must generally be arrived at through trial and error, we perform several tests to discriminate among alternate choices of state definitions to determine which choice results in the most *useful* master equation model. We consider terminally-blocked alanine peptide in explicit solvent as a model system where the slow degrees of freedom, the (ϕ, ψ) torsions, are known *a priori*. Since there are only two slow degrees of freedom[2], we construct a two-dimensional potential of mean force (PMF)

[2]Simulations of alanine dipeptide examining the committor distribution have implicated solvent coordinates as the next-slowest degrees of freedom [17, 111], but we have previously verified that ϕ and ψ torsions form a sufficient basis

in these coordinates by simulation, and visually identify the metastable conformational substates. In this way, we can separate the problem of finding a good decomposition into metastable states from that of computing transition rates and assessing the resulting master equation model.
As in Ref. [174], we first obtain a broad sampling of conformation space by way of a parallel tempering simulation, from which we also obtain the PMF and from this, the state decomposition. A number of short trajectories are then initiated from an equilibrium distribution within each state obtained from the parallel tempering simulation in order to collect state-to-state transitions. These simulations can then be used to construct an interstate transition matrix and determine the timescale for which the underlying rate matrix becomes an adequate description of the dynamics.
This paper is organized as follows: In Section 4.2, we briefly review the the discrete-state master equation and Markov chain theory and the basic concepts behind its construction from simulation data. Section 4.3 introduces the terminally-blocked alanine in explicit solvent model system (the same system considered in a previous study [25]) that will be used to assess the utility of various tests of Markovianity. Finally, Section 4.4 describes a number of such tests and their observed performance in determining the time for the emergence of Markovian behavior for this model system.

4.2 Master equation and Markov models

4.2.1 The discrete-state master equation.

The dynamical evolution of a discrete-state, continuous-time process that is Markovian on all timescales is governed by the so-called *master equation* (see, for example, van Kampen [190] for an excellent review) which can be written as

$$\frac{\partial}{\partial t}\boldsymbol{p}(t) = \boldsymbol{K}\boldsymbol{p}(t). \tag{4.1}$$

Here, $\boldsymbol{p}(t) \in \Re^M$, where $p_i(t)$ denotes the probability of state i being occupied at time t, requiring $p_i \geq 0$ and $\sum_{i=1}^{M} p_i = 1$. $\boldsymbol{K}$ is the matrix of rate constants, with K_{ji} denoting the rate constant associated with the transition from state i to state j if $j \neq i$, and the diagonal elements determined such that the columns sum to zero to ensure conservation of total probability, *i*.e., $K_{ii} = -\sum_{j\neq i}^{M} K_{ji}$. The equilibrium (or *stationary*) distribution $\boldsymbol{p}_{\mathrm{eq}}$ is given by

$$\frac{\partial}{\partial t}\boldsymbol{p}_{\mathrm{eq}} = \boldsymbol{K}\boldsymbol{p}_{\mathrm{eq}} = \boldsymbol{0}. \tag{4.2}$$

for the slow degrees of freedom on timescales of 6 ps and greater [25].

For the systems we consider here, we shall presume that only one stationary distribution exists. In problems of physical interest, the rate matrix $\boldsymbol{K}$ statisfies the condition of *detailed balance*

$$K_{ji}\, p_{\mathrm{eq},i} = K_{ij}\, p_{\mathrm{eq},j}. \tag{4.3}$$

As a result, $\boldsymbol{K}$ is related to a symmetric matrix $\tilde{\boldsymbol{K}}$ by a similarity transformation

$$\tilde{\boldsymbol{K}} = \boldsymbol{D}^{-1/2}\boldsymbol{K}\boldsymbol{D}^{1/2} = \boldsymbol{D}^{1/2}\boldsymbol{K}^{\mathrm{T}}\boldsymbol{D}^{-1/2} = \tilde{\boldsymbol{K}}^{\mathrm{T}} \tag{4.4}$$

where $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{p}_{\mathrm{eq}})$ is the diagonal matrix of equilibrium probabilities. $\tilde{\boldsymbol{K}}$ is therefore orthogonally diagonalizable

$$\tilde{\boldsymbol{K}} = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{U}}^{\mathrm{T}} \tag{4.5}$$

with $\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^{\mathrm{T}} = \boldsymbol{I}$ and $\tilde{\boldsymbol{\Lambda}} = \mathrm{diag}(\tilde{\lambda_1}, \tilde{\lambda_2}, \ldots, \tilde{\lambda_M})$ is the diagonal matrix of eigenvalues, sorted such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M$. $\boldsymbol{K}$ shares the same eigenvalues as $\tilde{\boldsymbol{K}}$:

$$\boldsymbol{K} = \boldsymbol{D}^{1/2}\tilde{\boldsymbol{K}}\boldsymbol{D}^{-1/2} = \left(\boldsymbol{D}^{1/2}\tilde{\boldsymbol{U}}\right)\tilde{\boldsymbol{\Lambda}}\left(\boldsymbol{D}^{1/2}\tilde{\boldsymbol{U}}\right)^{-1} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1} \tag{4.6}$$

but the eigenvectors differ by a factor of $\boldsymbol{D}^{1/2}$. All the eigenvalues are real, and exactly one eigenvalue, λ_1, is zero — all others are negative.

The time evolution of an initial probability vector $\boldsymbol{p}(0)$ is given by the solution to Eq. 4.1

$$\boldsymbol{p}(t) = e^{\boldsymbol{K}t}\boldsymbol{p}(0) \equiv \boldsymbol{T}(t)\boldsymbol{p}(0) \tag{4.7}$$

where $e^{\boldsymbol{K}t}$ denotes the matrix exponential, and $\boldsymbol{T}(t)$ is called the *transition matrix*, as element $T_{ji}(t)$ denotes the probability of observing the system in state j a time t after initially observing it in state i[3]. $\boldsymbol{T}(\tau)$ is sometimes also referred to as the *propagator*, as its operation on a probability vector $\boldsymbol{p}(t)$ evolves the distribution by a fixed time τ. The eigenvalues of $\boldsymbol{T}(t)$ and $\boldsymbol{K}$ are simply related

$$\boldsymbol{T}(t) = \boldsymbol{U}\boldsymbol{M}\boldsymbol{U}^{-1} = e^{\boldsymbol{K}t} = \boldsymbol{U}e^{\boldsymbol{\Lambda}t}\boldsymbol{U}^{-1} \tag{4.8}$$

where $\boldsymbol{M} = \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_M)$ is the matrix of eigenvalues of the transition matrix, which lie in the interval $[0, 1]$. Because $\boldsymbol{U} = \boldsymbol{D}^{1/2}\tilde{\boldsymbol{U}}$, this decomposition allows us to write $\boldsymbol{T}(t)$ as an expansion over the eigenvectors of $\tilde{\boldsymbol{K}}$ as

$$\boldsymbol{T}(t) = \sum_{k=1}^{M} \boldsymbol{D}^{1/2}\, \tilde{\boldsymbol{u}}_k\, e^{\lambda_k t}\, \tilde{\boldsymbol{u}}_k^{\mathrm{T}}\, \boldsymbol{D}^{-1/2}. \tag{4.9}$$

[3] We adopt the notation of a *column-stochastic* transition matrix, where the columns of $\boldsymbol{T}(t)$ sum to unity. In mathematical literature on Markov chains, and in some publications cited here, it is often common to instead see *row-stochastic* transition matrices, where the rows sum to unity, and evolution is governed by the equation $[\boldsymbol{p}(t)]^{\mathrm{T}} = [\boldsymbol{p}(0)]^{\mathrm{T}}\boldsymbol{T}(t)$

For physical systems, we do not expect Eq. 4.7 to hold for all times once we have coarse-grained space into states. Instead, the rates K_{ji} reflect *phenomenological rates* that may only correctly describe interstate transitions after some coarse-grained *internal equilibration time* τ_{int} [20]. Because of this, states that do not share a boundary in configuration space may still have nonzero phenomenological transition rates connecting them [191]. Additionally, for physical systems in which all states are accessible, there is always some probability that a system prepared in one state will have a momentum that will carry it to another state some finite time τ later, so that all transition elements $T_{ji}(\tau)$ should be nonzero, though they may be very small.

4.2.2 Construction from molecular dynamics simulation.

Here, we briefly review the construction of a discrete-state master equation model from simulation. We obtain the same expressions as in Ref. [174], but our exposition and notation differs slightly. We denote a point in phase space by $\boldsymbol{z} = (\boldsymbol{q}, \boldsymbol{p})$, where $\boldsymbol{q}, \boldsymbol{p} \in \Re^{3N}$ are the coordinates and momenta, respectively. A trajectory in phase space of time length T is denoted $\boldsymbol{z}(t)$, $t \in [0, T]$, with t denoting the time index. The system has Hamiltonian

$$H(\boldsymbol{z}) = \frac{1}{2}\boldsymbol{p}^{\mathrm{T}}\boldsymbol{M}^{-1}\boldsymbol{p} + V(\boldsymbol{q}) \tag{4.10}$$

where $\boldsymbol{M}$ is the diagonal matrix of atomic masses, and $V(\boldsymbol{q})$ is the potential function. The equilibrium average of a phase function or mechanical observable $A(\boldsymbol{z})$ is given by the expectation over the canonical ensemble

$$\langle A \rangle = \frac{\int d\boldsymbol{z}\, e^{-\beta H(\boldsymbol{z})}\, A(\boldsymbol{z})}{\int d\boldsymbol{z}\, e^{-\beta H(\boldsymbol{z})}} \tag{4.11}$$

where $\beta = (k_B T)^{-1}$ denotes the inverse temperature. Similarly, the average over a functional of trajectories $A[\boldsymbol{z}(t)]$, which depends on the value of $\boldsymbol{z}(t)$ at multiple times, is given by

$$\langle A[\boldsymbol{z}(t)] \rangle \equiv \frac{\int \mathcal{D}[\boldsymbol{z}(t)]\, \mathcal{P}[\boldsymbol{z}(t)]\, A[\boldsymbol{z}(t)]}{\int \mathcal{D}[\boldsymbol{z}(t)]\, \mathcal{P}[\boldsymbol{z}(t)]} \tag{4.12}$$

where $\mathcal{D}[\boldsymbol{z}(t)]$ is the differential over constant-energy trajectories $\boldsymbol{z}(t)$ and $\mathcal{P}[\boldsymbol{z}(t)]$ is the probability density of paths. We presume dynamics is governed by Hamilton's equations of motion[4]

$$\frac{\partial}{\partial t}\boldsymbol{q} = \nabla_{\boldsymbol{p}} H \; ; \; \frac{\partial}{\partial t}\boldsymbol{p} = -\nabla_{\boldsymbol{q}} H.$$

[4] In this study and the previous one [25], we employ dynamical simulations with bond length constraints for efficiency purposes. Technically, this causes dynamics to be non-Hamiltonian [63,185], but we assume this difference is unimportant for the purposes of our analyses. In future, multiple timestep integrators (*e.g.* [184]) could be used to void loss in efficiency while eliminating the need for bond length constraints.

which allows us to rewrite the expectation over functionals of trajectories as

$$\langle A[\boldsymbol{z}(t)]\rangle = \frac{\int d\boldsymbol{z}(0)\, e^{-\beta H(\boldsymbol{z}(0))}\, A[\boldsymbol{z}(t)]}{\int d\boldsymbol{z}\, e^{-\beta H(\boldsymbol{z})}}. \tag{4.13}$$

We suppose that we have already followed some procedure to generate a decomposition of phase space into a complete set of M non-overlapping regions $\mathcal{S}_i$, which we term *states*, that together form a complete covering of phase space. We define an *indicator function* $\chi_i(\boldsymbol{z})$ for state i such that

$$\chi_i(\boldsymbol{z}) = \begin{cases} 1 & \text{if } \boldsymbol{z} \in \mathcal{S}_i \\ 0 & \text{otherwise.} \end{cases}$$

Since the states form a complete covering of phase space, we have

$$\sum_{i=1}^{M} \chi_i(\boldsymbol{z}) = 1 \;\forall\, \boldsymbol{z}.$$

Furthermore, we presume these states are only functions of configuration $\boldsymbol{q}$, that is, $\chi_i(\boldsymbol{z}) = \chi_i(\boldsymbol{q})$. As the phenomenological transition rates K_{ji} are not directly observable, and may not even exist at short times, our procedure instead involves computing the observed state-to-state transition probabilities $T_{ji}(\tau)$ [174]. We shall often refer to the evolution time τ as the *lag time* because it refers to the time between subsequent observations of the system. Formally, this transition probability is defined as

$$T_{ji}(\tau) \;\equiv\; \frac{C_{ji}(\tau)}{\sum\limits_{j'=1}^{M} C_{j'i}(\tau)} = \frac{C_{ji}(\tau)}{p_{\mathrm{eq},i}} \tag{4.14}$$

where we have defined the state-state time correlation function $C_{ji}(\tau)$ by

$$\begin{aligned} C_{ji}(\tau) &\equiv \langle \chi_j(\boldsymbol{z}(\tau))\, \chi_i(\boldsymbol{z}(0))\rangle \\ &= \langle \chi_i(\boldsymbol{z}(0))\, \chi_j(\boldsymbol{z}(\tau))\rangle \\ &= C_{ij}(\tau). \end{aligned} \tag{4.15}$$

The second equality follows by time-reversal symmetry of Newtonian trajectories and the invariance of the $\chi_i(\boldsymbol{z})$ to inversion of the momenta.

4.3 Terminally-blocked alanine peptide in explicit solvent as a model system

Here, we describe the terminally-blocked alanine in explicit solvent model system considered in this study to evaluate various tests of Markovianity. Previously, it was determined that a Markov chain model constructed for this system using a lag time of 6 ps was able to describe the statistical dynamical evolution over long times, as verified by an independent set of 100 ps trajectories [25]. For convenience, we describe the protocol used here to generate the simulation data (the same data as considered in Ref. [25]) and the methods used to estimate transition probabilities.

4.3.1 System setup and equilibration.

Using the LEAP program from the AMBER7 molecular mechanics package [18], a terminally-blocked alanine peptide (sequence Ace-Ala-Nme, see Figure 4.1) was generated in the extended conformation and subsequently solvated with 431 TIP3P water molecules [90] in a truncated octahedral simulation box whose dimensions were chosen to ensure a minimum of 7 Å distance from the peptide to the box faces. Peptide force field parameters were taken from the AMBER parm96 parameter set [95]. All molecular dynamics simulations were conducted using the `sander` program from the AMBER7 package. Default nonbonded cutoffs were used, bonds to hydrogen were constrained with SHAKE using a tolerance of 10^{-8} Å [148], and long-range electrostatics were treated by the particle-mesh Ewald (PME) method [34] with the default settings. The system was first subjected to 50 steps of steepest descent energy minimization, followed by 1000 steps of conjugate gradient optimization. To equilibrate the explicit solvent system to the appropriate volume, a 100 ps molecular dynamics simulation was performed with the temperature adjusted to 300 K and the pressure to 1 atm by the Berendsen weak-coupling algorithm [10] with temperature and pressure relaxation time constants of 1 ps and 0.2 ps, respectively. The simulation box was fixed at the final size obtained from this equilibration step, with a volume of 13 232 Å^3, in all subsequent simulations.

4.3.2 Parallel tempering simulation.

In order to broadly explore the configuration space of the peptide and ensure that all important conformational substates were located, a parallel tempering (also known as replica-exchange among temperatures) molecular dynamics simulation [171] was conducted using a parallel `Perl`

wrapper for the `sander` program[5]. Forty replicas were used, with replica temperatures exponentially distributed over the range 273–600 K, yielding an average exchange acceptance probability of about 50%. All momenta were reassigned from the Maxwell-Boltzmann distribution at the appropriate replica temperature after each exchange attempt, and constant-energy, constant-volume molecular dynamics with a 2 fs timestep was performed between exchange attempts. The algorithm used to select pairs of replicas for temperature exchange attempts starts from the highest-temperature replica and attempts to swap the configuration for the next-lowest temperature replica using the Metropolis-like criteria, and proceeds down the temperatures in this manner. On the next iteration, swapping attempts start from the lowest temperature and proceed upward, and this alternation in direction is continued in subsequent pairs of iterations.
Starting all replicas from the 300 K NPT-equilibrated configuration described above, 100 iterations were conducted with 1 ps between exchange attempts to equilibrate the replicas to their respective temperatures. This equilibration run was followed by a production run of 500 iterations with 20 ps between exchange attempts, producing a total of 10 ns/replica, or 400 ns in aggregate, for the production run. The longer time between exchanges for the production run was chosen so as not to hinder transitions by frequent velocity randomization. Solute snapshots and potential energies were stored during the production run every 0.1 ps for the purposes of estimating the free energies of each state (resulting in $4 \cdot 10^6$ saved solute configurations), and full-system restart files were saved every 1 ps (yielding $4 \cdot 10^5$ restart files) for the purpose of starting shooting simulations from these configurations, described below in Section 4.3.4.

4.3.3 State decomposition.

In more complex molecular systems, it is generally a difficult problem to identify the "slow" degrees of freedom involved in conformational transitions between metastable states. Here, we postulate that only a subset of the solute degrees of freedom will be necessary to accurately describe the long-time dynamics of the system. We presume the solvent will be important in determining the inter-state rate constants, but plays no real role in determining the gross structure of the conformational substates at long timescales. To this end, the potential of mean force of the two backbone torsion angles ϕ and ψ at 302 K was estimated from the parallel tempering data using the weighted histogram analysis method (WHAM) [26, 100] and is shown in Figure 4.1. Six free energy basins are readily visible, and rectangular regions around these basins were manually

[5] A copy of this Perl wrapper to perform replica-exchange simulations using AMBER7 and AMBER8 can be obtained from `http://www.dillgroup.ucsf.edu/~jchodera/code/rex`.

chosen as previously [25] to serve as a "good" decomposition of all of configuration space into a set of six states for the construction of a discrete-state master equation model. These states are depicted in Figure 4.1, alongside a "poor" 6-state decomposition, in which the boundaries have been significantly displaced so as to include internal barriers within states[6]. Two additional state decompositions were considered, both of which were "lumpings" of the good 6-state decomposition. A "good" lumping, which is intended to be minimally perturbative on the long timescales, was constructed by lumping states 1 and 2, as well as states 3 and 4, as each pair of states is only separated by low free energy barriers. A "poor" lumping was also constructed by lumping states 1, 4, and 5 together into a single state containing large internal barriers, which is expected to disrupt the ability to produce a good model of the statistical dynamics from short simulations.

It should be noted that this approach to state decomposition is only reasonable if the PMF in the complete set of slow degrees of freedom can be computed and examined — projection onto an arbitrary set of structural observables or order parameters (such as the number of native contacts and radius of gyration) and identification of basins as metastable states is not the same thing. When arbitrary order parameters are used, there is no guarantee that connected regions in the low-dimensional order parameter space correspond to single, connected regions in phase space, a requirement for a Markovian model of dynamics. Additionally, the projection must account for the Jacobian, so as not to distort space so much that free energy barriers are not artificially introduced or eliminated [79, 80].

4.3.4 Shooting simulations.

In order to construct a discrete-state master equation model once the states have been identified, we must obtain reliable estimates of the interstate transition probabilities. While a number of equilibrium trajectories could be used for this purpose, states of high free energy would be visited only infrequently, resulting in the transition elements of these states being poorly determined. Swope *et al.* proposed the use of *selection cells*, (potentially overlapping) regions which together form a complete covering of configuration space from which sets of trajectories are initiated [174]. The purpose of these selection cells is to allow more trajectories from high free energy regions to

[6]The poor partitioning was defined as follows: (1) $\phi \in [(179, -135]$, $\psi \in (98, 48]$; (2) $\phi \in (-135, -60]$, $\psi \in (98, 48]$; (3) $\phi \in (179, -135]$, $\psi \in (48, 98]$; (4) $\phi \in (-135, -60]$, $\psi \in (48, 98]$; (5) $\phi \in (-60, 179]$, $\psi \in (98, -45]$; (6) $\phi \in (-60, 179]$, $\psi \in (-45, 98]$. Specified intervals denote intervals on the torus, which is continuous from -180 to +180. All torsion angles are specified in degrees.

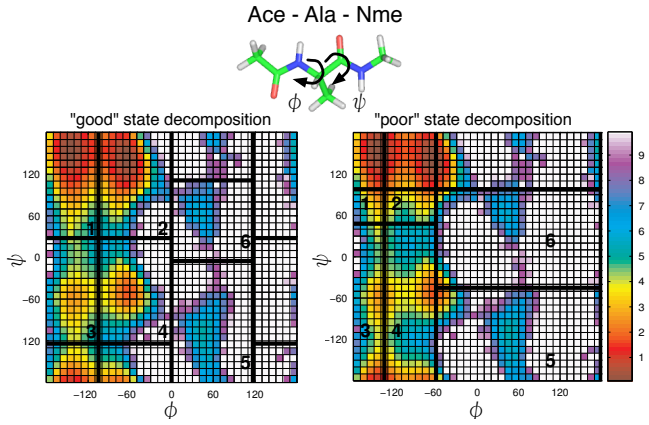


Figure 4.1: **Terminally-blocked alanine and potential of mean force at 302 K.** Top: The terminally-blocked alanine peptide with (ϕ, ψ) torsions labeled. Bottom: The potential of mean force in the (ϕ, ψ) torsions at 302 K estimated from the parallel tempering simulation, with "good" (left) and "poor" (right) state decompositions labeled, and colorbar graded in units of $k_B T$ and truncated at $10 k_B T$ (white).

be harvested, allowing the transition matrix to be determined to sufficient precision.

Here, we choose to initiate sets of trajectories from a set of selection cells that are coincident with the six states of the "good" manual decomposition described above. For this case, we may decompose the expectation of any property that is a functional of trajectories, $\langle A[\mathbf{z}(t)] \rangle$, into a weighted sum over conditional expectations:

$$\begin{aligned}
\langle A[\mathbf{z}(t)] \rangle &= \frac{\int d\mathbf{z}(0)\, e^{-\beta H(\mathbf{z}(0))}\, A[\mathbf{z}(t)]}{\int d\mathbf{z}\, e^{-\beta H(\mathbf{z})}} \\
&= \frac{\int d\mathbf{z}(0) \left[\sum_{i=1}^{M} \chi_i(\mathbf{z}(0)) \right] e^{-\beta H(\mathbf{z}(0))}\, A[\mathbf{z}(t)]}{\int d\mathbf{z} \left[\sum_{i=1}^{M} \chi_i(\mathbf{z}(0)) \right] e^{-\beta H(\mathbf{z})}} \\
&= \frac{\sum_{i=1}^{M} \left[\frac{\int d\mathbf{z}\, \chi_i(\mathbf{z})\, e^{-\beta H(\mathbf{z})}}{\int d\mathbf{z}\, e^{-\beta H(\mathbf{z})}} \right] \left[\frac{\int d\mathbf{z}(0)\, \chi_i(\mathbf{z}(0))\, e^{-\beta H(\mathbf{z}(0))}\, A[\mathbf{z}(t)]}{\int d\mathbf{z}\, \chi_i(\mathbf{z})\, e^{-\beta H(\mathbf{z})}} \right]}{\sum_{i=1}^{M} \left[\frac{\int d\mathbf{z}\, \chi_i(\mathbf{z})\, e^{-\beta H(\mathbf{z})}}{\int d\mathbf{z}\, e^{-\beta H(\mathbf{z})}} \right]} \\
&= \frac{\sum_{i=1}^{M} w_i \langle A[\mathbf{z}(t)] \rangle_i}{\sum_{i=1}^{M} w_i}
\end{aligned} \tag{4.16}$$

where the weights w_i are proportional to the equilibrium state probabilities $p_{\mathrm{eq},i} = \langle \chi_i \rangle$, and the

conditional expectation $\langle A[\mathbf{z}(t)]\rangle_i$ is the expectation computed over trajectories of length T originating in state i. Here, the weights are obtained from the free energies estimated from the parallel tempering simulation[7]. For convenience, we require the weights to sum to unity, *i.e.*, $\sum_{i=1}^{M} w_i = 1$.

A set of 5000 NVE trajectories 10 ps in length were initiated from each state, for an aggregate simulation time of 300 ns. Initial configurations were selected from the entire pool of conformations generated from the production replica-exchange simulation, with each snapshot chosen with a likelihood appropriate to its probability of being sampled at 302 K, as computed by WHAM. Initial velocities were randomly assigned from the Maxwell distribution at the same temperature. Estimation of various properties (except for the transition probabilities, described below) used weights $w_i = e^{-\beta F_i}$, where the state free energies F_i estimated from the parallel tempering simulation were used [25].

4.3.5 Uncertainty estimation.

To assess uncertainties in quantities computed from the sampled set of trajectories, bootstrap resampling was performed. In bootstrap, the distribution of finite size samples over replications of the experiment is estimated by resampling from the original set of independent measurements. In applying the method here, we presume the shooting trajectories originating from each state are independent, and resample from this set. For each bootstrap trial, an artificial sample containing the same number of shooting trajectories from each state as the original set is constructed by drawing trajectories at random, with replacement, from this set of trajectories. Because entire trajectories are considered as independent measurements, the time-correlation within each trajectory is properly captured. All expectations (e.g. transition probabilities, eigenvalues of the transition matrix, lifetimes) are computed from this set of trajectories, weighting trajectories originating from each state as prescribed by the selection cell method described in Section 4.3.4 above. Forty such trials are conducted, the means computed and reported, and the distribution of expectations stored to indicate the uncertainty in these values, presented either as a standard deviation or plotted as a 68% confidence interval.

[7] Note that it may be desirable to update the weights by self-consistent iteration to ensure that $w_i \propto \langle \chi_i \rangle$, but this was not done here.

4.3.6 Estimation of transition probabilities.

To estimate the transition matrix $\boldsymbol{T}(\tau)$ from the set of 10 ps shooting trajectories, we compute the expectation $C_{ji}(\tau) = \langle \chi_j(t)\chi_i(0)\rangle$ by Eq. 4.16 above, taking advantage of the time-reversibility of Hamiltonian trajectories, the invariance of the state indicator functions upon momentum reversal, and stationarity:

$$\begin{aligned} C_{ji}(\tau) \;\approx\; & \sum_{k=1}^{M} w_k \frac{1}{N} \sum_{n=1}^{N} \frac{1}{T-\tau} \sum_{t_0=1}^{T-\tau} \\ & \frac{1}{2}\left[\chi_j(\mathbf{z}_{kn}(t_0+\tau))\chi_i(\mathbf{z}_{kn}(t_0)) \right. \\ & \left. +\chi_j(\mathbf{z}_{kn}(t_0))\chi_i(\mathbf{z}_{kn}(t_0+\tau))\right] \end{aligned} \tag{4.17}$$

where $\mathbf{z}_{kn}(t)$ denotes the nth trajectory initiated from state k. The transition matrix is then estimated from the correlation functions by Eq. 4.14 above.

With well-chosen selection cells that are nearly coincident with the true metastable states, as is the case here, is possible that the parallel tempering simulation generated configurations that are distributed from equilibrium *within* each state, but that the relative weights of the states w_i have high uncertainties. This is especially possible if the cells are separated by kinetic barriers, such that there are few transitions between the states observed in the simulation. The shooting trajectories, here a total of 300 ns at 302 K, provide an abundance of information about the relative populations of the various states at the temperature of interest, since a number of transitions out of each state are sampled. Because of this, we might trust the equilibrium probabilities estimated from the stationary eigenvector of the transition matrix more than the parallel tempering simulation, which only contained 10 ns of data at 302 K, though many replicas contribute to the estimate of the state free energies through the WHAM procedure. As the transition matrix itself depends on the weights, a simple way to ensure this self-consistency would be to iteratively update the vector of weights by the relationship

$$\boldsymbol{w}^{(n)} \;=\; \boldsymbol{T}(\tau)\,\boldsymbol{w}^{(n-1)} \tag{4.18}$$

to obtain a new estimate of the unnormalized weights $\boldsymbol{w}^{(n)}$, with which a new estimate of the transition matrix $\boldsymbol{T}(\tau)$ is computed, and so on, until the weights converge to within some tolerance. Here, we employed this procedure in the estimation of transition matrices, where a relative tolerance of 10^{-4} was used to determine convergence.

4.4 Tests for Markovianity.

As Hamiltonian dynamics is deterministic, it is by definition Markovian at all times in the full phase space of the system. However, when the momenta are neglected and phase space is coarse-grained into a finite number of states by aggregating contiguous regions of configuration space, specification of the state i currently occupied by the system does not provide sufficient information to uniquely determine the current phase space point, and hence the consequent dynamics of the system. As a result, the dynamical evolution can only be described *statistically*, in that it is only known that a certain fraction $T_{ji}(\tau)$ of trajectories will be found in a state j at some time τ later. Additionally, the dynamics in this state space will possess *memory*, where the *entire history* of states visited contains information on the probability of finding the system in a particular state j at time τ [141]. In Ref. [25], this memory manifested itself by the inability of models constructed from short lag times τ to properly describe dynamics over long times.
Fortunately, due to the presence of a viscous solvent, metastable conformational states, high dimension, and local roughness of the potential energy surface, this memory may appear to be be relatively short, such that the system *behaves* as if it were a Markov process in the discrete state space when observed only at intervals larger than some time τ_{int}. Physically, this may occur if the system spends most of its time trapped in metastable states, punctuated by infrequent transitions to other states, and the imposed decomposition defined by the $\{\chi_i(\boldsymbol{z})\}$ is chosen to be nearly coincident with these true metastable states. The minimal lag time τ_{int} at which dynamics appears to be Markovian for a given state decomposition $\{\chi_i\}$ has been denoted as the *internal equilibration time* [23]. Poor choices of state decomposition will mean the system appears Markovian only on long times. For particularly poor choices of state decomposition, τ_{int} can approach the time for the entire system to relax to equilibrium, τ_{eq}. If τ_{int} exceeds the timescale of the process of interest, then the model will not be useful in studying this process.
Unfortunately, direct calculation of the internal equilibration time τ_{int} is difficult[8]. We are therefore forced to instead examine various *consequences* of this Markovian behavior for which it is easy to determine their validity at various lag times (to within statistical uncertainty) from the same simulation dataset that is used to construct the model. Below, we enumerate a number of tests for Markovian behavior, many of which have been proposed elsewhere [174]. In Section 4.3, we apply these tests to a set of short trajectories of the terminally blocked alanine peptide to assess

[8] It has been suggested that consideration of restrictions of Markov chains within individual states would allow a lower bound to be estimated, in the absence of statistical uncertainty [60, 118]

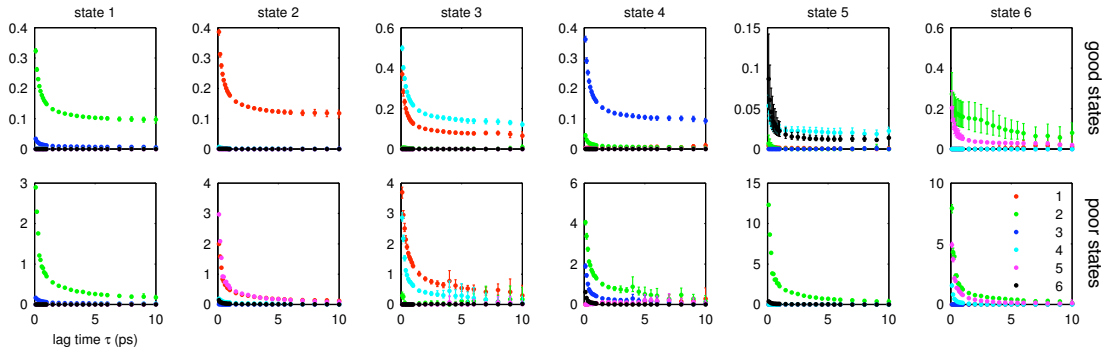


Figure 4.2: **Implied transition rates as a function of lag time.** For each state, the implied transition rates to all other states are shown. Top: Rate matrix elements implied by the observed transition matrix as a function of lag time for the "good" state decomposition. Bottom: Rate matrix elements implied by the observed transition matrix as a function of lag time for the "poor" state decomposition.

their utility in determining the time for emergence of Markovian behavior.

4.4.1 The implied rate matrix.

If the state-to-state dynamics appears Markovian after a time τ_{int}, we expect the relationship

$$\boldsymbol{T}(\tau) = e^{\boldsymbol{K}\tau} \;\; \forall \;\; \tau \geq \tau_{\text{int}} \tag{4.19}$$

where $\boldsymbol{T}(\tau)$ is the observed transition matrix, will hold (neglecting statistical uncertainty).
At short lag times τ, it is known that rate estimates will be too fast; in the limit where $\tau \to 0^+$, we have the same problem as in transition state theory where the neglect of recrossings results in the rate estimate to exceed the phenomenological rate constant [20]. As the lag time increases beyond the time required for the system to equilibrate within the states, the longest of which is termed the *internal equilibration time*, the implied rates will eventually stabilize [23]. However, given an arbitrary coarse-graining, the internal equilibration time may actually exceed the fast timescales in the system. In fact, in particularly poor choices for states, the internal equilibration time may exceed the longest timescales in the system, rendering the model useless in describing the time-evolution of the system [174].
This suggests that a way to determine τ_{int} may be to compute the rate matrix $\boldsymbol{K}$ implied by the observed transition matrix $\boldsymbol{T}(\tau)$ for every lag time τ, and identify the smallest τ after which the implied $\boldsymbol{K}$ appears to be invariant. At some sufficient lag time τ_{eq}, however, these rates should

stabilize and a Markov description will be appropriate for all times t larger than τ_{eq} [174]. It is hoped that this time will be short compared to the relaxation times of the slow processes of interest in the system, such as protein folding.

Examination of Eqs. 4.8 and 4.9, shows that the rate matrix can be computed from a transition matrix $\boldsymbol{T}(\tau)$ that satisfies detailed balance by

$$\boldsymbol{K} \;=\; \boldsymbol{U}(\tau^{-1}\log \boldsymbol{M})\boldsymbol{U}^{-1} \tag{4.20}$$

where $\boldsymbol{U}$ is the matrix of eigenvectors of $\boldsymbol{T}(\tau)$ and $\boldsymbol{M}$ the diagonal matrix of eigenvalues. Unfortunately, due to statistical uncertainty in the observed transition matrix $\boldsymbol{T}(\tau)$, two things may occur that result in a rate matrix that does not satisfy the conditions enumerated in Section 4.2.1: (1) some of the rates $K_{ji},\ j \neq i$ may be negative; (2) some of the eigenvalues of the observed transition matrix μ_k may be negative. As a result of (2), some elements of the resulting rate matrix are complex; here, we only consider the real part. To ensure that the conditions of nonnegativity and reality of rates are met, one may use more sophisticated methods to recover rate matrices that only approximately satisfy Eq. 4.19 yet satisfy these conditions [74, 167].

We computed the implied rate matrix as a function of lag time from the observed transition matrices using Eq. 4.20 for the 10 ps shooting trajectories. The implied rates for good and poor partitionings are shown in Figure 4.2. As expected, the rates approach their transition state theory estimates as the lag time τ approaches zero and decay as the lag time increases. For the good state decomposition, the rates appear to approach a constant value (to within statistical uncertainty) at lag times around 6 ps. State 6 is clearly the most problematic, with the uncertainties for transition to state 4 being especially large and only converging to a constant value very late. This may suggest that this state in particular could benefit from refinement of the state boundaries. The poor decomposition has much faster rates, consistent with a model with shorter aggregate timescales. It is not evident if the rates converge over 10 ps — they appear to be still diminishing near lag times of 10 ps, though at a slow rate.

4.4.2 Eigenvalues of the implied rate matrix.

As the number of states, N, increases, the number of rate matrix elements increases as N^2. Examination of all of these for signs of convergence may be impractical. Additionally, we are generally concerned with only the *slowest* processes in the system. Note that the time evolution of

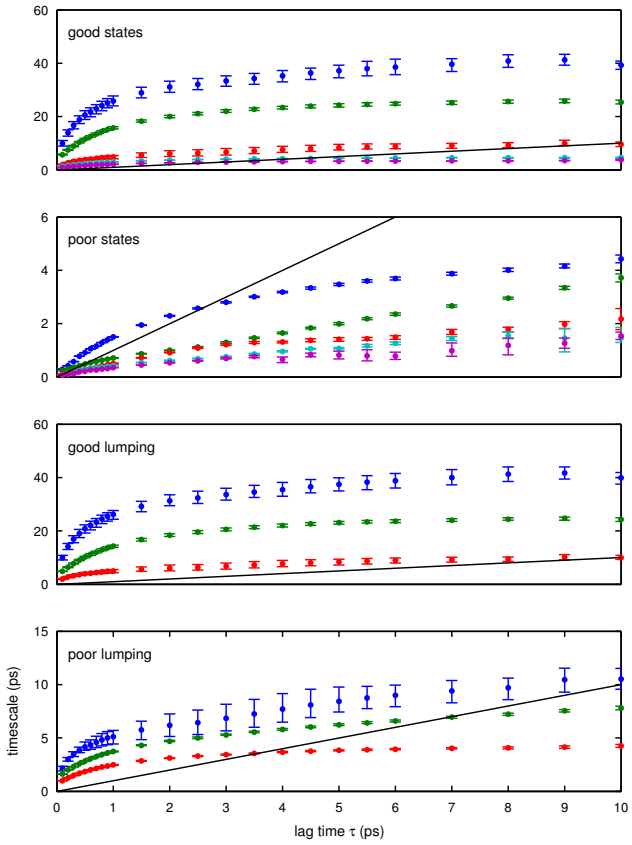


Figure 4.3: **Implied timescales of the rate matrix as a function of lag time.** Implied timescales and associated uncertainties are shown for good decomposition (top), poor decomposition (middle upper), good lumping (middle lower), and poor lumping (lower). Timescales are colored by order from longest (blue) to shortest (purple). The black line denotes $\tau_k = \tau$, such that processes whose timescales fall below this line occur on times shorter than the lag time.

an observable $A(\mathbf{q})$ can be written in terms of the eigenvector expansion (Eq. 4.9) as

$$\begin{aligned} \langle A \rangle(t) &= \sum_{k=1}^{N} (\boldsymbol{A}^{\mathrm{T}} \boldsymbol{D}^{1/2} \tilde{\boldsymbol{u}}_k) \, (\tilde{\boldsymbol{u}}_k \boldsymbol{D}^{1/2} \boldsymbol{p}(0)) \, e^{\lambda_k t} \\ &= \sum_{k=1}^{N} c_k \, e^{-t/\tau_k} \end{aligned} \tag{4.21}$$

where $\boldsymbol{A}$ is the vector of characteristic spectroscopic observables over each individual state, τ_k is a characteristic timescale for the process governed by eigenvector $\boldsymbol{u}_k$, and c_k is its (potentially negative) amplitude. It may be that some non-Markovity is tolerated if it only disrupts the fastest processes in the system but leaves the slowest ones, which may serve to determine the behavior of $\langle A \rangle(t)$ over the timescales of interest, unaffected. In fact, if the model is only Markovian on timescales greater than τ_{int}, processes with characteristic timescales $\tau_k < \tau_{\mathrm{int}}$ will not correctly be described by the model anyway. Additionally, because we only extract $N - 1$ values from the transition matrix instead of N^2, these values may be better determined statistically than the individual rates.

As a result, Swope *et al.* [174] suggested that the implied timescales of the rate matrix, computed from the observed transition matrix as a function of lag time, could be used to determine the onset of Markovian behavior. These timescales are determined from the eigenvalues λ_k of the rate matrix, which can in turn be determined directly from the eigenvalues of the transition matrix through Eq. 4.8:

$$\tau_k = -\lambda_k^{-1} = -\tau [\log \mu_k]^{-1} \ , \ \ k = 2, \ldots, N \tag{4.22}$$

If the elements of the implied rate matrix converge to some fixed value for lag times $\tau \geq \tau_{\mathrm{int}}$, then these implied timescales will also converge. While the converse is not true, we may be unlikely to encounter a case where the timescales have converged but many of the individual rates are still changing significantly.

We computed the timescales as a function of lag time for the 10 ps shooting trajectories for all four state decompositions, and depicted in Figure 4.3. The timescales of the good state decomposition all appear to fluctuate about a constant value (to within statistical uncertainty) at lag times of 6 ps or greater, where three of the timescales are larger than the lag time at this point. By contrast, the slowest timescales for the poor decomposition do not appear to stabilize within 6 ps, and all processes appear *faster* than the lag time after lag times of 3 ps. The good lumping does not appear to disrupt the longest timescales — three timescales greater than the lag time are again observed at

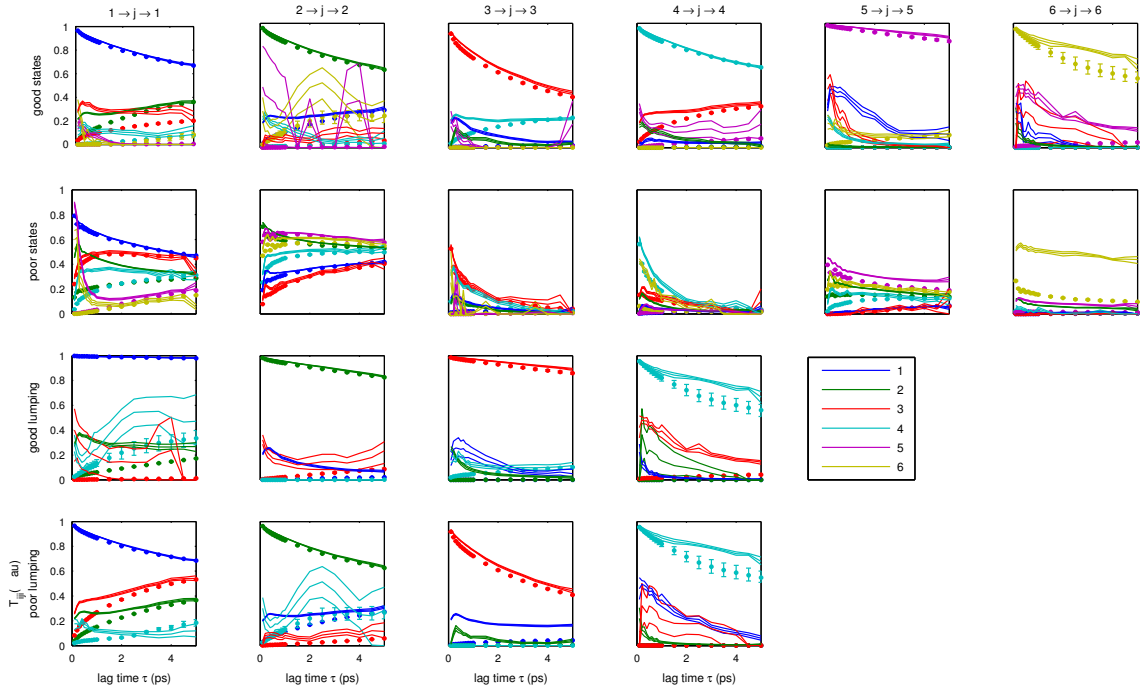


Figure 4.4: **Second-order transition probabilities compared with products of first-order transition probabilities as a function of lag time.** Observed second-order transition probabilities $T_{k|ji}(\tau)$ are shown as solid lines with the envelope indicating a 68% confidence interval, while first-order transition probabilities are shown as points with error bars. Because there are $6^3 = 216$ second-order transition probabilities, and many of them are not well estimated by the limited dataset, only the 36 $T_{i|ji}(\tau)$ transition probabilities are shown. Each plot shows 6 transition probabilities originating from a single state i, occupying a state j at time τ, and returning to i at time 2τ.

a lag time of 6 ps. The poor lumping, on the other hand, diminished the timescales such that by 9 ps, when it appears the timescales *may* have reached a plateau, only one timescale is barely longer than the lag time.

4.4.3 Second-order transition probabilities.

A more direct test for the effects of history dependence is the examination of second-order transition probabilities $T_{k|ji}(\tau)$, which denotes the probability of observing the system in state k at time 2τ given that it was initially in j at time τ and i at time 0 [174]. These transition probabilities

can be estimated from three-time correlation functions $C_{kji}(\tau) = \langle \chi_k(2\tau)\chi_j(\tau)\chi_i(0)\rangle$ by

$$T_{k|ji}(\tau) \quad = \quad \frac{C_{kji}(\tau)}{\sum_{k'=1}^{M} C_{k'ji}(\tau)}. \tag{4.23}$$

If the Markovian property holds at lag time τ, the probability of observing a transition from state j to state k should be independent of the previous state i:

$$T_{k|ji}(\tau) = T_{kj}(\tau) \ \forall \ \tau \geq \tau_{\text{int}}. \tag{4.24}$$

Ideally, we would verify that all observed second-order transition probabilities are equal to the corresponding first-order transition probabilities to within statistical uncertainty.
Second-order transition matrices were computed from correlation functions $C_{kji}(\tau)$ estimated from the shooting data using Eq. 4.23 above, and these are shown in Figure 4.4 together with their corresponding first-order transition probabilities estimated from the same dataset. Because there are $N^3 = 216$ possible three-time correlation functions, and because the observed data for transitions where $i \neq j \neq k$ will be sparse, we choose to only examine those where $k = i$. One obvious limitation of this approach is the requirement that the trajectories be at least 2τ in length, so we are not even able to reach the 6 ps lag time required for Markovianity. It is also evident that there is a large amount of statistical uncertainty in the resulting second-order transition matrix, so much so that it is difficult to discern whether disagreement is meaningful, as in the $2 \rightarrow j \rightarrow 2$ plot for the good partitioning.

4.4.4 Chapman-Kolmogorov equation.

Due to the large statistical uncertainty in the three-time correlation functions $C_{kji}(\tau)$ used to compute the second-order transition probabilities, and the N^3-dependence of $T_{k|ji}(\tau)$ on the number of cells, it may be useful to instead consider whether the Chapman-Kolmogorov equation is satisfied at some minimum τ:

$$\boldsymbol{T}(2\tau) \quad = \quad [\boldsymbol{T}(\tau)]^2 \tag{4.25}$$

Here, only N^2 elements need be compared, and intermediate states are effectively ummed over. Figure 4.5 shows a comparison of the left-hand side and right-hand side estimated independently from the 10 ps shooting trajectories. It is again apparent that this metric, too, requires trajectories of minimal length 2τ to test the satisfaction of Eq. 4.25 for a lag time of τ. Somewhat surprising is

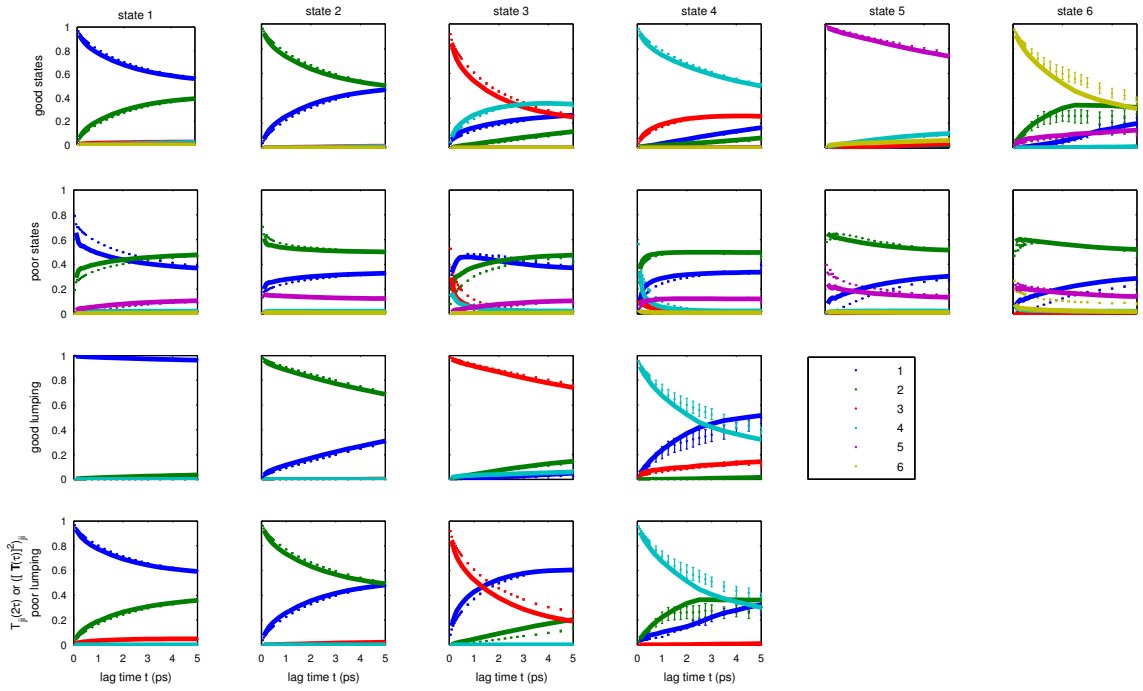


Figure 4.5: **Test of the Chapman-Kolmorov equation.** Observed transition probabilities $T_{ji}(2\tau)$ are shown as points with error bars, while predicted transition probabilities from $([\boldsymbol{T}(\tau)]^2)_{ji}$ are shown as lines enveloping a 68% confidence interval.

the observation that, for the good partitioning, the Chapman-Kolmogorov property appears to be satisfied within statistical uncertainty at very *early* times, up to perhaps 1 ps. This is counterintuitive, as it is clear that models constructed from such short lag times poorly reproduce dynamics [25]. It is hard to determine whether the condition is satisfied at longer times, as state 6 appears particularly problematic in the good decomposition. Surprisingly, the poor decomposition does not appear much worse than the good decomposition over the span of times considered, though the poor lumping shows more systematic deviation than the good lumping throughout lumped state 3.

4.4.5 Discrete lifetime distribution.

For a discrete- or continuous-time master equation model, if the system is initially in state i at time $t = 0$, the probability that it will be observed to remain in state i for exactly $L - 1$ additional consecutive observations with observation interval τ is given by the geometric distribution [174], with probability mass function

$$P(L;\tau) \quad = \quad (1 - T_{ii}(\tau))[T_{ii}(\tau)]^{L-1}\,, L \in \{1, 2, \ldots, \}. \tag{4.26}$$

and cumulative distribution function

$$P(L \leq K;\tau) \quad = \quad 1 - [T_{ii}(\tau)]^{K}. \tag{4.27}$$

When $\log P(L;\tau)$ is plotted as a function of L for fixed lag time τ, the trend should be linear for $\tau \geq \tau_{\mathrm{eq}}$ with slope $\log T_{ii}(\tau)$, though its behavior may differ at short times before Markovianity has been achieved.

Figure 4.6 shows the observed $\log P(L;\tau_{\mathrm{sample}})$ for each state, determined by constructing a histogram of the number of successive snapshots the system is observed to remain in state i, along with the geometric distribution with the same mean lifetime. At large L, all states show geometric lifetime distributions, but at short times, more complex behavior is observed. States 5 and 6 of the good manual decomposition, for example, appear to exhibit two distinct geometric phases. In all cases, the onset of the final linear behavior is very early — within enough sampling intervals such that the remainder of the plot is linear for $L\tau_{\mathrm{sample}} > 2$ ps. This is, of course, too early for the onset of Markovian prediction. It is possible that examination of the complete set of plots for various sampling intervals τ would allow one to better estimate the onset of Markovian behavior, but this may be difficult, as the number of these grows as $NT/\tau_{[}\mathrm{samplet}$.

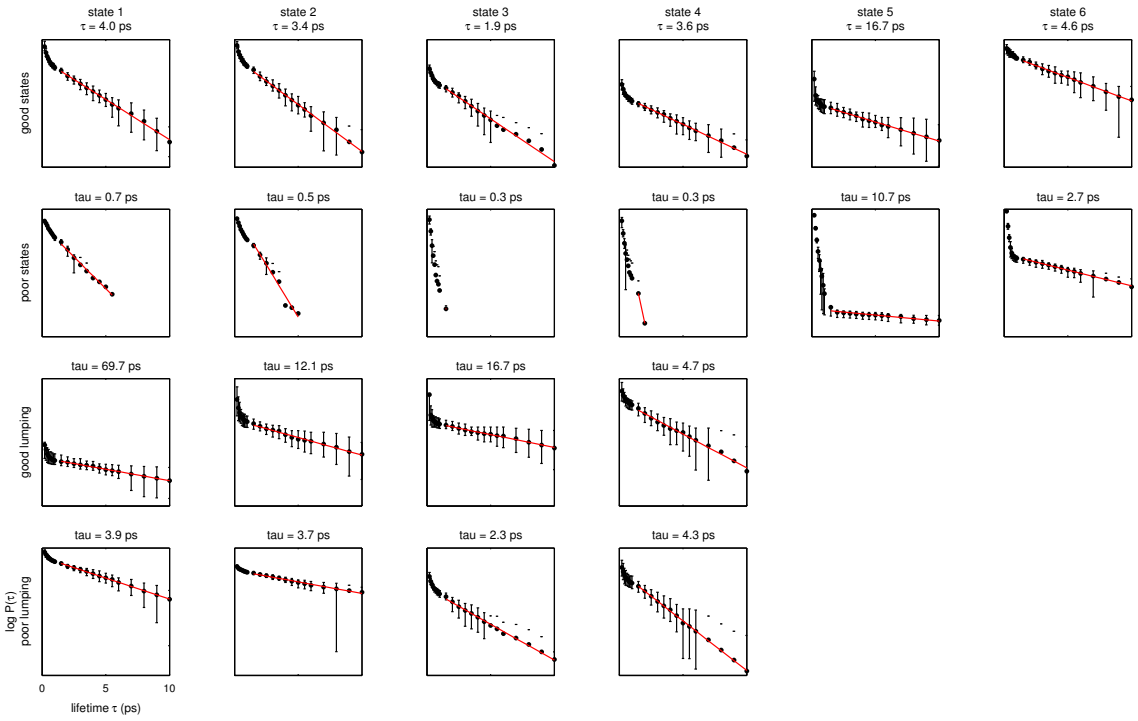


Figure 4.6: **Observed and geometric discrete lifetime probability distributions.** The logarithm of the discrete lifetime probability mass function $P(L;\tau)$ for each state, which indicate the probability of observing the system to remain within a given state for exactly a a number of consecutive observation intervals L, is shown as a function of $L\tau$ along with the characteristic lifetimes obtained from a linear fit to the log probability over the interval $L\tau \in [1.5, 10]$ ps, evaluated with sampling interval τ_{sample} = 0.1 ps. The number of intervals L for which the system remains in one state has been multiplied by the sampling interval τ so that the x-axis appears in units of time. Top: good states; middle top: poor states; middle bottom: good lumping; bottom: poor lumping.

4.5 Discussion and Conclusions

Here, we have examined the issue of how one may verify whether a Markov chain or master equation model constructed from a set of short trajectories will be able to reproduce dynamics over long times without necessitating comparison with a separate set of long trajectories. Due to the difficulty of computing the internal equilibration time τ_{int}, the time required for the system to lose memory within the states, we considered a number of tests of Markovianity based on determining whether conditions corresponding to consequences of Markovian behavior were satisfied to within our ability to resolve by statistical uncertainty. Not all of these tests were of equal utility. Some tests, such as comparison of second-order transition probabilities with first-order transition probabilities, were too noisy to provide useful information directly. Summation over intermediate states to give a test of the Chapman-Kolmogorov equation reduced the statistical uncertainty, but this condition was met to within statistical uncertainty at lag times that were too short to accurately reproduce dynamics. Examination of discrete lifetime distributions revealed clear non-Markovian behavior at short times, but this test too gave indication that Markovian behavior was achieved at lag times that were too short. Examination of the implied rate matrix elements or timescales was informative of the time of emergence of Markovian behavior, but as the number of rate matrix elements increases as N^2, and some non-Markovitianity might be tolerated if the slowest processes are preserved, examination of the slowest timescales was deemed the most useful of these metrics for verifying Markov behavior.

Many challenges to efficiently constructing useful master equation or Markov models of protein dynamics remain. The selection of an appropriate state space, a topic addressed in Ref. [23], is perhaps the most difficult of these. Even given an optimal decomposition into a desired number of states for a desired number of states, whether the timescale on which this model becomes Markovian is accessible and the quantity of simulation data needed to construct accurate models is not know. Additionally, efficient methods for computing interstate transition rates are needed. The success of these models in accurately describing the dynamics of small solvated peptides is encouraging, but their utility in characterizing complex biomolecular dynamics of large macromolecules remains to be seen.

4.6 Acknowledgements

The authors happily thank Nina Singhal, Sanghyun Park, Vijay Pande, and Hans Andersen (Stanford) for many enlightening discussions, and Libusha Kelly (UCSF) for critical comments on this manuscript. JDC was supported by Howard Hughes Medical Institute and IBM predoctoral fellowships. WS acknowledges support from NSF MRSEC Center on Polymer Interfaces and Macromolecular Assemblies DMR – 0213618, and KD the support of NIH grant GM34993.

Chapter 5

An automatic state decomposition algorithm

The material in this chapter will be submitted to the *Journal of Chemical Physics* for publication as

Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics

John D. Chodera[+1], Nina Singhal[+2], Vijay S. Pande[3], Ken A. Dill[4], and William C. Swope[51]

[1]*Graduate Group in Biophysics and* [4] *Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143*

[2] *Department of Computer Science and* [3] *Department of Chemistry, Stanford University, Stanford, CA 94305*

[5] *IBM Almaden Research Center, 650 Harry Road, San Jose, CA 95120*

[+] *These authors contributed equally to this work.*

Abstract

To meet the challenge of modeling the conformational dynamics of biological macromolecules over long timescales, much recent effort has been devoted to constructing stochastic kinetic models, often in the form of discrete-state Markov models, from short molecular dynamics simulations. To construct useful models that faithfully represent dynamics at the timescales of

[1]Author to whom correspondence should be addressed: William C. Swope <swope@us.ibm.com>

interest, it is necessary to decompose configuration space into a set of kinetically metastable states. Previous attempts to define these states have relied upon either prior knowledge of the slow degrees of freedom or on the application of conformational clustering techniques which assume that conformationally distinct clusters are also kinetically distinct. Here, we present a first version of an *automatic* algorithm for the discovery of kinetically metastable states that is generally applicable to solvated macromolecules. Given molecular dynamics trajectories initiated from a well-defined starting distribution, the algorithm discovers long-lived, kinetically metastable states through successive iterations of partitioning and aggregating conformation space into kinetically related regions. We apply this method to three peptides in explicit solvent — terminally blocked alanine, the engineered 12-residue β-hairpin trpzip2, and the 21-residue helical F_s peptide — to assess its ability to generate physically meaningful states and faithful kinetic models.

5.1 Introduction

Many biomolecular processes are fundamentally dynamic in nature. Protein folding, for example, involves the ordering of a polypeptide chain into a particular topology over the course of microseconds to seconds, a process which can go awry and lead to misfolding or aggregation, causing disease [43]. Enzymatic catalysis may involve transitions between multiple conformational substates, only some of which may allow substrate access or catalysis [15, 50, 200]. Post-translational modification events, ligand binding, or catalytic events may alter the transition kinetics among multiple conformational states by modulating catalytic function, allowing work to be performed, or transducing a signal through allosteric change [21, 65, 114]. A purely static description of these processes is insufficient for mechanistic understanding — the dynamical nature of these events must be accounted for as well.

Unfortunately, these processes may involve molecular timescales of microseconds or longer, placing them well outside the range of typical detailed atomistic simulations employing explicit models of solvent. Many of these systems are very large, limiting the length of trajectories that can be generated by molecular dynamics simulation. However, due to the presence of many energetic barriers on the order of the thermal energy, the uncertainty in initial microscopic conditions, and the stochasticity introduced into the system by the surrounding solvent in contact with a heat bath, any suitable description of conformational dynamics must *by necessity* be statistical in nature. This has motivated the development of stochastic kinetic models of macromolecular dynamics which might conceivably be constructed from short dynamics simulations, yet provide a useful and

accurate statistical description of dynamical evolution over long times.

Several approaches have been used to construct of these models. *Transition interface sampling* (TIS) [123], *milestoning* [54], and methods based on commitment probability distributions [11, 144] attempt to describe dynamics along a one dimensional reaction coordinate, but these approaches are valid only if an appropriate reaction coordinate can be identified such that relaxation transverse to this coordinate is fast compared to diffusion along it. Discrete-state, continuous-time master equation models, characterized by a matrix of phenomenological rate constants describing the rate of interconversion between states [190], can be constructed by identifying local potential energy minima as states and estimating interstate transition rates by transition state theory [8, 32, 52, 101, 105, 124, 125]. Unfortunately, the number of minima, and hence the number of states, grows exponentially with system size, making the procedure prohibitively expensive for larger proteins or systems containing explicit solvent molecules. Others have suggested that stochastic models of dynamics can be constructed by expansion of the appropriate dynamical operator in a basis set [156, 157, 186], but this approach appears to be limited by the great difficulty of choosing rapidly-convergent basis sets for large molecules, a process that is not fundamentally different from identifying the slow degrees of freedom.

Instead, much work has focused on the construction of discrete- or continuous-time Markov models to describe dynamics among a small number of states which may each contain many minima within large regions of configuration space [2, 35, 51, 74, 141, 151, 161, 162, 164, 167, 175]. In these models, it is hoped that a separation of timescales between fast *intrastate* motion and slow *interstate* motion allows the statistical dynamics to be modeled by stochastic transitions among the discrete set of metastable conformational states governed by first-order kinetics. Such a separation of timescales would be a natural consequence of the widely held belief that the nature of the energy landscape of biomacromolecules is hierarchical [4, 7, 9, 105, 106]. If the system reaches local equilibrium *within* the state before attempting to exit, the probability of transitioning to any other state will be independent of all but the current state. This allows the process to be modeled with either a discrete-time Markov chain (e.g. Ref. [162]) or a continuous-time master equation model with coarse-grained time (e.g. Ref. [167]). In either model, processes occurring on timescales faster than a coarse-graining time, determined by the time to reach equilibrium within each state, cannot be resolved.

Markov models embody a concise description of the various kinetic pathways and their relative likelihood, facilitating comparison with experimental data and providing a powerful tool for mechanistic insight. Once the model is constructed and the timescale for Markovian behavior

determined, it can be used to compute the stochastic temporal evolution of either a single macromolecule or a population of noninteracting macromolecules, allowing direct comparison of simulated and experimental observables for both single-molecule or ensemble kinetics experiments. In addition, useful properties difficult to access experimentally, such as state lifetimes [174], relaxation from experimentally inaccessible prepared states [25], mean first-passage times [162], the existence of hidden intermediates [137], and P_{fold} values or transmission coefficients [103], can easily be obtained. This allows for both a thorough understanding of mechanism and the generation of new, experimentally testable hypotheses.

To build such a model, it is necessary to decompose configuration space into an appropriate set of metastable states. If the low-dimensional manifold containing all the slow degrees of freedom is known a priori, then this can be partitioned into free energy basins to define the states, such as by examination of the potential of mean force [25, 51, 164, 167, 175]. In the absence of this knowledge, others have turned to conformational clustering techniques to identify conformationally distinct regions which may also be kinetically distinct [2, 35, 92, 162].

Instead, we adopt a strategy first suggested for the discovery of metastable states in biomolecular systems by researchers at the *Konrad-Zuse-Zentrum für Informationstechnik* [152]. The principal idea is this: If configuration space could be decomposed into a large number of small cells, the probability of transitioning between these cells in a fixed evolution time could be measured. This probability is a measure of *kinetic connectivity* among the cells, which allows the identification of aggregates of these cells that approximate true metastable states [153]. Unfortunately, the choice of how to divide configuration space into cells is not straightforward. Suppose one is to consider the analysis of some fixed amount of simulation data. If configuration space is decomposed very finely, the boundaries between metastable states can in principle be well-approximated, but the estimated cell-to-cell transition probabilities will become statistically unreliable. On the other hand, if configuration space is decomposed too coarsely, the transition probabilities may be well-determined, but the boundaries between metastable states cannot be clearly resolved, potentially disrupting or destroying the Markovian behavior of interstate dynamics. An optimal choice would ultimately require knowledge of the metastable regions in order to determine the best decomposition of space into cells.

In this work, we propose an iterative procedure to determine both the choice of cells and their aggregates to approximate the desired metastable states. We use a conformational clustering method to carve configuration space into an initial crude set of cells (*splitting*), and a Monte Carlo simulated annealing procedure to collect metastable collections of cells into states (*lumping*). This

cycle is repeated, with the splitting procedure now applied individually to each state to generate a new set of cells, and the lumping procedure applied to the entire set of cells to redefine states until further application of this procedure leaves the approximations to metastable states unchanged. This procedure allows state boundaries to be iteratively refined, as regions that mistakenly have been included in one state can be split off and regrouped with the proper state. Throughout this process, we require that the cells never become so small that estimation of the relevent transition matrix elements is statistically unreliable. Our proposed method is efficient, of $\mathcal{O}(N)$ complexity in the number of stored configurations, and can be easily parallelized.

This paper is organized as follows: In Section 5.2, we give an overview of the Markov chain model and its construction, elaborate on desirable properties of an algorithm to partition configuration space into states, and outline the principles underlying the algorithm we present here. In Section 5.3, we provide a detailed description of the automatic state decomposition algorithm and its implementation. In Section 5.4, we apply this algorithm to three model peptide systems in explicit solvent to assess its performance: alanine dipeptide, the 12-residue engineered trpzip2 hairpin, and the 21-residue F_s helix-forming peptide. Finally, in Section 5.5, we discuss the advantages and shortcomings of our algorithm, with the hope that future state decomposition algorithms can address the remaining challenges.

5.2 Theory

Some discussion of the stochastic model of kinetics considered here and the theory underlying the method is appropriate before describing the algorithmic implementation in detail. First, in Section 5.2.1, we review Markov chain and master equation models of conformational dynamics. Next, in Section 5.2.2, we describe their construction from equilibrium molecular dynamics trajectories given any state partitioning. Section 5.2.3 enumerates a number of requirements for a useful state partitioning. Finally, Section 5.2.4 discusses possible methods for validating a given state decomposition. The actual implementation of the algorithm used here is described in detail in Section 5.3.

5.2.1 Markov chain and master equation models of conformational dynamics.

Consider the dynamics of a macromolecule immersed in solvent, where the solvent is at equilibrium at some particular temperature of interest. We presume that all of configuration space

has already been decomposed into a set of nonoverlapping regions, or *states*, which together form a complete decomposition of configuration space. The method by which these states are identified is described in subsequent sections.

If we observe the evolution of this system at times $t = 0, \tau, 2\tau, \ldots$, where τ denotes the observation interval, we can represent this sequence of observations in terms of the state the system visits at each of these discrete times. The sequence of states produced is a realization of a *discrete-time stochastic process*. For this process to be described by a Markov chain, it must satisfy the *Markov property*, whereby the probability of observing the system in any state in the sequence is independent of all but the previous state. For a stationary process on a finite set of L states, this process can be completely characterized by an $L \times L$ *transition matrix*[2] $\boldsymbol{T}(\tau)$ dependent only on the observation interval, or *lag time*, τ. The element $T_{ji}(\tau)$ denotes the probability of observing the system in state j at time t given that it was previously in state i at time $t - \tau$. If this process satisfies detailed balance (which we will assume to be the case for physical systems of the sort we consider here [190]) we additionally have the requirement

$$T_{ji} p_{\mathrm{eq},i} = T_{ij} p_{\mathrm{eq},j} \tag{5.1}$$

where $p_{\mathrm{eq},i}$ denotes the equilibrium probability of state i.

The vector of probabilities of occupying any of the L states at time t (here also referred to as the vector of state populations, such as in an experiment involving a population of noninteracting macromolecules) can be written as $\boldsymbol{p}(t)$. If the initial probability vector is given by $\boldsymbol{p}(0)$, we can write the probability vector at some later time $t = n\tau$ as

$$\boldsymbol{p}(n\tau) \;=\; \boldsymbol{T}(n\tau)\boldsymbol{p}(0) \;=\; [\boldsymbol{T}(\tau)]^n \boldsymbol{p}(0). \tag{5.2}$$

This is a form of the *Chapman-Kolmogorov equation*.

Alternatively, the process can be characterized in *continuous* time by a matrix of phenomenological rate constants $\boldsymbol{K}$, where the element K_{ji}, $j \neq i$ denotes the nonnegative phenomenological rate from state i to state j. The diagonal elements are determined by $K_{ii} = -\sum_{j \neq i} K_{ji}$ to ensure the columns sum to zero so as to conserve probability mass. Time evolution is then governed by the equation

$$\dot{\boldsymbol{p}}(t) \;\;=\;\; \boldsymbol{K}\boldsymbol{p}(t) \tag{5.3}$$

[2] We adopt the notation for a *column-stochastic* transition matrix, in which the columns sum to unity. This differs from the notation in some previously-cited references, which use a *row stochastic* transition matrix, equal to the transpose of the column stochastic matrix used here.

where the dot represents differentiation with respect to time. This evolution equation has formal solution

$$\boldsymbol{p}(t) \quad = \quad e^{\boldsymbol{K}t}\boldsymbol{p}(0), \tag{5.4}$$

where the exponential denotes the formal matrix exponential. Eq. 5.3 is often referred to as a *master equation* [135, 190] describing evolution among a discrete set of states in continuous time. It is important to note that, despite the fact that $\boldsymbol{p}(t)$ is formally defined for all times t, we do not expect Eq. 5.4 to hold for *all* times t for physical systems of the sort we consider here. In particular, for states of finite extent in configuration space, there exists a corresponding limit for the time resolution for which dynamics will appear Markovian; processes that occur on timescales shorter than this will be be incorrectly described by the master equation. We will return to this topic in detail in subsequent sections.

There is an obvious relationship between the transition matrix $\boldsymbol{T}(\tau)$ and the rate matrix $\boldsymbol{K}$ evident from comparison of Eqs. 5.2 and 5.4:

$$\boldsymbol{T}(\tau) \quad = \quad e^{\boldsymbol{K}\tau}. \tag{5.5}$$

If the process can be described by a continuous-time Markov process at all times, then this process can be equivalently described at discrete time intervals by the corresponding transition matrix. The converse may not always be true due to sampling errors in $\boldsymbol{T}(\tau)$, though methods exist to recover rate matrices $\boldsymbol{K}$ consistent with the observed data and the requirements of detailed balance and nonnegativity rates [74, 167].

The transition and rate matrices have eigenvalues $\mu_k(\tau)$ and λ_k, respectively, and share corresponding right eigenvectors $\boldsymbol{u}_k$. The detailed balance requirement additionally ensures that all eigenvalues are real, and we here presume them to be sorted in descending order. $\mu_k(\tau)$ and λ_k are related by

$$\mu_k(\tau) \quad = \quad e^{\lambda_k \tau}. \tag{5.6}$$

The eigenvalues each imply a timescale corresponding to an inverse aggregate rate

$$\tau_k = -\lambda_k^{-1} = -\tau[\ln \mu_k(\tau)]^{-1} \tag{5.7}$$

and the associated eigenvector gives information about the aggregate conformational transitions that are associated with this timescale [83, 152, 154, 155]. In particular, the components of $\boldsymbol{u}_k$ sum

to zero for each $k \geq 2$, and the aggregate dynamical mode can be identified with transitions from microstates with positive eigenvector components interconverting with the set of microstates with negative components, and vice-versa, with the degree of participation in the mode governed by the magnitude of the eigenvector component. This fact can be useful in deducing the conformational transitions among aggregated regions of configuration space that govern relaxation to equilibrium, which is achieved once all processes have exponentially damped out.

For the remainder of this manuscript, we will refer exclusively to the discrete-time Markov chain model picture without loss of generality (Eq. 5.2), except for use of the timescales implied by the transition matrix, as described above.

5.2.2 Construction from simulation data given a state partitioning.

Once a statistical-mechanical ensemble describing equilibrium and a microscopic model describing dynamical evolution in phase space have been selected, the transition matrix $\boldsymbol{T}(\tau)$ can be estimated from molecular dynamics simulations. For a system in which dynamical evolution is Newtonian and, at equilibrium, configurations are distributed according to a canonical distribution at a given temperature, Swope *et al.* [174] show that the transition probability $T_{ji}(\tau)$ can be written as the following ratio of canonical ensemble averages:

$$T_{ji}(\tau) = \frac{\int d\boldsymbol{z}(0)\, e^{-\beta H(\boldsymbol{z}(0))}\, \chi_j(\boldsymbol{z}(\tau))\, \chi_i(\boldsymbol{z}(0))}{\int d\boldsymbol{z}(0)\, e^{-\beta H(\boldsymbol{z}(0))}\, \chi_i(\boldsymbol{z}(0))} \tag{5.8}$$

$$= \frac{\langle \chi_j(\tau) \chi_i(0) \rangle}{\langle \chi_i \rangle} \tag{5.9}$$

where $\boldsymbol{z(t)}$ denotes a point in phase space visited by a trajectory at time t, $\chi_i(\boldsymbol{z})$ denotes the indicator function for state i (which assumes a value of unity if $\boldsymbol{z}$ is in state i, and zero otherwise), $\beta \equiv (k_B T)^{-1}$ the inverse temperature, $H(\boldsymbol{z})$ the Hamiltonian, and $\langle A \rangle$ the canonical ensemble expectation of a phase function $A(\boldsymbol{z})$ at inverse temperature β.

Given a set of simulations initiated from an equilibrium distribution, the expectations in Eq. 5.9 can be computed independently by standard analysis methods [1]. Estimation of the correlation function in the numerator can make use of both the stationarity of an equilibrium distribution (by considering overlapping intervals of time τ), and the microscopic reversibility (by considering also time-reversed versions of the simulations) of Newtonian trajectories. Alternatively, if an equilibrium distribution within each state can be prepared, one can also directly estimate a column of transition matrix elements by computing the fraction of trajectories initially at equilibrium within state i that terminate in state j a time τ later. More elaborate methods based on equilibrium

ensembles prepared within special *selection cells* that are not coincident with the states [174, 175] or *partition of unity* restraints [194] can also be used to compute transition matrix elements efficiently.

5.2.3 Requirements for a useful Markov model.

For any given state partitioning, the dynamics of the system will be Markovian on some time scale. For example, if the lag time τ is so long as to approach the time for the system to relax to an equilibrium distribution from any arbitrary nonequilibrium starting distribution, a single application of the transition matrix $\boldsymbol{T}(\tau)$ carries any arbitrary initial probability distribution directly to the invariant equilibrium distribution. However, if this τ exceeds the timescale of the process of interest, our model is not useful[3] for describing it, and therefore it is advantageous to attempt to find a state decomposition that is Markovian on a shorter timescale in order to extract useful dynamical information about this process.

For a given state i, we will define its internal equilibration time, $\tau_{\mathrm{int},i}$, as the characteristic time one must wait before the system, initially in a configuration within state i, generates a new *uncorrelated* configuration within the state by dynamical evolution. This internal equilibration time, or *memory time*, closely related to the molecular relaxation timescale τ_{mol} in Chandler's reactive flux formulation of transition state theory [20], depends, of course, on the choice of state decomposition. We can denote the longest of these times over all states by τ_{int}. This is not to be confused with the time it takes an arbitrary nonequilibrium distribution to relax to global equilibrium, but rather, the minimum lag time for which dynamics will appear to be Markovian using this particular state decomposition. If the lag time is longer than τ_{int}, we will expect the system to have lost memory of its previous location within *any* state it may have been in, either remaining within that state or transitioning to a new one, and for dynamics on this set of states to be independent of history. On the other hand, for lag times shorter than τ_{int}, we cannot guarantee that transition probabilities are independent of history everywhere. This suggests a way in which the utility of various decompositions can be measured. For a fixed number of states, the most useful model will partition configuration space to yield the shortest τ_{int}, as this model can be used to study the widest range of dynamical processes.

In addition to producing transition probabilities that are history independent at a relevant lag time,

[3] Equilibrium probabilities can still be extracted from the stationary eigenvector (the eigenvector of corresponding to an eigenvalue of unity) of such a transition matrix, which may have some utility if one had constructed the transition matrix from trajectories not initiated from distributions at equilibrium globally.

we impose additional conditions on our states to ensure the resulting model also provides physical and chemical insight. Because we are primarily interested in macromolecular dynamical motion such as protein folding, we first require that states be consistent with a chemical intuition for a macromolecular *conformational* substate and, therefore, exist as constructs exclusively in the configuration space of the macromolecule. In solvated systems, we expect relaxation and decorrelation of momenta to be much faster than any of the dynamical behaviors of interest, and so we ignore momenta in defining our states. Furthermore, we presume reorganization of the solvent is faster than processes of interest, and therefore ignore coordinates associated with the surrounding solvent[4].

Also, we seek conformational states that exhibit a *separation of timescales*. If states can be constructed where the timescale for equilibration *within* each state is much shorter than the timescale for transitions *among* the states, we would expect interstate dynamics to be well-modeled by a Markov chain after sufficiently long observation intervals. Consider, for example, the isomerization of butane, which has three main metastable conformational states (*gauche-plus*, *gauche-minus*, and *trans*). At sufficiently low temperature, dynamics is dominated by long dwell times *within* each of these three states, punctuated by infrequent transitions between them. The slow interstate transition process is well-described by first order reaction kinetics for observation intervals longer than the fast molecular relaxation time for intrastate dynamics due to the presence of a separation of timescales [20].

In order for the states to be defined such that equilibration within a state is rapid, we further require that the region of configuration space defining each state be *compact* and *connected*. A state composed of two or more unconnected regions of phase space defies the assumption that equilibration within the state is much faster than the characteristic time to leave it.

5.2.4 Validation of Markov models.

Once a decomposition of configuration space is chosen, we are faced with the task of determining the observation time interval τ at which dynamics in this state space appears Markovian. Unfortunately, we cannot directly compute the internal macrostate equilibration times, though examination of the eigenvalues of the transition matrix restricted to a state may give a lower bound on this time in the absence of statistical uncertainty [118]. The most rigorous test for Markovian

[4]We recognize that solvent coordinates may be critical in some phenomena, but dealing with solvent degrees of freedom would also require accounting for the indistinguishability of solvent molecules upon their exchange. We leave this to further iterations of the algorithm.

behavior would be a direct test for history independence. The simplest test of this type is to compute second order transition probabilities and compare them to the appropriate products of the first order transition probabilities to see if their disagreement is statistically significant, though this would miss possible yet unlikely higher order history dependencies. While it is possible to estimate these from the simulation data, this requires the estimation of three-time correlation functions, which often possess statistical uncertainties so large as to render them useless for this kind of test [24].

Raising the transition matrix to a power n (hence summing over the intermediate states) and comparing with the observed transition probabilities for a lag time of $n\tau$, such that one is effectively determining whether the Chapman-Kolmogorov equation (Eq. 5.2) is satisfied, helps to reduce the uncertainty so that the test becomes practical. This is equivalent to propagating the population in time out of a probability distribution confined to each state i initially, and comparing the model evolution with the observed transition probabilities over times much longer than τ_{int}. This serves as a check to ensure that the model is at least consistent with the dataset from which it was constructed, to within the statistical uncertainty of the transition matrices obtained from the dataset. This method was employed, for example, in Refs. [25, 174].

Another approach, from Park, *et al.*, [141] uses concepts from information theory to compute the *conditional mutual information* conveyed by the second-to-last state, which quantifies the discrepancy between observed second-order transition probabilities and the estimate modeled from first-order transition probabilities. The result of this analysis is a scalar that quantifies the degree of history dependence. For a pure first-order Markov process, the mutual information will be zero, as no additional information is gained by including additional history. While this method also requires computing three-time correlation functions, which may individually have substantial uncertainties, the weighted combination of these into a single value reduces the uncertainty in the resulting metric. Unfortunately, there is no rigorous criteria for how small this measure must be in order for the model to be considered acceptably Markovian.

Swope, *et al.*, [174] suggested a number of additional tests for signatures of Markov behavior, the most sensitive of which appears to be examining the behavior of the *implied timescales* of the transition matrix $\boldsymbol{T}(\tau)$, which can be computed from the eigenvalues of the transition matrix by Eq. 5.7, as a function of increasing lag time τ [24]. At sufficiently large τ, the implied timescales will be independent of τ, implying that exponentiation of the transition matrix is nearly identical to constructing the transition matrix using longer observation time intervals (Eq. 5.2). The shortest observation time interval for which this holds can be correlated with the internal equilibration time

τ_{int}, and descriptions of the behavior of the system using that state decomposition should be Markovian for all lag times $\tau \geq \tau_{\mathrm{int}}$. This is also a test of whether the Chapman-Kolmogorov equation holds, but as it computes only L numbers and orders them by timescale, it allows emphasis to be placed on the longest timescales in the system.

Unfortunately, this method has some drawbacks. First, small uncertainties in the eigenvalues of the transition matrix can induce very large uncertainties in the implied timescales. With increasing lag time τ, the number of statistically independent observed transitions, from which $\boldsymbol{T}(\tau)$ is estimated, diminishes, and the statistical uncertainty in the implied timescales τ_k will grow. Second, while stability of the implied timescales with respect to lag time is a *necessary* consequence of history independence, it is not itself *sufficient* to guarantee history independence, though we may be unlikely to encounter physical systems for which this is problematic. However, tests on simple models indicate that the information theoretic metric suggests the emergence of Markovian behavior on similar lag times to this method, suggesting some degree of fundamental equivalence [141].

In this work, the analysis of implied timescales as a function of lag time will be our primary test for the emergence of Markovianness.

5.3 The automatic state decomposition algorithm

Based on the theory above, we provide a list of practical considerations for an automatic state decomposition algorithm and then present an algorithm that meets the criteria proposed below. The algorithm operates on an ensemble of molecular dynamics trajectories where conformations (the Cartesian coordinates of all atoms of the macromolecule) have been stored at regular intervals. In this work, we apply the method to a set of *equilibrium* trajectories at the temperature of interest, but the algorithm can in principle be applied to trajectories generated from *biased* initial conditions, provided the unbiased transition probabilities between regions of configuration space can be computed. We stress that the algorithm presented here is simply a first attempt at a truly general and automatic algorithm for use with biomacromolecules.

5.3.1 Practical considerations for an automatic state decomposition algorithm.

There are several desirable properties that a state decomposition should possess to be both useful and practical:

1. It is not uncommon for simulations conducted on supercomputers such as Blue Gene [62, 73], distributed computing platforms such as Folding@Home [140, 158], or even computer clusters to generate datasets that may contain 10^5 to 10^7 configurations in up to 10^4 trajectories, therefore prohibiting the use of any algorithm with a time complexity greater than $\mathcal{O}(N \log N)$ in the number of configurations.

2. Molecules may have symmetries under permutation of atoms, such as aromatic rings, the protons on methyl groups, and the oxygens of carboxylate groups that should be accounted for in some way.

3. The state decomposition algorithm should produce a decomposition for which dynamics appears to be Markovian at the shortest possible lag time τ, so as to produce the most useful model.

4. The resulting model should not generate so many states so that the elements of the transition matrix will be statistically unreliable.

5.3.2 Sketch of the method.

A state decomposition algorithm intended to produce the most *useful* models, as discussed in Section 5.2.3 above, would generate models that minimize the internal equilibration time τ_{int}, the minimum time for which the model behaves in a Markovian fashion. Unfortunately, τ_{int} is difficult to determine directly, so we are instead forced to identify some surrogate quantity whose maximization will hopefully lead to improved separation between fast intrastate and slow interstate timescales. Following the approach of Ref. [84], we define a measure of the *metastability* Q of a partitioning into L *macrostates* as the sum of the self-transition probabilities for a given lag time τ:

$$Q \equiv \sum_{i=1}^{L} T_{ii}(\tau) \tag{5.10}$$

For $\tau = 0$, $Q = L$, and decays to unity as τ grows large enough for the self-transition probabilities T_{ii} to reach the equilibrium probabilities of each macrostate. Poor partitionings into weakly metastable states will result in a small Q, as trajectories started in some states will rapidly exit; conversely, good partitionings into strongly metastable states will result in a large Q, as trajectories will remain in each macrostate for long times. In the absence of statistical uncertainty, Q is bounded from above by the sum of the L largest eigenvalues of the true dynamical propagator for the system [84].

The goal of our algorithm is to identify a partitioning into L contiguous macrostates that maximizes the metastability Q. While in principle, the boundaries between these macrostates can be varied directly to optimize Q, in analogy to variational transition state theory [182], a complicated parameterization may be necessary to describe the potentially highly convoluted hypersurfaces separating the states, and Q may have multiple maxima in these parameters. Instead, we choose an approach based on *splitting* the conformation space into a large number of small contiguous *microstates* and then *lumping* these microstates into macrostates in such a way that maximizes the metastability.

This approach is very similar to the approach of Schütte and coworkers described in Ref. [152], but with a substantial difference. In their work, each degree of freedom of the molecule (such as a torsion angle) is subdivided independently to produce a multidimensional grid. As the number of states is exponential in the number of degrees of freedom, this approach quickly becomes intractable for macromolecules that possess large numbers of degrees of freedom, even if the sparsity of the transition matrix is taken into account. Instead, we choose to let the data define the low-dimensional manifold of configuration space accessible to the macromolecule, and we can apply any clustering algorithm that is no worse than $\mathcal{O}(N \log N)$ in the number of configurations to decompose the sampled conformation space into a set of K contiguous microstates. This step corresponds to the first *split* step in Figure 5.1.

Once the conformation space is divided into K microstates, we *lump* the microstates together to produce $L < K$ macrostates with high metastability, Q. This corresponds to the first *lump* step in Figure 5.1. The difficulty here is that the uncertainty in the metastability of a partitioning can be large if any macrostate contains very few configurations. Since a macrostate may consist of a single microstate, the microstates must be large enough for the self-transition elements to be statistically well-determined. This comes at a price: with large microstates, the procedure may have difficulty accurately determining the boundaries between macrostates because the resolution of partitioning is limited by the finite extent of the microstates. Additionally, the choice of decomposition into microstates is arbitrary, whereas we would like the state decomposition algorithm to produce equivalent sets of macrostates regardless of how good the initial partitioning was.

To overcome these difficulties, we *iterate* the aforementioned procedure. After microstates are combined into macrostates, each macrostate is again fragmented into a new set of microstates (the second *split* step in Figure 5.1). The refined set of all microstates is then lumped to form refined macrostates (the second *lump* step in Figure 5.1). In this way, the boundaries between macrostates are iteratively refined, and regions incorrectly lumped in previous iterations may be split off and

lumped with the correct macrostate in subsequent iterations. At convergence, the same set of macrostates will simply be split and lumped back together in the same way — no further shuffling of conformations between macrostates will occur.

There is unfortunately no unambiguous way to choose the number of states L. If there is a clean separation of timescales, examination of the eigenvalue spectrum of the microstate transition matrix may suggest an appropriate value of L [153]. In a hierarchical system, there will be many gaps in the eigenvalue spectrum and many of choices of L will lead to good Markovian models of varying complexity. There is, however, a tradeoff between the number of states and the amount of data needed to obtain a model with the same statistical precision. It may be necessary to apply the algorithm with multiple choices of L to produce a model sufficient for resolving the timescales of interest.

5.3.3 Implementation.

There are a number of implementation choices to be made in the algorithm given above, and here we briefly summarize and justify our selections.

For the split step, we choose to apply K-medoid clustering [81] because of its $\mathcal{O}(KN)$ time complexity (where K can be taken to be constant) and ease of parallelization. Additionally, K-medoid clustering has an advantage over the more popular K-means clustering [113] in this application, as it does not require averaging over conformations, which may produce nonsensical constructs when drastically different conformations are included in the average. Splitting by K-medoid clustering is initiated from a random choice of K unique conformations to function as *generators*. All conformations are assigned to the microstate identified by the generator they are closest to by some distance metric (defined below). Next, an attempt is made to update the generator of each microstate. K members of the microstate, drawn at random, are evaluated to see if they reduce the intrastate variance of some distance metric from the generator. If so, the configuration for which the intrastate variance is minimal is assigned as the new generator. All conformations are then reassigned to the closest generator, and the process of updating the generators is repeated. In standard K-medoid applications, this procedure is iterated to convergence, but since the purpose of the splitting phase is simply to divide the sampled manifold of configuration space into contiguous states, ensuring that each state is significantly populated, only five iterations of this procedure were used.

For the distance metric, we selected the root-mean squared deviation (RMSD), computed after a

minimizing rigid body translation and rotation using the rapid algorithm of Theobald [178]. In the first splitting iteration, only C_α atoms were used to compute the RMSD due to the expense of having to cluster all conformations in the dataset; in subsequent iterations, all heavy atoms (excepting those indistinguishable by symmetry) were used, as well as sidechain polar hydrogens. This metric was chosen because it possesses all the qualities of a proper distance metric [169], accounts for both local similarities between pairs of conformations as well as global ones, and runs in time proportional to the number of atoms, as opposed to a metric such as distance matrix error (DME or dRMSD), which scales as the square of the number of atoms. In molecules with additional symmetry, the distance metric can be adjusted accordingly. Our choice of distance metric is not the only one that would suffice; any distance metric which can distinguish between kinetically distinct conformations is sufficient for this algorithm. For example, backbone RMSD would ignore potentially relevant sidechain kinetics.

Lumping to L states so as to maximize the metastability Q of the macrostates proceeds in two stages. In the first stage, information on the metastable state structure contained in the slowest eigenvectors [41, 83, 153, 154] is used to construct an initial guess at the optimal lumping. Because the eigenvectors contain statistical noise, this initial guess may not actually be optimal; because of this, we include a second stage that uses a Monte Carlo simulated annealing (MCSA) optimization algorithm to attempt to further improve the metastability. Though the MCSA algorithm could in principle be used without the first stage to find optimal lumpings, we find its convergence is greatly accelerated by use of the initial guess.

In the first stage, a transition matrix among microstates is computed (using Eq. 5.9) taking advantage of both stationarity and time-reversibility for a short lag time τ, typically the shortest interval at which configurations were stored. Motivated by the Perron cluster cluster analysis (PCCA) algorithm of Deuflhard *et al.* [41], an initial guess for the optimal lumping of microstates to macrostates is generated using the *left* eigenvectors[5] associated with the largest eigenvalues of the microstate transition matrix. We begin by assigning all microstates to a single macrostate. For each eigenvalue, the corresponding eigenvector contains information about an aggregate transition between the set of microstates with positive eigenvector components and the set with negative components, with a timescale determined by the eigenvalue; equilibration within each set must occur on a faster timescale, provided the eigenvalues are non-degenerate. We can therefore use this information to identify one macrostate to divide in two. We select the macrostate with the largest

[5]The left eigenvector $\boldsymbol{v}_k$ is simply related to the right eigenvector $\boldsymbol{u}_k$ by $(\boldsymbol{v}_k)_i = p_{\mathrm{eq},i}^{-1} (\boldsymbol{u}_k)_i$ [135].

L_1 norm of the vector formed from the eigenvector components that belong to that macrostate, after subtracting the mean of this vector, as the state to split. In Ref. [41], the sign structure alone was used to split these sets, but we find it more stable to split about the mean. This procedure is performed for eigenvectors $k = 2, \ldots, L$ in order, which should correspond to the slowest processes in the system, generating a total of L macrostates.

Due to statistical noise in the eigenvectors and near-degeneracy in the eigenvalues, this procedure does not always result in the lumping with the maximal metastability Q. Therefore, in the second stage, the metastability was maximized using a Monte Carlo simulated annealing (MCSA) algorithm, using the eigenvector-generated lumping as an initial seed. In each step of the Monte Carlo procedure, a microstate was selected with uniform probability and assigned to a random macrostate. If this proposed move would leave a macrostate empty or did not change the partitioning, it was rejected immediately. The proposed partitioning was accepted with probability $\min\{1, e^{\beta \Delta Q}\}$, where the metastability Q of the proposed lumping was rapidly computed by combining elements of the matrix of inter-microstate transition counts. The effective inverse temperature parameter β was set to be equal to the step number, and the MCSA procedure run for 20 000 steps. Twenty independent MCSA runs were initiated from the initial eigenvector-based partitioning, and the partitioning with the highest metastability sampled in any run was selected to define the lumping into macrostates.

It should be noted that the metastability Q is not the only surrogate that could be optimized in order to produce a useful state decomposition. Many choices may be possible, especially when one considers the problem of lumping as an attempt to preserve the L longest timescales (determined by the eigenvalues of the transition matrix near unity) present in the microstate transition matrix. One could choose to maximize the fastest eigenvalue or timescale of the lumped transition matrix, the product of eigenvalues (which would give more weight to faster timescales), or even a weighted sum of the eigenvalues, where the weights might be due to the equilibrium importance of the eigenmode in dynamics or in modeling a process of interest. Unfortunately, these quantities all necessitate computing some eigenvalues or the determinant of the lumped transition matrix for every proposed lumping to be evaluated by the MCSA algorithm, which would add significant computational burden. Alternatively, other quantities could be computed from the transition matrix directly, such as the state lifetimes estimated from the self-transition probabilities as $\tau_{L,i} = (1 - T_{ii})^{-1}$. However, the combination of computational and theoretical convenience makes the use of metastability a natural choice here.

For the remaining iterations, the K-medoid clustering is repeated independently on each

macrostate. We set a minimum expected microstate size (estimated by the population of the macrostate divided by K) to ensure statistical reliability of the transition probability matrix. This is set to 100 configurations (unless otherwise noted), though a more useful criteria may be to set a minimum number of statistically independent visits to the state. Each macrostate is split into a number of states such that the expected microstate population (assuming even division into microstates) is no smaller than this threshold, or a maximum of 10 microstates. The lumping step is then repeated on all resulting microstates. The entire procudure of splitting and lumping was repeated for a total of 10 iterations, which for the applications considered here was sufficient for convergence of the slowest timescales.

5.3.4 Validation.

To validate the model, we examine the largest implied timescales as a function of lag time, as computed for the eigenvalues of the transition matrix by Eq. 5.7. In particular, we attempt to determine the minimum lag time after which the implied timescales appear to be independent of lag time to within the estimated statistical uncertainty (see Section 5.2.4). To estimate the statistical uncertainty of these implied timescales, we perform a bootstrapping procedure [49] on the pool of independent trajectories. Forty bootstrap samples of a number of trajectories equal to the number of independent trajectories in the dataset pool are generated, drawn with replacement from the pool of trajectories, except for alanine dipeptide, where 100 bootstrap samples were used. The implied timescales are computed for each sample, and the set of computed timescales is used to estimate a confidence interval. In figures, uncertainties will always be shown as 68% symmetric confidence intervals about the mean of the bootstrap sample, while uncertainties in quantities printed as $a \pm b$ will indicate variances about the mean.

5.4 Applications

5.4.1 Alanine dipeptide.

We first demonstrate application of the automatic state decomposition algorithm to a simple model system, terminally blocked alanine peptide (sequence Ace-Ala-Nme) in explicit solvent. Because the slow degrees of freedom (ϕ and ψ torsions, labeled in Figure 5.2, left) are known *a priori*[6], it is

[6] Simulations of alanine dipeptide examining the committor distribution have implicated solvent coordinates as the next-slowest degree of freedom [17, 111], but we have previously verified that ϕ and ψ torsions form a sufficient basis for the slow degrees of freedom on timescales of 6 ps and greater [25].

relatively straightforward to manually identify metastable states from examination of the potential of mean force, making it a popular choice for the study of biomolecular dynamics [5, 17, 22, 25, 85, 125]. Previously, a master equation model constructed from a set of six manually identified states (Figure 5.2, right) was shown to reproduce dynamics over long times (with the time to reach equilibrium over 100 ps at 302 K) given trajectories only 6 ps in length [25]. We therefore determine whether the automatic algorithm can recover a model of equivalent utility to this manually constructed six-state decomposition for this system, as well as study its convergence properties.

Trajectories were obtained from the 400 K replica of a 20 ns/replica parallel tempering simulation described in Ref. [25], and consisted of an equilibrium pool of 1000 constant-energy[7], constant-volume trajectory segments 20 ps in length with configurations stored every 0.1 ps. Velocities were reassigned from a Maxwell distribution after each exchange attempt[8]. The peptide was modeled by the AMBER parm96 forcefield [95], and solvated in TIP3P water [90]. The previous study [25] considered the dynamics at 302 K, but resorted to a focused sampling strategy where a number of trajectories were initiated from equilibrium distributions within constricted *selection cells* [174] in order to obtain statistically reliable estimates of the transition matrix. Here, as the focus was on locating these metastable states from equilibrium data, we found it necessary to use equilibrium data from a higher temperature — here, the 400 K replica — in order to obtain sufficient numbers of trajectories covering the entirety of the landscape. A 2D potential of mean force (PMF) at 400 K in the (ϕ, ψ) backbone torsions was estimated from the parallel tempering simulation using the weighted histogram analysis method [26, 100] by discretizing each degree of freedom into 10° bins (Figure 5.2). Because the (ϕ, ψ) torsions are supposed to be the *only* slow degrees of freedom in the system, we can visually identify basins in the potential of mean force with metastable states in the PMF. The six such states identified from the 302 K PMF in the previous study [25], identified as dark lines in Figure 5.2, can be seen to still adequately separate the free energy basins observed at 400 K. We take this decomposition as our reference "gold standard", and compare state decompositions obtained from our automatic state decomposition algorithm with this one.

[7] Note that, because these trajectories are constant energy, the system (which includes macromolecule and a large bath of solvent) cannot exchange heat with its environment. A Markov model constructed from such a pool of trajectories therefore models the case where the system does not exchange a significant amount of heat with its environment during the course of transitions occurring on the Markov time.

[8] Note that only 10 ns/replica were used in Ref. [25] — the data presented here includes an additional 10 ns/replica of production simulation. Additionally, configurations containing *cis*-ω torsions discussed in the text were not observed in the first 10 ns/replica cited in the previous study – these conformations only appeared in the latter 10 ns/replica.

First, the automatic state decomposition method described in Section 5.3 was applied to this dataset in a fully automatic way to obtain six macrostates that could be compared with the "gold standard". Since there is only one C_α atom in the peptide, we opted to use the backbone RMSD (including the amide proton and carbonyl oxygen) in the first stage, splitting to 100 microstates; subsequent iterations used the distance metric and splitting procedure described in Section 5.3.3. A single sampling interval — 0.1 ps — was used for the calculation of the metastability metric Q used in lumping, as described in Section 5.3.2. Application of state decomposition to the entire dataset revealed a state that heavily overlapped with several others when projected onto the (ϕ, ψ) map, along with an extremely long timescale associated with its transitions (data not shown). Closer examination of the ensembles of configurations contained in this overlapping state revealed that the overlapping regions differed by a peptide bond isomerization; a small population of the trajectories contained an N-terminal ω peptide bond in the *cis* state, rather than the typical *trans* state. The number of trajectories that connected these states was extremely small. Examination of the parallel tempering data revealed that the majority of these transitions had occurred at much higher temperature, and that the *cis*-ω configurations found at 400 K had reached this temperature by annealing from higher temperature; in the majority of trajectories at 400 K that contained *cis*-ω configurations, the peptide remained in this state over the duration of the trajectory. This is a clear demonstration of how the automatic algorithm can discover additional slow degrees of freedom that the experimenters may not realize are important. For subsequent investigation, due to the extremely small number of transitions, trajectories containing conformations that included *cis*-ω bonds (a total of 25 trajectories) were removed from the set of trajectories used for analysis, leaving 975 trajectories.

The results of the automatic state decomposition algorithm applied to this reduced dataset can be seen in Figure 5.3, in comparison with the "gold standard" manual state decomposition from Ref. [25] and a "poor" manual decomposition that is expected to fail to reproduce kinetics well because its states include internal kinetic barriers[9]. Independent applications of the automatic method were observed to yield two distinct decompositions with metastabilities within statistical uncertainty, both of which slightly exceeded the metastability of the manual decomposition (Figure 5.3, bottom two plots). In the automatic decomposition with slightly larger metastability, six states in the same general locations as the manual "gold standard" decomposition are observed, though

[9]The poor partitioning was defined as follows: (1) $\phi \in [(179, -135]$, $\psi \in (98, 48]$; (2) $\phi \in (-135, -60]$, $\psi \in (98, 48]$; (3) $\phi \in (179, -135]$, $\psi \in (48, 98]$; (4) $\phi \in (-135, -60]$, $\psi \in (48, 98]$; (5) $\phi \in (-60, 179]$, $\psi \in (98, -45]$; (6) $\phi \in (-60, 179]$, $\psi \in (-45, 98]$. Specified intervals denote intervals on the torus, which is continuous from -180 to +180. All torsions are specified in degrees.

the boundaries are somewhat perturbed. However, the timescales as a function of lag time are not significantly different from those of the manual "gold standard" decomposition (Figure 5.3, right) . In the other automatic decomposition with nearly equal metastability, states 3 and 4 of the manual decomposition (numbering given in Figure 5.2) have been merged into a single state, and state 5 of the manual decomposition has been fragmented into two states. Despite this, the timescales as a function of lag time again appear to be statistically indistinguishable from those of the "gold standard", suggesting that this model may have equal utility. This suggests that the phenomenological rates may not be very sensitive to the exact choice of state boundaries after the Markov time, as recrossings will have been suppressed by this time. The fact that this lumping does not disrupt the behavior of the model substantially is not altogether surprising, because the barrier separating states 3 and 4 is rather small, and these states act like a single state even on timescales of a few picoseconds or greater. In contrast, the "poor" decomposition has extremely short timescales which do not appear to level off over the course of 10 ps.

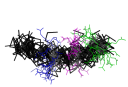
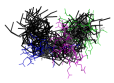
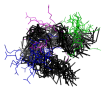
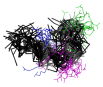
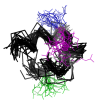
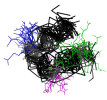
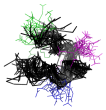
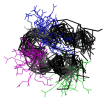
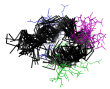
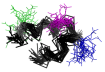
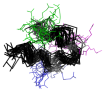
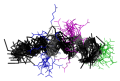
To examine the ability of the algorithm to recover optimal partitionings, the automatic state decomposition algorithm was applied to both the "gold standard" and "poor" manual decompositions (Figure 5.4) to see whether these partitionings would be maintained over the course of subsequent iterations. Ten iterations were conducted, with each macrostate split to ten microstates in the first iteration, rather than the entire configuration space being split into 100 states. In both cases, the algorithm converged to nearly equivalent partitionings after ten iterations (Figure 5.4), as verified by examination of the converged timescales (data not shown). This suggests the method yields partitionings that are relatively stable and optimal.

From the "poor" manual decomposition, however, a number of conformations in manual states 5 and 6 are incorrectly grouped with state 2, though this did not significantly affect the timescales. Further investigation showed that the algorithm never split these conformations from state 2, partly due to their comprising only 1 % of the population of the state. Splitting each macrostate into more microstates should alleviate this problem.

5.4.2 The F_s helical peptide.

To illustrate behavior of the automatic state decomposition method on a larger peptide system with fast kinetics, we applied it to the 21-residue helix-forming F_s peptide, which has been studied extensively both experimentally [102, 108, 109, 179, 197] and computationally [70, 163, 164, 202]. Since helix formation occurs on the nanosecond timescale, Sorin *et al.* were able to reach

Table 5.1: **Macrostates from a 20-state state decomposition of the F_s helical peptide.** The backbone is depicted in alpha carbon trace, and arginine sidechains are shown in blue (Arg10), magenta (Arg15), and green (Arg20) for clarity.

state	1	2	3	4	5
					
members	358 712	98 222	46 921	22 559	22 367
τ_{ac} (ns)	3.1	0.9	1.4	0.6	4.0
state	6	7	8	9	10
					
members	15 859	11 975	11 053	11 024	
τ_{ac} (ns)	1.3	1.6	2.2	2.0	
state	11	12	13	14	15
					
members	7 976	7 808	7 771	5 978	5 626
τ_{ac} (ns)	2.2	1.2	1.6	11.3	2.3
state	16	17	18	19	20
					
members	1 856	955	531	525	490
τ_{ac} (ns)	5.0	10.3	47.0	29.1	15.2

equilibrium from both helix and coil conformations and observe equilibrium conformational dynamics using ensembles of molecular dynamics trajectories on the distributed computing platform Folding@Home [164]. Two sets of 1000 trajectories at 302 K of varying length of the capped F_s peptide (sequence Ace-A_5[AAARA]$_3$A-Nme), one set initiated from an ideal helix and another from a random coil, were obtained from Sorin *et al.* [164]; details of the simulation protocol are available therein. The first 40 ns of each trajectory, a conservative overestimate of the time to reach equilibrium from either helix or coil, was discarded, and the two sets of trajectories combined to yield a total of 1689 trajectories varying in length from 10 ns to 95 ns with a sampling interval of 100 ps. In total, this equilibrium dataset contained nearly 65 μs of simulation data in 642 604 conformations. The peptide was modeled using the AMBER-99ϕ forcefield [164, 193]

and solvated in TIP3P water [90]. Though the Berendsen weak-coupling scheme [10] was employed for thermal and pressure control[10], we presume the trajectories still obey microscopic reversibility when only the coordinates of the macromolecular solute are considered for the purposes of computing transition probabilities.

We performed automatic state decomposition on this dataset to generate a set of 20 macrostates through 10 iterations of splitting and lumping. In the first iteration, the sampled region of conformation space was split into 400 microstates. In subsequent iterations, each macrostate was split into 50 microstates (or, if the expected microstate size would fall below 500 configurations, a number of microstates chosen to ensure the expected microstate size would remain above this threshold).

Automatic state decomposition produced a structurally diverse set of states (Table 5.1), ranging in size from over 350 000 members to 500 members, with the majority containing from 5 000 to 20 000 members. The states include a large extended helix/coil state (state 1 of Table 5.1), consisting of slightly over half the total conformations in the dataset; a pure helix state (state 15); a number of helix/coil states which are bent in half to different degrees to form tertiary contacts (states 2–14); and a number of smaller helical states which are bent into circles to form tertiary interactions (states 16–20). A previous analysis of this data clustered conformations into states based on dissimilarity in various order parameters: the number of helical residues, number of helical segments (stretches of helical residues), length of the longest helical segment, and radius of gyration [164]. We compared the macrostates generated by the automatic algorithm with these clusters, and found that while some states are similar, namely the bi-nucleated helices of different sizes, most were quite different. The most significant difference was the grouping of helix and coil conformations into a single macrostate in the lumping phase of the automatic algorithm; the order parameter-based clustering kept helix and coil states distinct [164]. When examining individual trajectories, we noticed conformations would rapidly flicker between helices and coils between consecutive frames of the trajectory, suggesting that their rapid interconversion justifies their lumping into a single macrostate. Additionally, the clustering based on helical order parameters was unable to distinguish certain structures that involved tertiary contacts, such as the bent and circular helical states. Interestingly, a previous study employing the related AMBER parm03 forcefield [45] identified similar configurations to those noted by the automatic state

[10] We note that thermal and pressure control, by design, modulate the velocities of molecules in the system, which may have a nonphysical effect on dynamics. In this particular application, however, we are only comparing our analysis with the original simulation data, rather than directly with experiment.

decomposition, terming these states helix (corresponding to our state 1), helix-turn-helix, adjusted helix-turn-helix, helix-wind-helix, globular helix (states 16–20), and helix tail (state 15) [202].
We then examined the implied timescales as a function of lag time (Figure 5.5). Lumping appeared to preserve the longest timescales found in the microstate transition matrix (data not shown), indicating that our lumping scheme had been successful in identifying a nondestructive lumping into kinetically metastable states at each iteration. Over the course of 10 iterations, the metastability (as optimized with a lag time of 100 ps) increased from 12.5 ± 0.3 to 14.5 ± 0.1, suggesting that the iterative refinement was actually improving the quality of the state decomposition. On the first iteration, the longest timescales increase nearly linearly with lag time, while on the last iteration, some of the longest timescales become stable by a lag time of 4 – 5 ns, suggesting Markovian behavior for some of the processes.
Using the interpretation of eigenvector components in terms of aggregate modes described in Section 5.2.1, the longest timescale was found to correspond to movement between the extended helix/coil state (state 1) and one of the twisted helix-turn-helix states (state 18) with only 500 members. We found, however, that state 18 appeared a small number of times in thirty trajectories, and over 450 times in a single trajectory. Further examination revealed that conformations belonging to this state were almost exclusively adjacent to conformations belonging to state 5, and structural comparison of conformations of these two states showed they were strikingly similar. This suggests that slight conformational differences between conformations in states 18 and 5 allowed the K-medoid clustering algorithm to partition between these states in a splitting step, and since state 18 was mainly isolated in a single trajectory, its self-transition probability was maximized by *not* lumping it with state 5, even though the two behaved in a similar kinetic fashion. Indeed, when we manually lump states 18 and 5, the longest timescale, corresponding to transitions involving state 18, disappears, but the remaining timescales are all preserved (data not shown).
A second potential cause of the increase with lag time observed in some of the other long timescales may be due to the finite length of trajectories. If the state is long-lived, and occurs near the trajectory beginning or end, then it can be seen that the estimated self-transition probability T_{ii} increases as a function of lag time. This effect is most pronounced when a state occurs in very few trajectories, and appears to be mitigated when the state occurs in many trajectories at random times within the trajectory.
In order to determine which states are poorly characterized, we estimated the number of statistically independent visits to each macrostate. Since sequential samples from a single trajectory are temporally correlated, we computed the integrated autocorrelation time [87, 173]

$\tau_{\mathrm{ac},i}$ for each macrostate i. Ignoring statistical uncertainty, this correlation time is an upper bound on the equilibration time within a state; long-lived states will necessarily have long autocorrelation times, but trajectories trapped within them may contain many uncorrelated samples if the internal equilibration time is short. In the absence of a convenient way to quantify the internal equilibration time for each state[11], the autocorrelation time provides a better estimate of the appropriate timescale than the time to reach global equilibrium τ_{eq}. As the correlation functions became statistically unreliable at times larger than 10 ns, a least squares linear fit to the log of the computed correlation function over the first 10 ns was used to estimate the tail at times greater than 10 ns, and this combined correlation function was integrated to obtain the autocorrelation time. The effective number of independent samples for each state was then estimated by summing the number of independent samples from each trajectory (which are assumed independent), where the effective number of independent samples of state i from trajectory n is computed as $N_{ni}^{\mathrm{eff}} \approx \min\{1, N_{ni}/g_i\}$, where N_{ni} is the number of configurations from trajectory n in state i, and $g_i = 1 + 2\tau_{\mathrm{ac},i}$ is the statistical inefficiency of state i.

Computed state autocorrelation times are given in Table 5.1. For many states, the correlation time was 1 – 2 ns, giving thousands of independent samples; however, for five states, including the four which were involved in the four longest timescales, the correlation times were between 10 and 50 ns, suggesting that the dataset contained less than 50 independent samples of these states.

Currently, in the automatic state decomposition algorithm, we try to reduce the statistical uncertainty in the transition matrix by limiting the expected population of each state to be greater than some minimum number of configurations. Since the conformations appearing within some states may be highly correlated, the number of conformations within a state is not the best measure of how statistically well-determined its transition elements are; instead, it may be advantageous to place a lower limit on the effective number of independent visits to each state, which is far less than the number of configurations it contains. Alternatively, it may be necessary to ensure better characterization of these states by conducting additional simulations from them, provided the equilibrium transition probabilities can still be computed.

We constructed a Markov model from the transition matrix estimated at a 5 ns lag time, where some (though apparently not all) of the timescales appear to have stabilized. Repeated application of this transition matrix to an initial probability distribution can be compared to the transition probabilities at longer lag times estimated directly from the data to assess how well the model

[11] There is some indication that consideration of restrictions of Markov chains to these macrostates may facilitate the computation of the internal equilibration time [118].

reproduces the observed kinetics. The time evolution of probability density out of three states (state 2, a populous state; state 13, a moderately populated state; and state 19, a sparsely populated state) over the course of 50 ns is shown in Figure 5.6. The Markov model appears to do a very reasonable job of predicting the time evolution of the system to within statistical uncertainty over many times longer than the lag time it was constructed for. In fact, the time evolution was well-modeled for evolution out of nearly all states, except for state 13, for which dynamics seemed to be particularly poorly reproduced. This state has a particularly long correlation time, and many trajectories seem to contain only a single configuration that is part of this state, suggesting its boundaries are simply poorly resolved. Regardless, the time evolution is generally well-modeled for this system.

5.4.3 The trpzip2 β-peptide.

As an illustration of the application of the state decomposition algorithm to a system with complex kinetics implying the existence of multiple metastable states [199], we considered the engineered 12-residue β-peptide trpzip2 [28]. A set of 323 10 ns constant-energy, constant-volume simulations of the unblocked peptide[12] simulated using the AMBER parm96 forcefield [95] in TIP3P water [90] was obtained from Pitera et al. [143]; details of the simulation protocol are provided therein. The trajectories were initiated from an equilibrium sampling of configurations at 425 K, a temperature high enough to observe repeated unfolding and refolding events at equilibrium. Configurations were sampled every 10 ps, giving a total of 3.23 μs of data in 323 000 configurations.

The automatic state decomposition method was applied to obtain a set of 40 macrostates in 10 iterations of splitting and lumping. In the first iteration, the conformations were split into 400 microstates, and in subsequent iterations, as described in Section 5.3.3.

Figure 5.7 depicts some of the final set of 40 macrostates compared with a set of states produced by consideration of backbone hydrogen bonding patterns in a previous study by Pitera *et al.* [143]. (The complete set of macrostates is shown in a figure included as Supplementary Information.) As the trajectories considered here were resampled to 10 ps intervals (rather than 1 ps in Ref. [143]) we found less than five examples of the +2 and -2 hydrogen bonding states identified in Ref. [143], and therefore do not include them in the comparison. The automatic state decomposition method recovers states corresponding to the native, +1C, and +1N hydrogen bonding patterns, and often further separates conformations based on the packing of the tryptophan sidechains (Figure 5.7, A,

[12] Note that the peptide studied experimentally in Refs. [28] and [199] was synthesized with an amidated C-terminus, whereas the termini of the simulated peptide in the dataset considered here were left zwitterionic.

C, D). However, the -1N hydrogen bonding pattern is not further resolved, and instead is grouped into a state of mostly disordered hairpins; further examination is necessary to determine whether the algorithm simply failed to resolve this state or if the state is simply not long-lived. In addition to recovering most of the manually identified misregistered states, the algorithm was also able to greatly resolve the state labeled as "unfolded" in Pitera *et al.* (in that it did not conform to any of the enumerated hydrogen bonding patterns) into substates which exhibit considerable structure (E–J). Some of these kinetically resolved states have distinct hydrogen bonding patterns, such as where both strands are rotated (H), causing the tryptophan sidechains to appear on the opposite face, or where the misregistration is greater than two residues (G, J). This demonstrates the utility of the method in identifying additional kinetically relevant states that were not initially part of the experimental hypothesis space.

Figure 5.8 depicts the implied timescales of the kinetic model as a function of lag time. The longest timescale ranges between 25 and 35 ns and appears to stabilize over the range of lag times considered, though the uncertainty is quite large. Eigenvector analysis (described in Sec. 5.2.1) shows that this timescale corresponds to transitions between the unfolded and disordered hairpin states (E) and the hairpin with both strands rotated (H). The states labeled H together totaled 935 conformations, but appeared in only 13 trajectories, with over 95% of the conformations appearing in a single trajectory. Correlation time analysis (Sec. 5.4.2) suggests there are less than 10 independent samples for each of the three states, so proper resolution of this timescale would require more data. The second longest timescale grows to about 15 ns and levels off by around 4 ns, and corresponds to transitions between the unfolded and disordered hairpin states (E) and the native backbone states (A). The states involved in this transition are much better characterized, with a total of over 25 000 conformations appearing in over half the trajectories. The next three longest timescales were all between 3 and 4 ns and correspond to movement between the unfolded state (E) and various sets of misregistered states, namely the newly identified misregistered states I and J, and the +1C state (C). Unfortunately, these timescales are on the order of the Markov time for the whole system, so it is difficult to characterize these transitions well.

5.5 Discussion

Markov models are expected to be effective and efficient ways to statistically summarize information about the pathways (mechanism) and timescales for heterogeneous biomolecular processes such as protein folding. The great challenge is in defining an appropriate state space.

Here, we have presented a new algorithm for automatically generating a set of configurational states that is appropriate for describing peptide conformational dynamics in terms of a Markov model, though we expect it to be applicable to macromolecular dynamics in general. The algorithm uses molecular dynamics simulations as input, and generates the state definitions using information about the temporal order of conformations seen in the trajectories. The importance of having an automatic algorithm, *i.e.*, one that requires little or no human intervention, is that without it, human bias may inadvertently produce incorrect interpretations of the mechanism of conformational change by imposing a particular view on the simulation data. Additionally, molecular simulation datasets are becoming so large and complex that effectively summarizing the data or extracting insight becomes increasingly impractical unless the experimenter analyzes the data with a specific hypothesis in mind. Construction of a Markov model, however, allows for a "hypothesis-free" investigation of conformational dynamics, provided that the state space is sufficiently well sampled.

Our algorithm is based on the availability of large numbers of molecular dynamics simulations of appropriate simulation length such as might be generated by a supercomputer or a large (possibly distributed) cluster. Current technology allows for the production of thousands of simulations that can be tens of nanoseconds in length, hundreds of trajectories of up to hundreds of nanoseconds in length, or dozens that are on the order of a microsecond in length. Since our goal has been to develop Markov models that accurately characterize the time evolution of ensembles of macromolecules over experimental timescales (that can range from microseconds to milliseconds) from short simulations of single molecules, our approach places strong emphasis on the longest timescales observed in molecular simulations. For example, recognizing that ill-formed states often result in artificially shortened timescales, we sought to find states that maximize the timescales implied by their corresponding transition matrix for a particular choice of lag time and number of states. This resulted in the maximization of the metastability as a computationally convenient surrogate for minimizing the internal equilibration time τ_{int}.

Nonetheless, for the three data sets to which we have applied the method, there have been a number of important successes. For alanine dipeptide, the algorithm discovered a distinct manifold of states that consisted of conformations containing a *cis*-ω peptide bond. This manifold was discovered because it was kinetically distinct, rather than structurally distinct. Also, for alanine dipeptide, the method produces states that are robust and structurally very similar to the best ones produced manually, as well as kinetically indistinguishable to within statistical uncertainty according to our validation metrics. The application of the method to the F_s peptide data set

produced a set of states somewhat different from those identified previously from the clustering of helical order parameters [164]. The states produced by the algorithm properly identified many very long lived (metastable) conformations whose lifetimes and kinetics might determine behavior on an experimental timescale. The Markov model produced from this state decomposition and a 5 ns transition matrix was shown to reproduce the observed state populations over 50 ns to within statistical uncertainty. Finally, for the application of the method to the trpzip2 peptide the states constructed were consistent with ones previously identified [143]. This was very encouraging since the previously constructed states used an intramolecular hydrogen bonding criterion and the automatic algorithm utilized different observables and metrics, heavy atom RMSD and kinetics, to resolve states. Moreover, the automatic algorithm more finely resolved what was considered to be the "unfolded" ensemble into metastable states that were not identified by the decomposition based on hydrogen bonding patterns.

Therefore, the algorithm is achieving many of its design objectives. It provides a method for identifying and characterizing the *slower* degrees of freedom of a molecular system. It correctly identifies metastable states, dividing structurally very similar conformations into multiple sets that have short times for intraconversion but long times for interconversion. It combines together conformations that rapidly interconvert even though they may be structurally diverse. This is a prerequisite to capturing a concise description of the pathways for conformational changes. Once meaningful states are identified, the transition matrix itself encapsulates the branching ratios for various pathways and the timescales for overall relaxation to equilibrium from any arbitrary starting ensemble.

Work is ongoing to establish standards for the amount and nature of simulation data (number and length of simulations) needed to develop useful and sufficiently precise Markov models as well as investigations of the effect of quality metrics other than the trace of the transition matrix on the nature of the resulting states and time scales. Metrics for assessing the quality of the resulting model also need to be examined to complement, or as alternatives to, seeking stability of the implied time scales with respect to lag time. A strong candidate for this includes information theoretic-based metrics cited earlier [141]. Finally, alternative approaches to performing this state decomposition are a further matter of current study, such as the method of Noé and coworkers appearing in this issue, motivated by much the same ideas of metastability but employing different methods for the construction of a microstate space [131].

A general observation about the models produced using states defined by our method is that Markovian behavior is not obtained until lag times that are less than an order of magnitude shorter

than the longest timescales. Recall that the *utility* of a state space depends to a large extent on how early Markovian behavior is observed compared to the processes of interest. There are multiple possibilities for why this might be the case. For some molecular systems, there may be no identifiable metastable states in the usual sense. The existence of experimentally observed metastable states in protein systems (*e.g.* native, intermediate, unfolded) combined with the observation of metastable states in even models of small solvated peptides [25] argues that this may be unlikely. It could be that statistical uncertainty could be undermining both the metastability quality metric and the tests for Markovian behavior. Alternatively, the way we establish boundaries between states may not flexible enough to adequately divide true metastable regions. It may also be that we simply need to allow more states to be produced, resulting in subdivision of states that have internal barriers, to reduce the Markov times. Both of these possibilities could in principle be easily addressed by allowing the creation of more states. However, the creation of more states, especially ones with low populations, leads inevitably to situations where transition probabilities become statistically unreliable given the current fixed quantity of equilibrium data.
Long time scales are ultimately the result of infrequent events, and for even large but finite equilibrium datasets these will be small in number, with resulting small off-diagonal transition probabilities that are statistically unreliable. This has placed us in the particularly difficult but unavoidable situation of attempting to optimize a statistically uncertain objective function. One solution to this problem, of course, is to consider this algorithm as only the first step of an iterative process where important states and transitions are identified, and then further simulations are performed to improve the characterization of important regions of conformation space. This will allow refinement of the state space and improved precision for important selected transition probabilities. Information from the subsequent simulations could be combined with that from the first set using the selection cell approach described previously [174]. Selection of states, or regions of configuration space, from which further simulations should be initiated could be chosen based on uncertainty considerations [161].

5.6 Supporting Information

A Fortran 90/95 implementation of the automatic state decomposition algorithm presented here is available for download as part of the Supplementary Information for this article. The latest version of the code, along with the alanine dipeptide dataset, can be obtained from `http://www.dillgroup.ucsf.edu/~jchodera/code/automatic-state-decomposition/`.

The trpzip2 dataset is available directly from WCS upon request (E-mail: swope@almaden.ibm.com). A gallery of the macrostates produced by the 40-state decomposition of the trpzip2 peptide is also available as part of the Supplementary Information for this article.

5.7 Acknowledgments

The authors would especially like to thank Jed W. Pitera (IBM) for insightful discussion and constructive comments on this manuscript, and for providing simulation data for trpzip2; Eric Sorin (Stanford) for providing simulation data for the F_s peptide; Hans C. Andersen (Stanford) and Frank Noé (IWR Heidelberg) for enlightening conversations on the nature of Markov chain models; Vishal Vaidyanathan for assistance with clustering algorithms; and Libusha Kelly and David L. Mobley (UCSF) for critical comments on this manuscript. JDC was supported by an Howard Hughes Medical Institute and an IBM predoctoral fellowship. WCS acknowledges support from NSF MRSEC Center on Polymer Interfaces and Macromolecular Assemblies DMR – 0213618, and KAD the support of NIH grant GM34993. NS and VSP acknowledge support from NSF grant 0317072.

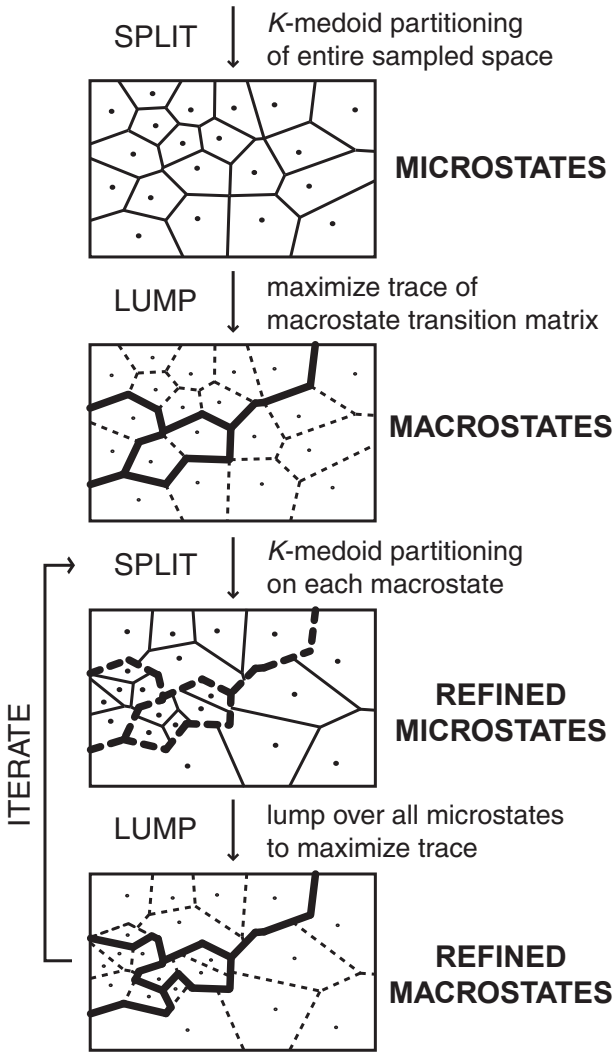


Figure 5.1: **Flowchart of the automatic state decomposition algorithm.**

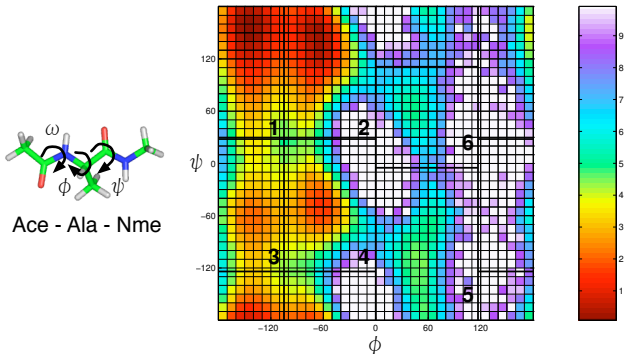


Figure 5.2: **Potential of mean force and manual state decomposition for alanine dipeptide.** Left: The terminally-blocked alanine peptide with ϕ, ψ, and ω backbone torsions labeled. Right: The potential of mean force in the (ϕ, ψ) torsions at 400 K estimated from the parallel tempering simulation, truncated at 10 $k_B T$ (white regions), with reference scale (far right) labeled in units of $k_B T$. Boundaries defining the six states manually identified in Ref. [25] from examining the 300 K PMF are superimposed, and the states labeled.

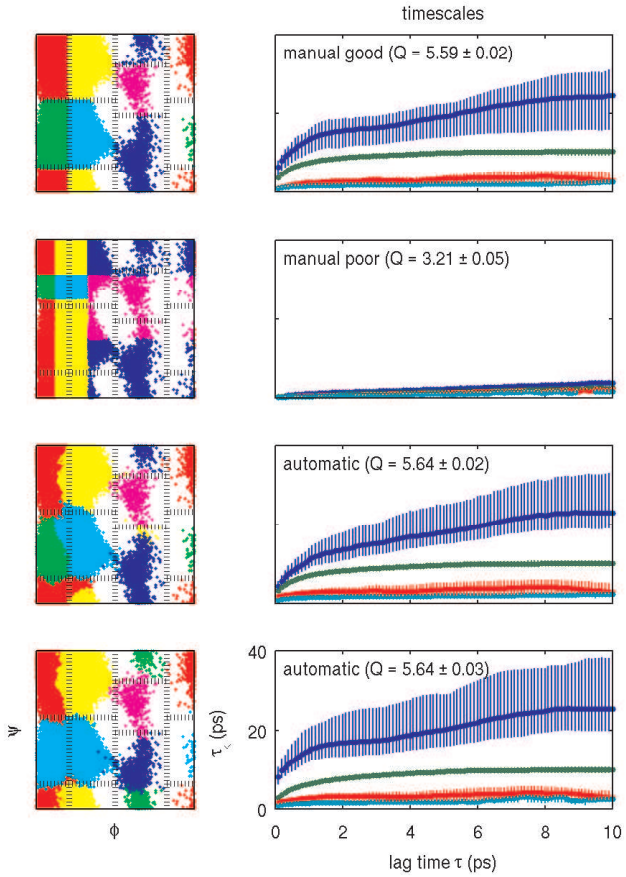


Figure 5.3: **Comparison of manual and automatic state decompositions for alanine dipeptide.** The left panels depict state partitionings, and the right panels the associated timescales (in picoseconds) as a function of lag time with uncertainties shown, as estimated from the procedure described in Section 5.3.4. Top two panels: Manual "good" or "gold standard" state decomposition from Ref. [25] and manual "poor" state decomposition, where the state boundaries are grossly distorted so as to include internal kinetic barriers within the states. Bottom two panels: Two nearly-equivalent partitionings obtained from the automatic state decomposition algorithm.

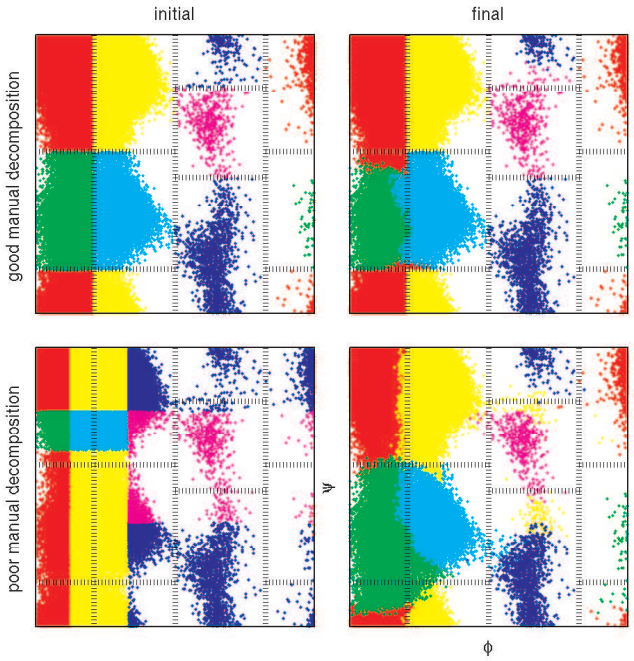


Figure 5.4: **Stability and recovery of optimal state decomposition for alanine dipeptide.** Top: Ten cycles of automatic state decomposition applied to a "good" manual partitioning (left) to yield an automatic partitioning (right). Bottom: Ten cycles of automatic state decomposition applied to a "poor" manual partitioning (left) to yield an automatic partitioning (right).

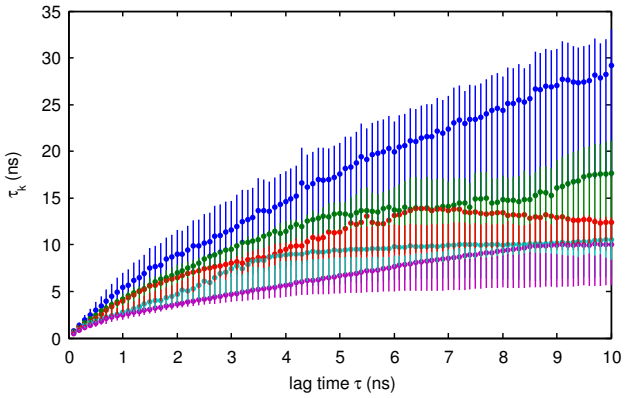


Figure 5.5: **Implied timescales of the F_s peptide as a function of lag time for 20-state automatic state decomposition.** The five longest timescales are shown. Circles represent the maximum likelihood estimate, and vertical bars depict 68% symmetric confidence intervals about the mean. Note the timescales associated with two processes appear to cross, but are here colored and uncertainties are estimated with bootstrapping by ordering them by rank. This may cause the uncertainties depicted here to be an underestimate of the true uncertainties of each process.

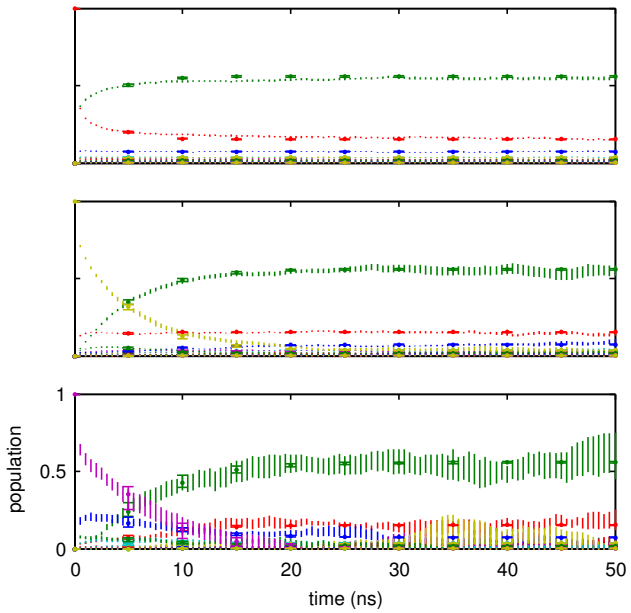


Figure 5.6: **Reproduction of observed state population evolution by Markov model for the $\mathbf{F}_s$ peptide.** The time evolution of the Markov model constructed from the 5 ns lag time transition matrix is shown by the filled circles with flat error bars, which denote the 68% confidence interval from realizations of a bootstrap sample of 40 transition matrices computed from a 5 ns lag time. Vertical bars without flat ends denote the 68% asymmetric confidence interval for the probability of finding the system in the 20 macrostates a given time after initial preparation in a specific state. The system was originally prepared in state 2 (top, red), 13 (middle, yellow), or 19 (bottom, purple). The most populous states are colored green (state 1), red (state 2), and blue (state 3).

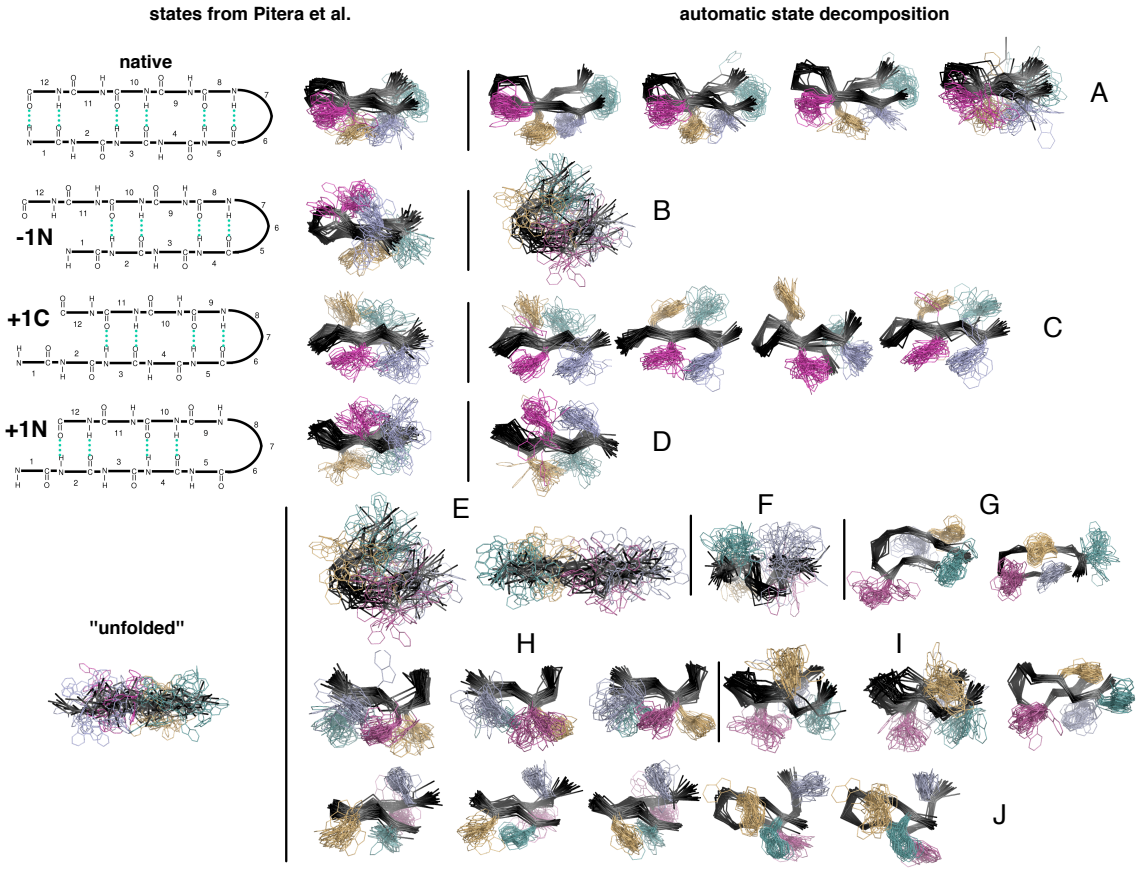


Figure 5.7: **Comparison of some trpzip2 macrostates found by automatic state decomposition with misregistered hydrogen bonding states identified in a previous study.** Left: The five hydrogen bonding patterns enumerated in Pitera *et al.* [143] that occurred in sufficient numbers in the subsampled trpzip2 dataset used here, with representative conformational ensembles. Right: A selection of macrostates discovered by automatic state decomposition that the contain the largest numbers of hydrogen bonding pattern states. The backbone is depicted in alpha carbon trace, and tryptophan sidechains are shown in light blue (Trp2), orange (Trp4), magenta (Trp9), and teal (Trp11). A complete set of macrostates obtained from the 40-state decomposition of the trpzip2 dataset is available as Supplementary Information.

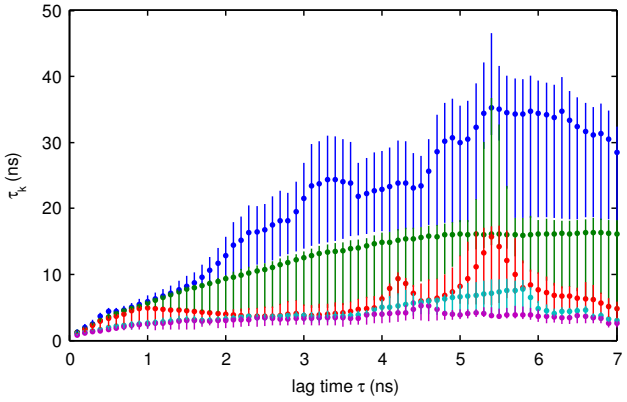


Figure 5.8: **Implied timescales of trpzip2 as a function of lag time for 40-state automatic state decomposition.** The five longest timescales are show. Vertical bars depict 68% confidence intervals.

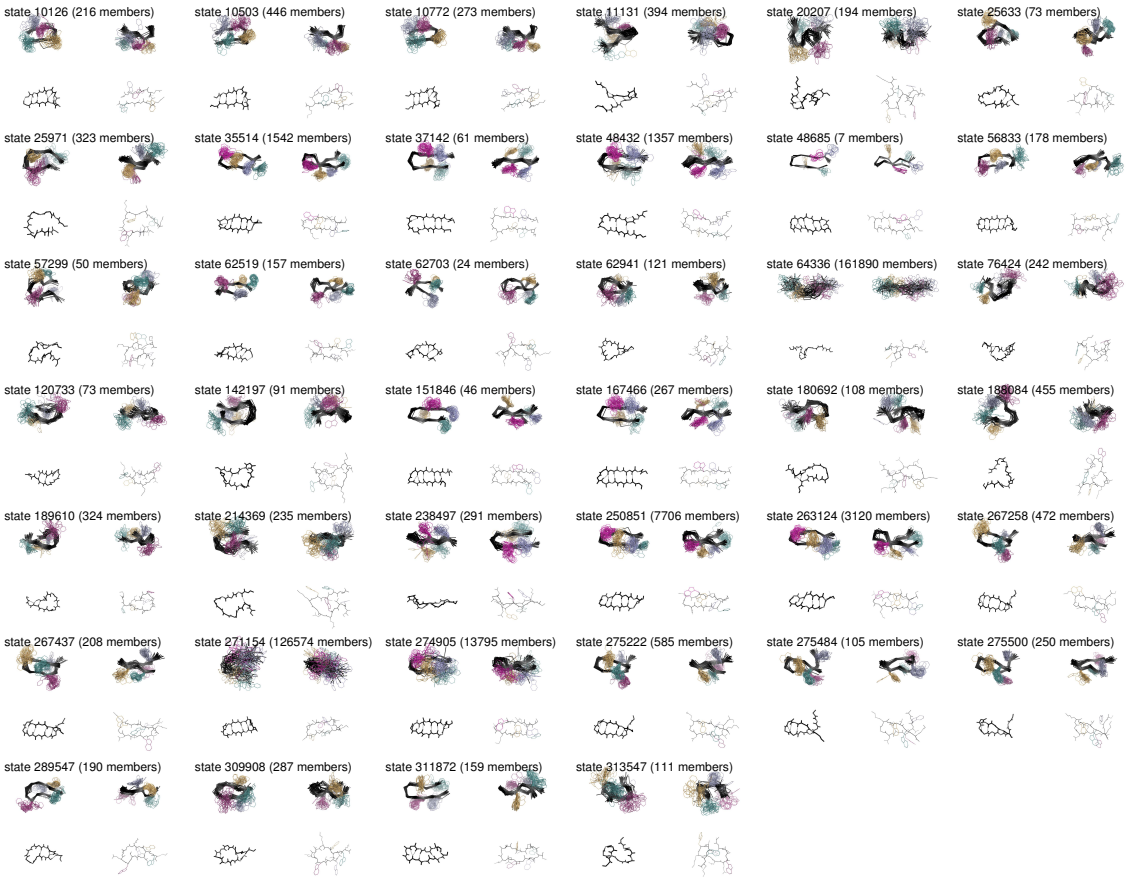


Figure 5.9: **Automatic state decomposition applied to trpzip2 to produce 40 macrostates.**

Chapter 6

Conclusion

In this dissertation, we have considered the problem of how the dynamics of biological macromolecules can be studied using discrete-state master equation or Markov models. The construction of these models requires two elements: a way of decomposing configuration space into states, and a method for computing transition rates or probabilities among these states. Chapter 3 contained a proof of concept, demonstrating that a model constructed from short trajectories for a model system, terminally-blocked alanine peptide in explicit solvent, was capable of accurately describing the statistical dynamics over long times. In Chapter 4, we considered a number of tests to establish the timescale at which the stochastic model would be an appropriate description of dynamics, an important prerequisite for evaluating various state decompositions. Finally, Chapter 5 presented a first attempt at an *automatic* algorithm for finding an optimal set of states given the number of states desired. These last two chapters describe the minimal essential ingredients for the construction of these models for problems of biological interest.

There are obviously many remaining challenges before the use of these models becomes widespread, and before it is possible to tackle the most interesting questions in biology. Currently, the quantity of data needed to construct these models requires the resources of massive computing projects, such as Blue Gene or Folding@Home, though it is becoming apparent that equilibrium datasets from these projects are also insufficient to construct well-determined Markov models. Future work will concentrate on multistage sampling techniques. There, an initial set of simulations is used to construct a crude state space, from which sets of trajectories are initiated. By use of well-defined starting distributions that completely cover configuration space (*e.g.* [98, 174, 194]), equilibrium transition probabilities can be reliably computed even if global equilibrium has never been achieved, as we saw in Chapter 4. Alternatively, automatic algorithms

to construct Markov models may start a number of simulations from any known structural information (*e.g.* crystal structures, NMR ensembles, or homology models) and iteratively construct and update the model, continually discovering new metastable states and reapportioning computational effort to always explore the most poorly characterized regions, perhaps using a method like the one described by Singhal *et al.* [161]. Transition path sampling [16, 37] could be employed to more efficiently compute transition rates or probabilities between states if at least one trajectory connecting the two has been found.

Further progress in the efficient construction of these models from biomolecular simulation data will lend insight into the biophysical processes of protein folding and dynamics. More efficient algorithms will not only allow more complex problems to be addressed, but also will allow these models to be constructed with modest computer clusters instead of distributed computing projects. There may come a time when the majority of molecular simulations are performed to construct these models. After all, what good is a single trajectory when the entire statistical dynamics can be characterized instead?

Bibliography

[1] M. P. Allen and D. J. Tildesley. *Computer simulation of liquids*. Clarendon Press, Oxford, 1991.

[2] Michael Andrec, Anthony K. Felts, Emilio Gallicchio, and Ronald M. Levy. Protein folding pathways from replica exchange simulations and a kinetic network model. *Proc. Natl. Acad. Sci. USA*, 102:6801–6806, 2005.

[3] C. B. Anfinsen. Principles that govern folding of protein chains. *Science*, 181(4096):223–230, 1973.

[4] Anjum Ansari, Joel Berendzen, Samuel F. Bowne, Hans Frauenfelder, Icko E. T. Iben, Todd B. Sauke, Erramilli Shyamsunder, and Robert D. Young. Protein states and proteinquakes. *Proc. Nat. Acad. Sci. USA*, 82:5000–5004, 1985.

[5] Joannis Apostolakis, Philippe Ferrara, and Amadeo Caflisch. Calculation of conformational transitions and barriers in solvated systems: Application to the alanine dipeptide in water. *J. Chem. Phys.*, 110(4):2099–2108, 1999.

[6] Ivet Bahar, Ali Rana Atilgan, and Burak Erman. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold. Des.*, 2:173–181, 1997.

[7] Y. S. Bai and M. D. Fayer. Time scales and optical dephasing measurements: Investigation of dynamics in complex systems. *Phys. Rev. B*, 39:11066–11084, 1989.

[8] Keith D. Ball and R. Stephen Berry. Realistic master equation modeling of relaxation on complete potential energy surfaces: Kinetic results. *J. Chem. Phys.*, 109(19):8557–8572, 1998.

[9] Oren M. Becker and Martin Karplus. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. *J. Chem. Phys.*, 106(4):1495–1517, 1997.

[10] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81(8):3684–3690, 1984.

[11] Alexander Berezhkovskii and Attila Szabo. One-dimensional reaction coordinates for diffusive activated rate processes in many dimensions. *J. Chem. Phys.*, 122:014503, 2005.

[12] Bernd A. Berg and Thomas Neuhaus. Multicanonical algorithms for first order phase transitions. *Phys. Lett. B*, 267:249–253, 1991.

[13] Bernd A. Berg and Thomas Neuhaus. Multicanonical ensemble: A new approach to simulate first-order phase transitions. *Phys. Rev. Lett.*, 68(1):9–12, 1992.

[14] Francisco Bezanilla. The voltage sensor in voltage-dependent ion channels. *Physiological Reviews*, 80(2):555–592, 2000.

[15] David D. Boehr, Dan McElheny, H. Jane Dyson, and Peter E. Wright. The dynamic energy landscape of dihydrofolate reductase catalysis. *Science*, 313:1638–1642, 2006.

[16] P. G. Bolhuis, D. Chandler, C. Dellago, and P. L. Geissler. Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.*, 53:291–318, 2002.

[17] Peter G. Bolhuis, Christoph Dellago, and David Chandler. Reaction coordinates of biomolecular isomerization. *Proc. Natl. Acad. Sci.*, 97(11):5877–5882, 2000.

[18] D. A. Case, D. A. Pearlman, J. W. Caldwell, T. E. Cheatham III, J. Wang, W. S. Ross, C. L. Simmerling, T. A. Darden, K. M. Merz, R. V. Stanton, A. L. Cheng, J. J. Vincent, M. Crowley, V. Tsui, H. Gohlke, R. J. Radmer, Y. Duan, J. Pitera, I. Massova, G. L. Seibel, U. C. Singh, P. K. Weiner, and P. A. Kollman. AMBER7, 2002.

[19] George Casella and Roger L. Berger. *Statistical inference*. Duxbury Press, Belmont, CA, 1998.

[20] David Chandler. Statistical mechanics of isomerization dynamics in liquids and the transition state approximation. *J. Chem. Phys.*, 68(6):2959–2970, 1978.

[21] Jean-Pierre Changeux and Stuart J. Edelstein. Allosteric mechanisms of signal transduction. *Science*, 308:1424–1428, 2005.

[22] Dmitry S. Chekmarev, Tateki Ishida, and Ronald M. Levy. Long-time conformational transitions of alanine dipeptide in aqueous solution: Continuous and discrete-state kinetic models. *J. Phys. Chem. B*, 108:19487–19495, 2004.

[23] John D. Chodera, Nina Singhal, Vijay S. Pande, Ken A. Dill, and William C. Swope. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *Submitted to J. Chem. Phys.*, 2006.

[24] John D. Chodera, William C. Swope, Jed W. Pitera, and Ken A. Dill. Describing protein folding kinetics by molecular dynamics simulations. 3. Validation of state space decomposition, with application to terminally-blocked alanine in explicit solvent. *Submitted to J. Chem. Phys. B*, 2006.

[25] John D. Chodera, William C. Swope, Jed W. Pitera, and Ken A. Dill. Long-time protein folding dynamics from short-time molecular dynamics simulations. *Multiscale Model. Simul.*, Special Issue on Biomolecular Simulation, to appear, 2006.

[26] John D. Chodera, William C. Swope, Jed W. Pitera, Chaok Seok, and Ken A. Dill. Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. *Submitted to J. Chem. Theor. Comput.*, 2006.

[27] Bruce W. Church and David Shalloway. Top-down free-energy minimization on protein potential energy landscapes. *Proc. Natl. Acad. Sci. USA*, 98(11):6098–6013, 2001.

[28] Andrea G. Cochran, Nicholas J. Skelton, and Melissa A. Starovasnik. Tryptophan zippers: Stable, monomeric β-hairpins. *Proc. Natl. Acad. Sci.*, 98(10):5578–5583, 2001.

[29] Julio F. Cordero-Morales, Luis G. Cuello, and Eduardo Perozo. Voltage-dependent gating at the KcsA selectivity filter. *Nat. Struct. Mol. Biol.*, 13(4):319–322, 2006.

[30] Julio F. Cordero-Morales, Luis G. Cuello, Yanxiang Zhao, Vishwanath Jogini, D. Marien Cortes, Benoît Roux, and Eduardo Perozo. Molecular determinants of gating at the potassium-channel selectivity filter. *Nat. Struct. Mol. Biol.*, 13(4):311–318, 2006.

[31] Glen Cowan. *Statistical Data Analysis*. Oxford University Press, 1998.

[32] Ryszard Czerminski and Ron Elber. Reaction path study of conformational transitions in flexible systems: Application to peptides. *J. Chem. Phys.*, 92(9):5580–5601, 1990.

[33] William Dably-Brown, Henrik H. Hansen, Mads P. G. Korsgaard, Naheed Mirza, and Soren-P Olesen. K_v7 channels: Function, pharmacology and channel modulators. *Curr. Topics Med. Chem.*, 6:999–1023, 2006.

[34] T. A. Darden, D. M. York, and L. G. Pedersen. Particle mesh Ewald – an $n \cdot \log(n)$ method for Ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10092, 1993.

[35] Bert L. de Groot, Xavier Daura, Alan E. Mark, and Helmut Grubmüller. Essential dynamics of reversible peptide folding: Memory-free conformational dynamics governed by internal hydrogen bonds. *J. Mol. Biol.*, 309:299–313, 2001.

[36] Christoph Dellago, Peter G. Bolhuis, and David Chandler. Efficient transition path sampling: Application to the Lennard-Jones cluster rearrangements. *J. Chem. Phys.*, 108(22):9236–9245, 1998.

[37] Christoph Dellago, Peter G. Bolhuis, Félix S. Csajka, and David Chandler. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.*, 108(5):1964, 1998.

[38] Susan D. Demo and Gary Yellen. The inactivation gate of the *Shaker* k^+ channel behaves like an open-channel blocker. *Neuron*, 7:743–753, 1991.

[39] Florin Despa and R. Stephen Berry. Inter-basin dynamics on multidimensional potential surfaces. i. escape rates on complex basin surfaces. *J. Chem. Phys.*, 115(18):8274–8278, 2001.

[40] Florin Despa, Ariel Fernández, and R. Stephen Berry. Interbasin motion approach to dynamics of conformationally constrained peptides. *J. Chem. Phys.*, 118(12):5673–5682, 2003.

[41] P. Deuflhard, W. Huisinga, A. Fischer, and Ch. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra Appl.*, 315(1-3):39–59, August 2000.

[42] Markus Dittrich and Klaus Schulten. PcrA helicase, a prototype ATP-driven molecular motor. *Structure*, 14:1345–1353, 2006.

[43] Christopher M. Dobson. Protein folding and misfolding. *Nature*, 426:884–890, 2003.

[44] Rose Du, Vijay S. Pande, Alexander Yu. Grosberg, Toyoichi Tanaka, and Eugene S. Shakhnovich. On the transition coordinate for protein folding. *J. Chem. Phys.*, 108(1):334–350, 1998.

[45] Yong Duan, Chun Wu, Shibasish Chowdhury, Mathew C. Lee, Guoming Xiong, Wei Zhang, Rong Yang, Piotr Cieplak, Ray Luo, Taisung Lee, James Caldwell, Junmei Wang, and Peter Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, 24:1999–2012, 2003.

[46] H. Jane Dyson and Peter E. Wright. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, 6(3):197–208, 2005.

[47] Weinan E, Weiqing Ren, and Eric Vanden-Eijnden. String method for the study of rare events. *Phys. Rev. B*, 66:052301, 2002.

[48] Barbara Eckert, Andreas Martin, Jochen Balback, and Franz X. Schmid. Prolyl isomerization as a molecular timer in phage infection. *Nat. Struct. Mol. Biol.*, 12(7):619–623, 2005.

[49] B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 1979.

[50] Elan Zohar Eisenmesser, Daryl A. Bosco, Mikael Akke, and Dorothee Kern. Enzyme dynamics during catalysis. *Science*, 295:1520–1523, 2002.

[51] Sidney P. Elmer, Sanghyun Park, and Vijay S. Pande. Foldamer dynamics expressed via Markov state models. II. State space decomposition. *J. Chem. Phys.*, 123:114903, 2005.

[52] David A. Evans and David J. Wales. Folding of the GB1 hairpin peptide from discrete path sampling. *J. Chem. Phys.*, 112(2):1080, 2004.

[53] H. G. Evertz. The loop algorithm. *Advances in Physics*, 52(1):1–66, 2003.

[54] Aton K. Faradjian and Ron Elber. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.*, 120(23):10880–10889, 2004.

[55] Fatih Yaşar, Tarik Çelik, Bernd A. Berg, and Hagai Meirovitch. Multicanonical procedure for continuum peptide models. *J. Comput. Chem.*, 21:1251–1261, 2000.

[56] Alan M. Ferrenberg, D. P. Landau, and Robert H. Swendsen. Statistical errors in histogram reweighting. *Phys. Rev. E*, 51(5):5092–5099, 1995.

[57] Alan M. Ferrenberg and Robert H. Swendsen. Optimized Monte Carlo data analysis. *Phys. Rev. Lett.*, 63(12):1195–1198, 1989.

[58] Alan M. Ferrenberg and Robert H. Swendsen. New Monte Carlo technique for studying phase transitions. *Phys. Rev. Lett.*, 61(23):2635–2638, 1998.

[59] Alan R. Fersht. On the simulation of protein folding by short time scale molecular dynamics and distributed computing. *Proc. Nat. Acad. Sci. USA*, 99:14122–14125, 2002.

[60] Alexander Fischer. *An Uncoupling-Coupling Method for Markov Chain Monte Carlo Simulations with an Application to Biomolecules*. PhD thesis, Institute of Mathematics II, Free University Berlin, 2003.

[61] Alexander Fischer, Frank Cordes, and Christof Schütte. Hybrid Monte Carlo with adaptive temperature in mixed-canonical ensemble: Efficient conformational analysis of RNA. *J. Comput. Chem.*, 1999:1689–1697, 1998.

[62] B. G. Fitch, R. S. Germain, M. Mendell, J. Pitera, M. Pitman, A. Rayshubskiy, Y. Sham, F. Suits, W. Swope, T. J. C. Ward, Y. Zhestkov, and R. Zhou. Blue Matter, an application framework for molecular simulation on Blue Gene. *J. Parallel Distrib. Comput.*, 63:759–773, 2003.

[63] Marshall Fixman. Classical statistical mechanics of constraints: A theorem and application to polymers. *Proc. Natl. Acad. Sci. USA*, 71(8):3050–3053, 1974.

[64] H. Flyvbjerg and H. G. Petersen. Error estimates on averages of correlated data. *J. Chem. Phys.*, 91(1):461–466, 1989.

[65] Hans Frauenfelder, Benjamin H. McMahon, Robert H. Austin, Kevin Chu, and John T. Groves. The role of structure, energy landscape, dynamics, and allostery in the enzymatic function of myoglobin. *Proc. Nat. Acad. Sci. USA*, 98(5):2370–2374, 2001.

[66] T. Galliat, W. Huisinga, and P. Deuflhard. Self-organizing maps combined with eigenmode analysis for automated cluster identification. In *Proceeding of the ICSC Symposia on Neural Computation*, Berlin, Germany, May 2000.

[67] Tobias Galliat. *Adaptive Multilevel Cluster Analysis by Self-Organizing Box Maps*. PhD thesis, FU Berlin, 2002.

[68] Emilio Gallicchio, Michael Andrec, Anthony K. Felts, and Ronald M. Levy. Temperature weighted histogram analysis method, replica exchange, and transition paths. *J. Phys. Chem. B*, 109:6722–6731, 2005.

[69] A. Gara, M. A. Blumrich, D. Chen, G. L.-T. Chiu, P. Coteus, M. E. Giampapa, R. A. Haring, P. Heidelberger, D. Hoenicke, G. V. Kopcsay, T. A. Liebsch, M. Ohmacht, B. D. Steinmacher-Burow, T. Takken, and P. Vranas. Overview of the Blue Gene/L system architecture. *IBM J. Res. & Dev.*, 49(2/3):195–212, 2005.

[70] Angel E. García and Kevin Y. Sanbonmatsu. α-helical stabilization by side chain shielding of backbone hydrogen bonds. *Proc. Nat. Acad. Sci. USA*, 99:2782–2787, 2002.

[71] Phillip L. Geissler, Christoph Dellago, and David Chandler. Kinetic pathways of ion pair dissociation in water. *J. Phys. Chem. B*, 103:3706–3710, 1999.

[72] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Stat. Sci.*, 7(4):457–472, 1992.

[73] Robert S. Germain, Blake Fitch, Aleksandr Rayshubskiy, Maria Eleftheriou, Michael C. Pitman, Frank Suits, Mark Giampapa, and T.J. Christopher Ward. Blue Matter on Blue Gene/L: Massively parallel computation for biomolecular simulation. In *CODES+ISSS '05: Proceedings of the 3rd IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis*, pages 207–212, New York, NY, USA, 2005. ACM Press.

[74] Helmut Grubmüller and Paul Tavan. Molecular dynamics of conformational substates for a simplfied protein model. *J. Chem. Phys.*, 101(6):5047–5057, September 1994.

[75] Martin Gruebele, Jobiah Sabelko, Richard Ballew, and John Ervin. Laser temperature jump induced protein refolding. *Acc. Chem. Res.*, 31:699–707, 1998.

[76] Taekjip Ha, Xiaowei Zhuang, Harold D. Kim, Jeffrey W. Orr, James R. Williamson, and Steven Chu. Ligand-induced conformational changes observed in single rna molecules. *Proc. Natl. Acad. Sci. USA*, 96:9077–9082, 1999.

[77] Ulrich H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281:140–150, 1997.

[78] Ulrich H. E. Hansmann and Yuko Okamoto. Prediction of peptide conformation by multicanonical algorithm: New approach to the multiple-minima problem. *J. Comput. Chem.*, 14(11):1333–1338, 1993.

[79] Carsten Hartmann and Christof Schütte. Free energy calculations in many dimensions. *Submitted for publication*, 2004.

[80] Carsten Hartmann and Christof Schütte. A geometric approach to free energy calculations. *Submitted for publication*, 2004.

[81] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2001.

[82] Keiko Hirose, Erika Akimaru, Toshihiko Akiba, Sharyn A. Endow, and Linda A. Amos. Large conformational changes in a kinesin motor catalyzed by interaction with microtubules. *Mol. Cell*, 23:913–923, 2006.

[83] Wilhelm Huisinga. *Metastability of Markov systems: A transfer operator based approach in application to molecular dynamics*. PhD thesis, Free University of Berlin, Berlin, Germany, May 2001.

[84] Wilhelm Huisinga and Bernd Schmidt. *Advances in Algorithms for Macromolecular Simulation*, chapter Metastability and dominant eigenvalues of transfer operators. Lecture Notes in Computational Science and Engineering. Springer, 2005.

[85] Gerhard Hummer and Ioannis G. Kevrekidis. Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. *J. Chem. Phys.*, 118(23):10762–10773, June 2003.

[86] Tatyana I. Igumenova, Kendra King Frederick, and A. Joshua Wand. Characterization of the fast dynamics of protein amino acid side chains using NMR relaxation in solution. *Chem. Rev.*, 106:1672–1699, 2006.

[87] Wolfhard Janke. Statistical analysis of simulations: Data correlations and error estimation. In J Grotendorst, D Marx, and A Murmatsu, editors, *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*, volume 10, pages 423–445. John von Neumann Institute for Computing, 2002.

[88] Oleg Jardetzky. On the nature of molecular conformations inferred from high-resolution nmr. *Biochimica et Biophysica Acta*, 621:227–232, 1980.

[89] Sheila S. Jaswal, Stephanie M. E. Truhlar, Ken A. Dill, and David A. Agard. Comprehensive analysis of protein folding activation thermodynamics reveals a universal behavior violated by kinetically stable proteases. *J. Mol. Biol.*, 347:355–366, 2005.

[90] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926, 1983.

[91] Taketoshi Kambara and Mitsuo Ikebe. A unique ATP hydrolysis mechanism of single-headed processive myosin, myosin IX. *J. Biol. Chem.*, 281(8):4949–4957, 2006.

[92] Mary E. Karpen, Douglas J. Tobias, and Charles L. Brooks III. Statistical clustering techniques for the analysis of long molecular dynamics trajectories: Analysis of 2.2-ns trajectories of YPGDV. *Biochemistry*, 32:412–420, 1993.

[93] Allong Ke, Kalhong Zhou, Fang Ding, Jamie H. D. Cate, and Jennifer A. Doudna. A conformational switch controls hepatitis delta virus ribozyme catalysis. *Nature*, 429:201–205, 2004.

[94] M. Khalil, N. Demirdöven, and A. Tokmakoff. Coherent 2D IR spectroscopy: Molecular structure and dynamics in solution. *J. Phys. Chem. A*, 107:5258–5279, 2003.

[95] P. A. Kollman, R. Dixon, W. Cornell, T. Vox, C. Chipot, and A. Pohorille. The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of *a*b initio calculations and experimental data. In A. Wilkinson, P. Weiner, and W. F. van Gunsteren, editors, *Computer Simulation of Biomolecular Systems*, volume 3, pages 83–96. Kluwer/Escom, 1997.

[96] Dmitry A. Kondrashov, Qiang Cui, and George N. Phillips Jr. Optimization and evaluation of a coarse-grained model of protein motion using x-ray crystal data. *Biophys. J.*, 91:2760–2767, 2006.

[97] S. Kriminski, M. Kazmierczak, and R. E. Thorne. Heat transfer from protein crystals: Implications for flash-cooling and X-ray beam heating. *Acta Cryst.*, D59:697–708, 2003.

[98] Susanna Kube and Marcus Weber. Identification of metastabilities in monomolecular conformation kinetics. *ZIB-Report*, 2006.

[99] Jan Kubelka, James Hofrichter, and William A. Eaton. The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.*, 14:76–88, 2004.

[100] Shankar Kumar, Djamal Bouzida, Robert H. Swendsen, Peter A. Kollman, and John M. Rosenberg. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. *J. Comput. Chem.*, 13(8):1011–1021, 1992.

[101] Ralph E. Kunz and R. Stephen Berry. Statistical interpretation of topographies and dynamics of multidimensional potentials. *J. Chem. Phys.*, 103(5):1904–1912, 1995.

[102] Igor K. Lednev, Anton S. Karnoup, Mark C. Sparrow, and Sanford A. Asher. Transient UV Raman spectrocopy finds no crossing barrier between the peptide α-helix and fully random coil conformation. *J. Am. Chem. Soc.*, 123:2388–2392, 2001.

[103] Peter Lenz, Bojan Zagrovic, Jessica Shapiro, and Vijay S. Pande. Folding probabilities: A novel approach to folding transitions and the two-dimensional ising-model. *J. Chem. Phys.*, 120(14):6769–6778, April 2004.

[104] Irwin B. Levitan. Modulation of ion channels by protein phosphorylation and dephosphorylation. *Annu. Rev. Physiol.*, 56:193–212, 1994.

[105] Yaakov Levy, Joshua Jortner, and Oren M. Becker. Dynamics of hierarchical folding on energy landscapes of hexapeptides. *J. Chem. Phys.*, 115(22):10533–10547, 2001.

[106] Yaakov Levy, Joshua Jortner, and R. Stephen Berry. Eigenvalue spectrum of the master equation for hierarchical dynamics of complex systems. *Phys. Chem. Chem. Phys.*, 4:5052–5058, 2002.

[107] Kresten Lindorff-Larsen, Robert B. Best, Mark A. DePristo, Christopher M. Dobson, and Michele Vendruscolo. Simultaneous determination of protein structure and dynamics. *Nature*, 433:128–132, 2005.

[108] David J. Lockhart and Peter S. Kim. Internal Stark effect measurement of the electric field of the amino terminus of an α helix. *Science*, 257:947–951, 1992.

[109] David J. Lockhart and Peter S. Kim. Electrostatic screening of charge and dipole interactions with the helix backbone. *Science*, 260:198–202, 1993.

[110] A. P. Lyubartsev, A. A. Martsinovski, S. V. Shevkunov, and P. N. Vorontsov-Velyaminov. New approach to Monte Carlo calculation of the free energy: Method of expanded ensembles. *J. Chem. Phys.*, 96:1776–1783, 1992.

[111] Ao Ma and Aaron R. Dinner. Automatic method for identifying reaction coordinates in complex systems. *J. Phys. Chem. B*, 109:6769–6779, 2005.

[112] Hairong Ma and Martin Gruebele. Kinetics are probe-dependent during downhill folding of an engineered λ_{6-85} protein. *Proc. Nat. Acad. Sci. USA*, 102:2283–2287, 2005.

[113] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical statistics and probability*, pages 281–297. University of California Press, 1967.

[114] Nazli Maki, Karobi Moitra, Pratiti Ghosh, and Saibal Dey. Allosteric modulation bypasses the requirement for ATP hydrolysis in regenerating low affinitiy transition state conformation of human P-glycoprotein. *J. Biol. Chem.*, 281(16):10769–10777, 2006.

[115] Lidia M. Mannuzzu, Mario M. Moronne, and Ehud Y. Isacoff. Direct physical measure of conformational rearrangement underlying potassium channel gating. *Science*, 271(5246):213–216, 1996.

[116] Neelan J. Marianayagam, Nicolas L. Fawzi, and Teresa Head-Gordon. Protein folding by distributed computing and the denatured state ensemble. *Proc. Natl. Acad. Sci.*, 102(46):16684–16689, 2005.

[117] E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.*, 19(6):451–458, 1992.

[118] Eike Meerback, Christof Schütte, and Alexander Fischer. Eigenvalue bounds on restrictions of reversible nearly uncoupled Markov chains. *Lin. Alg. Appl.*, 398:141–160, 2005.

[119] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.

[120] Ayori Mitsutake and Yuko Okamoto. Replica-exchange simulated tempering method for simulations of frustrated systems. *Chem. Phys. Lett.*, 332:131–138, 2000.

[121] Ayori Mitsutake and Yuko Okamoto. Replica-exchange extensions of simulated tempering method. *J. Chem. Phys.*, 121(6):2491–2504, 2004.

[122] Ayori Mitsutake, Yuji Sugita, and Yuko Okamoto. Replica-exchange multicanonical and multicanonical replica-exchange monte carlo simulations of peptides. i. formulation and benchmark test. *J. Chem. Phys.*, 118(14):6664–6675, 2003.

[123] Daniele Moroni, Titus S. van Erp, and Peter G. Bolhuis. Investigating rare events by transition interface sampling. *Physica A*, 340:395–401, 2004.

[124] Paul N. Mortenson, David A. Evans, and David J. Wales. Energy landscapes of model polyalanines. *J. Chem. Phys.*, 117(3):1363–1376, 2002.

[125] Paul N. Mortenson and David J. Wales. Energy landscapes, global optimization and dynamics of the polyalanine Ac(ala)$_8$NHMe. *J. Chem. Phys.*, 114(14):6443–6454, 2001.

[126] Victor Mu?noz, Peggy A. Thompson, James Hofrichter, and William A. Eaton. Folding dynamics and mechanism of β-hairpin formation. *Nature*, 390:196–199, 1997.

[127] H. Müller-Krumbhaar and K. Binder. Dynamic properties of the Monte Carlo method in statistical mechanics. *J. Stat. Phys.*, 8(1):1–23, 1973.

[128] Nobuyuki Nakajima, Haruki Nakamura, and Akinori Kidera. Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. *J. Phys. Chem. B*, 101:817–824, 1997.

[129] M. E. J. Newman and R. G. Palmer. Error estimation in the histogram Monte Carlo method. *J. Stat. Phys.*, 97(5/6):1011–1026, 1999.

[130] Michael Nilges, Angela M. Gronenborn, Axel T. Brünger, and G. Marius Clore. Determination of the three-dimenional structures of proteins by simulated annealing with interproton distance restraints. application to crambin, potato carboxypeptidase inhibitor and barley serine protease inhibitor 2. *Protein Eng.*, 2(1):27–38, 1988.

[131] Frank Noé, Illia Horenko, Christof Schütte, and Jeremy C. Smith. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *Submitted to J. Chem. Phys.*, 2006.

[132] Hiroyuki Noji, Ryohel Yasuda, Masasuke Yoshida, and Kazuhiko Kinosita Jr. Direct observation of the rotation of f_1-ATPase. *Nature*, 386:299–302, 1997.

[133] James M. Ogle, Ditlev E. Brodersen, William M. Clemons Jr., Michael J. Tarry, Andrew P. Carter, and V. Ramakrishnan. Recognition of cognate transfer RNA by the 30S ribosomal subunit. *Science*, 292:897–902, 2001.

[134] Yuko Okamoto. Generalized-ensemble algorithms: Enhanced sampling techniques for monte carlo and molecular dynamics simulations. *J. Mol. Graph. Model.*, 22:425–439, 2004.

[135] Irwin Oppenheim, Kurt E. Shuler, and George H. Weiss, editors. *Stochastic processes in chemical physics: The master equation*. MIT Press, 1977.

[136] S. Banu Ozkan, Ivet Bahar, and Ken A. Dill. Transition states and the meaning of ϕ-values in protein folding kinetics. *Nature Struct. Biol.*, 8(9):765–769, 2001.

[137] S. Banu Ozkan, Ken A. Dill, and Ivet Bahar. Fast-folding protein kinetics, hidden intermediates, and the sequential stabilization model. *Protein Science*, 11:1958–1970, 2002.

[138] Emanuele Paci, Andrea Cavalli, Michele Vendruscolo, and Amedeo Caflisch. Analysis of the distributed computing approach applied to the folding of a small β peptide. *Proc. Natl. Acad. Sci. USA*, 100(14):8217–8222, 2003.

[139] Vijay S. Pande, Ian Baker, Jarrod Chapman, Sidney P. Elmer, Siraj Khaliq, Stefan M. Larson, Young Min Rhee, Michael R. Shirts, Christopher D. Snow, Eric J. Sorin, and Bojan Zagrovic. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, 68:91–109, 2003.

[140] Vijay S. Pande, Ian Baker, Jarrod Chapman, Sidney P. Elmer, Siraj Khaliq, Stefan M. Larson, Young Min Rhee, Michael R. Shirts, Christopher D. Snow, Eric J. Sorin, and Bokan Zagrovic. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. *Biopolymers*, 68:91–109, 2003.

[141] Sanghyun Park and Vijay S. Pande. Validation of Markov state models using Shannon's entropy. *J. Chem. Phys.*, 124:054118, 2006.

[142] Eduardo Perozo. Gating prokaryotic mechanosensitive channels. *Nat. Rev. Mol. Cell Biol.*, 7:109–119, 2006.

[143] Jed W. Pitera, Imran Haque, and William C. Swope. Absence of reptation in the high-temperature folding of the trpzip2 β-hairpin peptide. *J. Chem. Phys.*, 124:141102, 2006.

[144] Young Min Rhee and Vijay S. Pande. One-dimensional reaction coordinate and the corresponding potential of mean force from commitment probability distribution. *J. Phys. Chem. B*, 109:6780–6786, 2005.

[145] Elizabeth Rhoades, Mati Cohen, Benjamin Schuler, and Gilad Haran. Two-state folding observed in individual protein molecules. *J. Am. Chem. Soc.*, 126:14686–14687, 2004.

[146] Wolfgang Rieping, Michael Habeck, and Michael Nilges. Inferential structure determination. *Science*, 309:303–306, 2005.

[147] I. K. Roterman, M. H. Lambert, K. D. Gibson, and H. A. Scheraga. A comparison of the CHARMM, AMBER and ECEPP potentials for peptides. II. $\phi - \psi$ maps for N-acetyl alanine N'-methyl amide: Comparisons, contrasts and simple experimental tests. *J. Biomol. Struct. Dyn.*, 78(3):421, 1989.

[148] J. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *J. Comput. Phys.*, 23(3):327–341, 1977.

[149] K. Y. Sanbonmatsu. Energy landcape of the ribosomal decoding center. *Biochimie*, 88:1053–1059, 2006.

[150] K. Y. Sanbonmatsu and A. E. Garcia. Structure of met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics. *Proteins: Structure, Function, and Genetics*, 46:225–234, 2002.

[151] Verena Schultheis, Thomas Hirschberger, Heiko Carstens, and Paul Tavan. Extracting markov models of peptide conformational dynamics from simulation data. *J. Chem. Theor. Comput.*, 2005.

[152] Ch. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on Hybrid Monte Carlo. *J, Comput. Phys.*, 151:146–168, 1999.

[153] Ch. Schütte and W. Huisinga. Biomolecular conformations can be identified as metastable states of molecular dynamics. In P. G. Ciaret and J.-L. Lions, editors, *Handbook of Numerical Analysis - special volume on computational chemistry*, volume X. Elsevier, 2002.

[154] Christof Schütte. *Conformational dynamics: Modelling, theory, algorithm, and application to biomolecules*. PhD thesis, Konrad Zuse Zentrum Berlin, Berlin, Germany, 1999.

[155] Christof Schütte and Wilhelm Huisinga. Biomolecular conformations can be identified as metastable sets of molecular dynamics. In P. G. Ciaret and J.-L. Lions, editors, *Handbook of Numerical Analysis - special volume on computational chemistry*. Elsevier, in press.

[156] David Shalloway. Macrostates of classical stochastic systems. *J. Chem. Phys.*, 105(22):9986–10007, 1996.

[157] Min-yi Shen and Karl F. Freed. Long time dynamics of met-enkephalin: Tests of mode-coupling theory and implicit solvent models. *J. Chem. Phys.*, 118(11):5143–5156, 2003.

[158] Michael Shirts and Vijay S. Pande. Screen savers of the world unite! *Science*, 290(5498):1903–1904, December 2000.

[159] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, New York, NY, 1986.

[160] Nina Singhal, John D. Chodera, Jed W. Pitera, Vijay S. Pande, Ken A. Dill, and William C. Swope. An automatic state decomposition method for the construction of discrete-state markov models of protein dynamics. *Manuscript in preparation.*, 2006.

[161] Nina Singhal and Vijay S. Pande. Error analysis and efficient sampling in Markovian state models for molecular dynamics. *J. Chem. Phys.*, 123:204909, 2005.

[162] Nina Singhal, Christopher D. Snow, and Vijay S. Pande. Using path sampling to build better markovian state models: Predicting the folding rate and mechanism of a tryptophan zipper beta hairpin. *J. Chem. Phys.*, 121(1):415–425, 2004.

[163] Eric J. Sorin and Vijay S. Pande. Empirical force-field assessment: The interplay between backbone torsions and noncovalent term scaling. *J. Comput. Chem.*, 26:682–690, 2005.

[164] Eric J. Sorin and Vijay S. Pande. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. *Biophys. J.*, 88:2472–2493, 2005.

[165] Marc Souialle and Benoît Roux. Extension to the weighted histogram analysis method: combining umbrella sampling with free energy calculations. *Comp. Phys. Commun.*, 135:40–57, 2001.

[166] Chris A. E. M. Spronk, Sander B. Nabuurs, Alexandre M. J. J. Bonvin, Elmar Krieger, Geerten W. Vuister, and Gert Vriend. The precision of NMR structure ensembles revisited. *J. Biomol. NMR*, 25:225–234, 2003.

[167] Saravanapriyan Sriraman, Ioannis G. Kevrekidis, and Gerhard Hummer. Coarse master equation from Bayesian analysis of replica molecular dynamics simulations. *J. Phys. Chem. B*, 109:6479–6484, 2005.

[168] Peter J. Steinbach, Roxana Ionescu, and C. Robert Matthews. Analysis of kinetics using a hybrid maximum-entropy-nonlinear-least-squares method: Application to protein folding. *Biophys. J.*, 82:2244–2255, 2002.

[169] Boris Steipe. A revised proof of the metric properties of optimally superimposed vector sets. *Acta Cryst.*, A58:506, 2002.

[170] Lubert Stryer. *Biochemistry*. W. H. Freeman, New York, 2002.

[171] Yugi Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141–151, 1999.

[172] Yuji Sugita and Yuko Okamoto. Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape. *Chem. Phys. Lett.*, 2329:261–270, 2000.

[173] William C. Swope, Hans C. Andersen, Peter H. Berens, and Kent R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.*, 76(1):637–649, 1982.

[174] William C. Swope, Jed W. Pitera, and Frank Suits. Describing protein folding kinetics by molecular dynamics simulations: 1. Theory. *J. Phys. Chem. B*, 108:6571–6581, 2004.

[175] William C. Swope, Jed W. Pitera, Frank Suits, Mike Pitman, Maria Eleftheriou, Blake G. Fitch, Robert S. Germain, Aleksandr Rayshubski, T. J. C. Ward, Yuriy Zhestkov, and Ruhong Zhou. Describing protein folding kinetics by molecular dynamics simulations: 2. Example applications to alanine dipeptide and a beta-hairpin peptide. *J. Phys. Chem. B*, 108:6582–6594, 2004.

[176] Karen E. S. Tang and Ken A. Dill. How experiments see fluctuations of native proteins: Perspective from an exact model. *Intl. J. Quantum Chem.*, 75:147–164, 1999.

[177] John R. Taylor. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*, chapter 3.11, pages 73–79. University Science Books, 2nd ed. edition, 1996.

[178] Douglas L. Theobald. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. *Acta Cryst.*, A61:478–480, 2005.

[179] Peggy A. Thompson, William A. Eaton, and James Hofrichter. Laser temperature jump study of the helix⇌coil kinetics of an alanine peptide interpreted with a 'kinetic zipper' model. *Biochem.*, 36:9200–9210, 1997.

[180] Peggy A. Thompson, Victor Muñoz, Gouri S. Jas, Eric R. Henry, William A. Eaton, and James Hofrichter. The helix-coil kinetics of a heteropeptide. *J. Phys. Chem. B*, 104:378–389, 2000.

[181] Makoto Tominaga and Michael J. Caterina. Thermosensation and pain. *J. Nerobiol.*, 61:3–12, 2004.

[182] Donald G. Truhlar, Bruce C. Garrett, and Stephen J. Klippenstein. Current status of transition-state theory. *J. Phys. Chem.*, 100:12771–12800, 1996.

[183] V. Tsui and D. A. Case. Molecular dynamics simulations of nucleic acids using a generalized Born solvation model. *J. Am. Chem. Soc.*, 122:2489–2498, 2000.

[184] M. Tuckerman, B. J. Berne, and G. J. Martyna. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.*, 97(3):1990–2001, 1992.

[185] Mark E. Tuckerman, Yi Liu, Giovanni Ciccotti, and Glenn J. Martyna. Non-Hamiltonian molecular dynamics: Generalizing Hamiltonian phase space principles to non-Hamiltonian systems. *J. Chem. Phys.*, 115(4):1678–1702, 2001.

[186] Alex Ulitsky and David Shalloway. Variational calculation of macrostate transition rates. *J. Chem. Phys.*, 109(5):1670–1686, August 1998.

[187] Alex Ulitsky and David Shalloway. Erratum on "variational calculation of macrostate transition rates". *J. Chem. Phys.*, 110(10):4975, March 1999.

[188] Debbie van den Hemel, Ann Brigé, Savvas N. Savvides, and Jozef Van Beeumen. Ligand-induced conformational changes in the capping subdomain of bacterial old yellow enzyme homologue and conserved sequence fingerprints provide new insights into substrate binding. *J. Biol. Chem.*, 281(38):28152–28161, 2006.

[189] Titus S. van Erp, Daniele Moroni, and Peter G. Bolhuis. A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.*, 118(17):7762–7774, 2003.

[190] N. G. van Kampen. *Stochastic processes in physics and chemistry*. Elsevier, second edition, 1997.

[191] Arthur F. Voter and Jimmie D. Doll. Dynamical corrections to transition state theory for multistate systems: Surface self-diffusion in the rare-event regime. *J. Chem. Phys.*, 82(1):80–92, 1985.

[192] M. P. Wand. Data-based choice of histogram bin width. *The American Statistician*, 51(1):59, 1997.

[193] Junmei Wang, Piotr Cieplak, and Peter A. Kollman. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J. Comput. Chem.*, 21(12):1049–1074, 2000.

[194] Marcus Weber. *Meshless methods in conformation dynamics*. PhD thesis, Free University of Berlin, 2006.

[195] M. Weik, G. Kryger, A. M. M. Schreurs, B. Bouma, I. Silman, J. L. Sussman, P. Gros, and J. Kroon. Solvent behavior in flash-cooled protein crystals at cryogenic temperatures. *Acta Cryst.*, D57:566–573, 2001.

[196] Keith E. Willard and Donald P. Connelly. Nonparametric probability density estimation: Improvements to the histogram for laboratory data. *Computers and Biomedical Research*, 25:17–28, 1992.

[197] Skip Williams, Timothy P. Causgrove, Rudolf Gilmanshin, Karen S. Fang, Robert H. Callender, William H. Woodruff, and R. Brian Dyer. Fast events in protein folding: Helix melting and formation in a small peptide. *Biochem.*, 35:691–697, 1996.

[198] Gerburg Wulf, Greg Finn, Futoshi Suizu, and Kun Ping Lu. Phosphorylation-specific prolyl isomerization: is there an underlying theme? *Nature Cell Biol.*, 7(5):435–441, 2005.

[199] Wei Yuan Yang and Martin Gruebele. Detection-dependent kinetics as a probe of folding landscape microstrucure. *J. Am. Chem. Soc.*, 126:7758–7759, 2004.

[200] Ben Youngblood and Norbert O. Reich. Conformational transitions as determinants of specificity for the DNA methyltransferase EcoRI. *J. Biol. Chem.*, 281(37):26821–26831, 2006.

[201] Bojan Zagrovic, Christopher D. Snow, Michael R. Shirts, and Vijay S. Pande. Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. *J. Mol. Biol.*, 323:927–937, 2002.

[202] Wei Zhang, Hongxing Lei, Shibasish Chowdbury, and Yong Duan. Fs-21 peptides can form both single helix and helix-turn-helix. *J. Phys. Chem. B*, 108:7479–7489, 2004.

[203] Xiaowei Zhuang, Harold Kim, Miguel J. B. Pereira, Hazen P. Babcock, Nils G. Walter, and Steven Chu. Correlating structural dynamics and function in single ribozyme molecules. *Science*, 296:1473–1476, 2002.

[204] Robert Zwanzig and Narinder K. Ailawadi. Statistical error due to finite time averaging in computer experiments. *Phys. Rev.*, 182(1):280–283, 1969.