PhD Thesis
Master equation models of macromolecular dynamics from atomistic simulation

John D. Chodera
University of California, San Francisco
San Francisco, CA 94143

This document is available at

`http://www.dillgroup.ucsf.edu/~jchodera/thesis/`

The author can be reached at

`jchodera@gmail.com`

Master equation models of macromolecular dynamics from atomistic simulation

A 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

Fall, 2006

by

John D. Chodera

B.S. (California Institute of Technology) 1999

Committee in charge:

Chair: Professor Ken A. Dill, UCSF
Professor Matthew P. Jacobson, UCSF
Professor Vijay S. Pande, Stanford University

The research reported in this thesis was carried out at the University of California, San Francisco.

Dedicated to the memory of Peter A. Kollman,

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

Contents

List of Symbols

Page numbers refer to where the symbols are introduced.

General notation

$\langle A \rangle_\beta$	configuration space average in the canonical ensemble at inverse temperature β	**??**
$\langle A \rangle$	mathematical expectation, such as average over repeated realizations of an experiment	**??**
$\overline{A}$	time average over a trajectory	**??**
$\hat{A}$	estimator of $\langle A \rangle$	**??**
$\delta^2 \hat{A}$	squared uncertainty in $\hat{A}$	**??**

Vectors

$\mathbf{q}$	coordiantes	
$\mathbf{p}$	momenta	
$\mathbf{z}$	phase space point $\mathbf{z} = (\mathbf{q}, \mathbf{p})$	

Scalars

E	total energy	9
T	kinetic energy	**??**
V	potential energy	**??**

Scalar functions

$H(\mathbf{z}) = H(\mathbf{q}, \mathbf{p})$	Hamiltonian	9
$T(\mathbf{p})$	kinetic energy	**??**
$V(\mathbf{q})$	Hamiltonian	**??**
$\Omega(E)$	total energy density of states	10
$\Omega(T)$	kinetic energy density of states	
$\Omega(V)$	potential energy density of states	
$Z(\beta)$	canonical partition function	11

Vector-valued functions

$\boldsymbol{F}(\mathbf{q})$	force	7

Operators

$\mathcal{L}$	Liouvillean	13
$\mathcal{P}_t$	propagator	14

Abbreviations

WHAM	weighted histogram analysis method
PMF	potential of mean force

Introduction

Perspective

Conformational dynamics is integral to 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 [143]. 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 [38]. 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 [73]. Conformational changes of folded proteins can be critical for protein [65] or substrate binding [157] and catalytic function [14, 44, 169]. Together with order-disorder transitions [41], ordered sequences of conformational changes are integral to the transformation of chemical energy into mechanical work by motor proteins [37,76] or the reverse by ATP synthases [109]. 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 [19, 54, 98]. Slow prolyl *cis-trans* isomerization dynamics can be used as a molecular timer [42], which can be modulated by phosphorylation events for signaling purposes [165]. Ion channel gating, an inherently stochastic kinetic process involving conformational switching between multiple conductance states [28, 99], can be modulated by transmembrane voltage [13, 27], ligand binding at allosteric sites [31], temperature [151], mechanical pressure [118], and phosphorylation [89], and undergo time-dependent inactivation [33]. The conformational plasticity of RNA no doubt contributes to its many characterized biological roles in information transmission [110, 125], binding [61], and catalysis [78,173]. 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 bio-

logically relevant conformational fluctuations [7, 82], 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 [163] over times sufficiently long to allow for significant structural rearrangement [83]. 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 [147]) difficult, though work continues on improving this situation [92, 122]. Standard NMR refinement protocols [108] 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 [72]. 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 [138]. 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 [70], 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) [79] 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 [60]), 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 [121]. 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 [150], sums of exponentials [107, 141, 149, 167], and so-called *stretched exponentials* [96]. 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 [60]). 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 [58], 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 [115, 133] 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 [50, 100].

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 [36, 51, 51, 55, 56, 67, 127, 128] (and later Weber and Kube [84, 162]), as well as independent work by Shalloway and coworkers [25, 131, 155, 156], and Berry and coworkers [9, 34, 35, 91], 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 [5, 8, 10, 90, 91]. 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} [39] that aid in the understanding of mechanism [88], and the summarization of primary events in kinetics processes such as folding, most notably seen in the work of Banu Ozkan and Ken Dill [113, 114].

Synopsis

The aim of this dissertation is twofold.

First, we develop a statistical mechanical framework for modeling macromolecular conformational dynamics as stochastic transitions between metastable states for use as both a conceptual theoretical model for thinking about dynamics, interpreting and deigning experiments (both computational and experimental), and constructing efficient simulation algorithms. This framework provides a rigorous connection between the master equation model, which is often treated as a purely phenomenological construct, and the classical statistical dynamics governing the evolution of macromolecular systems, and allows the necessary and sufficient conditions under which the model faithfully represents dynamics to be enumerated.

Second, we build upon this framework by designing efficient algorithms for the construction of these stochastic models from atomistic molecular dynamics simulations in explicit solvent, in order to provide tools for the investigation of these dynamical processes and dynamics pictures from which experimental hypotheses can be generated. Just as structural models derived from X-ray or NMR data drove a revolution in biophysical and biochemical experimentation, these tools offer the possibility of generating models that are dynamical in nature, from which new insight into mechanistic models and experimental hypotheses may come. The eventual development of forcefields of sufficient accuracy to reliably model macromolecular interactions may even elevate molecular simulation to a tool on par with and complementary to existing experimental methods.

Chapter 1 contains a review the classical statistical mechanics describing the equilibrium and dynamical behavior of solvated biomacromolecular systems. Some readers may find this material elementary, but notation used throughout this dissertation is introduced.

Chapter 2 presents the case for modeling macromolecular dynamics as a series of stochastic transitions between metastable states, and sets out the formal theory for deriving this model from the underlying statistical mechanics. One approach based on the eigenfunction expansion of the Liouvillean is sketched, as it is useful for providing insight into experiments and simulations, and a detailed derivation using the Mori-Zwanzig projection operator formalism is presented. Various properties of the resulting master equation model are enumerated, and a compendium of useful things that can easily be computed from it given.

Chapter 3 is concerned with how these models, once constructed from simulation data, can be validated to determine the timescale for emergence of Markovian behavior. An model system where dynamics of the model can be compared directly with numerous long simulations,

terminally blocked alanine in explicit solvent, is introduced, and will be a recurring example used for illustrative purposes.

Chapter 4 describes progress toward algorithms for the discovery of metastable states without prior knowledge of the relevant degrees of freedom.

Chapter 5 focuses on the issue of efficiently computing transition probabilities between these states, once defined, and introduces the formalism of the *transition energy density of states*. Through this formalism, and exploiting an analogue of the weighted histogram analysis method, data from static equilibrium simulations, equilibrium trajectories, trajectories initiated from equilibrium distributions over the states, and trajectories harvested from transition path sampling simulations can be combined to obtain an optimal estimate of the transition probabilities and rates. Temperature-dependent transition probabilities and rates can also be obtained.

Finally, Chapter 6 suggests ways in which these pieces of technology might be combined to construct a multistage or iterative procedure for automatically constructing master equation models in an efficient manner given whatever structural data may be initially available.

Two appendices provide a review of the derivation of the weighted histogram analysis method, which is exploited extensively in this thesis, and a primer on the calculation of statistical uncertainties from molecular simulations.

1 The statistical mechanics of macromolecular dynamics

1.1 Modeling the macromolecular system

Consider a biomacromolecule in a solvated environment under typical biological or experimental conditions. Our interest lies in the conformational dynamics of this macromolecule. We presume the system can be treated entirely within the realm of classical statistical mechanics for the following reasons:

1. Because atomic nuclei are at least three orders of magnitude more massive than the electrons, the electronic degrees of freedom relax much more rapidly than the nuclei move, and can be considered as always occupying their ground state in the field of the nuclei. This is the celebrated Born-Oppenheimer approximation [16].

2. Quantized vibrations and zero-point energy effects are minimal due to the thermal kinetic energies under biological conditions[1].

3. In restricting our consideration to conformational dynamics and not chemical reactions, we require that chemical bonds are neither broken nor formed. This ensures that we do not need to treat nuclear wavefunctions, as might be the case in enzyme-catalyzed hydrogen transfer reactions [62].

The interatomic interactions can be completely described by some potential function $V(\mathbf{q})$ that depends only on the Cartesian coordinates[2] of the atomic nuclei $\mathbf{q} \in \Re^{3N}$. This potential is sometimes termed a *forcefield* due to the fact that it is the field whose negative gradient yields the instantaneous interatomic forces:

$$\boldsymbol{F}(\mathbf{q}) \quad = \quad -\nabla_{\mathbf{q}} V(\mathbf{q}) \tag{1.1}$$

The exact form of the potential is not of interest here, but there are several popular *molecular mechanics forcefields* in common use today, such as AMBER [29, 40, 81, 161], CHARMM [75], and GROMOS [111].

[1] Comparison with experiment will occasionally still require correction terms to account for quantum effects that cannot be treated classicaly, such as quantized vibration effects, as in [66]

[2] We note that this exposition could be extended to allow the use of generalized coordinates and momenta, but this would necessitate the use of a Jacobian correction factor in the resulting phase space densities and partition functions. Therefore, for simplicity, we leave this extension as an exercise to the reader, should it be necessary.

Although macromolecular crowding is known to have an effect on thermodynamics and conformational dynamics [20, 45, 102, 159, 171, 172], and many macromolecules primarily exist in complexes or assemblies [124], we restrict our consideration here to a single independent macromolecule[3]. This situation corresponds to a solution of macromolecules sufficiently dilute such that each can be considered as independent or noninteracting, which may be the case in some, but not all, experiments. Typically, the macromolecule is surrounded by a large amount of solvent (generally water) under near-atmospheric external pressures. In biological systems, as well as many experiments, the salt concentration ensures that some mobile counterions will surround the macromolecule[4]. Many macromolecules, especially nucleic acids, carry a net charge, so the nearby presence counterions is almost ensured so as to preserve net macroscopic neutrality. The environment in biological and experimental contexts is generally buffered to maintain near-constant pH, and the protonation states of titratable groups may change in response to conformational changes, ligand binding, or post-translational modifications, a switching behavior that is relevant to the biological function of several systems [140]. Additionally, the system might include the presence of small ligands or cosolvent molecules whose association with the macromolecule may have significant impact on its conformational dynamics.

Here, we only consider Hamiltonian systems *without* holonomic constraints. In molecular simulations, it is common practice to constrain bond lengths to hydrogen in the macromolecule [3, 123] or the water model [74]. These fast vibrations otherwise constrain the longest timestep that can be used to achieve stable dynamics to approximately 1 fs at 300 K. Bond constraints introduce a coupling between the coordinates and momenta, resulting in non-Hamiltonian dynamics with correspondingly more complex statistical mechanical properties[5] [53, 63, 154]. We will not consider constrained systems in the theoretical framework

[3]While the methods presented here may be extended to more complex systems of multiple interacting macromolecules and ligands, we restrict ourselves to consideration of a single macromolecule here, for simplicity.

[4]Standard practice in molecular simulations is to add only the minimal number of individual counterions to bring the system to net neutrality. This practice may result in an atypical effective salt concentration that is not representative of either *in vitro* or *in vivo* conditions. As an example, a recent simulation of the trpzip2 hairpin [119] employed a cubic simulation box (48 Å)^3 and a single chloride ion, for an effective salt concentration of 7–15 mM. The corresponding experiments [168] employed a 50 mM potassium phosphate buffer. A larger system, such as the 80-residue λ_{6-85}, might require an (88 Å)^3 simulation box, where the addition of a single neutralizing chloride counterion would give an effective salt concentration of 1–2 mM, while experiments employing 50 mM phosphate buffer [96] would require, on average, 21 pairs of counterions. In this system, addition or subtraction of a pair of neutralizing counterions only changes the effective salt concentration by 2 mM, so we might expect a broad distribution of counterion numbers if we examine comparable volumes in the experimental system. It should be noted that even this large simulation volume gives an unphysically large protein concentration — 15 mM for trpzip2 and 2 mM for λ_{6-85}, while the corresponding experiments employ concentrations of 1–5 μM [166] or 50–360 μM [96].

[5]In particular, the canonical partition function is no longer separable into a product of independent partition functions for coordinates and momenta. There is also some disagreement about whether molecular dynamics and Metropolis Monte Carlo simulations conducted in Cartesian coordinates require expectations to be estimated by including a weighting term for each snapshot based on the metric tensor factor. Simulation codes that conduct dynamics in torsion space, however, have employed such a correction for a number of years. This term can be computed from the determinant of the metric tensor which can be computed by an expression originally derived by Fixman [53] or from the constraint forces in constraint schemes based upon Lagrange multiplier methods [153]. This factor is universally ignored in standard simulation practice, though there is insufficient evidence to suggest whether this is, in fact, reasonable. An early study by Chandler and Berne demonstrated a significant discrepancy between the uncorrected and Fixman-corrected potential of mean force of the torsion angle of n-butane in a simulation where bond lengths and angles were constrained by SHAKE [18].

developed here[6].

We define the *Hamiltonian* of the system $H(\mathbf{z}) = H(\mathbf{q}, \mathbf{p}) = T(\mathbf{q}) + V(\mathbf{p})$ to denote the sum of the kinetic energy $T(\mathbf{p}) = \frac{1}{2}\mathbf{p}^{\mathrm{T}}\boldsymbol{M}^{-1}\mathbf{p}$, with $\boldsymbol{M}$ the diagonal mass matrix, and the potential energy $V(\mathbf{q})$. The value of the Hamiltonian at any phase space point $\mathbf{z}$ therefore gives the *total energy* of the system, $E = T + V$. In the absence of any thermostat or energy exchange, the system evolves according *Hamilton's equations of motion* which can be expressed succinctly in terms of the Hamiltonian as

$$\begin{aligned} \dot{\mathbf{q}} &= \nabla_{\mathbf{p}} H(\mathbf{q}, \mathbf{p}) \\ \dot{\mathbf{p}} &= -\nabla_{\mathbf{q}} H(\mathbf{q}, \mathbf{p}) \end{aligned} \tag{1.2}$$

where the dot denote differentiation with respect to time.

It can easily be seen that Hamilton's equations of motion reduce to an expression of Newton's Second Law:

$$\begin{aligned} \dot{\mathbf{q}} &= \nabla_{\mathbf{p}}[T(\mathbf{p}) + V(\mathbf{q})] = \boldsymbol{M}^{-1}\mathbf{p} \\ \dot{\mathbf{p}} &= -\nabla_{\mathbf{q}}[T(\mathbf{p}) + V(\mathbf{q})] = -\nabla_{\mathbf{q}} V(\mathbf{q}) = \boldsymbol{F}(\mathbf{q}) \end{aligned} \tag{1.3}$$

Because $\mathbf{p} = \boldsymbol{M}\boldsymbol{v}$, where $\boldsymbol{v}$ denotes the velocity, and $\dot{\mathbf{p}} = \boldsymbol{M}\dot{\boldsymbol{v}} = \boldsymbol{M}\boldsymbol{a}$, where $\boldsymbol{a}$ denotes the acceleration, we recover the familiar equation $\boldsymbol{F} = \boldsymbol{M}\boldsymbol{a}$.

1.2 Equilibrium statistical ensembles

The equilibrium distribution of the macromolecule and its environment can be modeled by several statistical mechanical ensembles. We shall give an overview here.

1.2.1 Equivalence of ensembles

[General comments on the equivalence of all thermodynamic ensembles in the limit of large systems.]

1.2.2 Microcanonical (NVE)

In an isolated, confined system of constant volume that does not exchange energy with its environment, the total system energy E will be constant. For sufficiently complex systems, the ergodic theorem will hold, and the macromolecule will be able to access all relevant regions of phase space [80].

The equilibrium phase space distribution function is given by

$$\begin{aligned} \rho(\mathbf{z}) &= [\Omega(E)]^{-1}\, \delta(H(\mathbf{z}) - E) \\ \Omega(E) &\equiv \int d\mathbf{z}\, \delta(H(\mathbf{z}) - E) \end{aligned} \tag{1.4}$$

[6] In subsequent sections, simulation data will be presented where these constraints were employed out of necessity due to the lack of adequate implementations of multiple timestep integrators in common molecular mechanics packages, as well as water models that have been parameterized for fully flexible use. It is hoped this situation can be rectified in the near future. If performance is a concern in unconstrained systems, multiple timestep integrators derived over a decade ago [153] allow for unconstrained simulations to be run without incurring the large computational penalty otherwise encountered.

Here, $\Omega(E)$ denotes the *total energy density of states*, which is a measure of the volume of phase space with total energy E, and $\delta(x)$ is the Dirac delta function.

The temperature can be determined by

$$\beta^{-1} \quad = \quad \left(\frac{\partial \Omega}{\partial E}\right) \tag{1.5}$$

In principle, we should write $\Omega(E, V, N)$, as the density of states also depends on the system volume and number of particles, but we suppress these variables in general unless we consider their variation, as in the calculation of pressure:

$$p \quad = \quad \beta \left(\frac{\partial \Omega}{\partial V}\right) \tag{1.6}$$

Because the Hamiltonian can be written $H(\mathbf{q}, \mathbf{p}) = T(\mathbf{p}) + V(\mathbf{q})$, the total energy density of states can be written in terms of the product of individual kinetic and potential energy densities of states $\Omega_T(T)$ and $\Omega_V(V)$, respectively. Because $T(\mathbf{p})$ has a simple functional form, we can further write

$$\begin{aligned}
\Omega_T(T) &= \int d\mathbf{p}\, \delta(T(\mathbf{p}) - T) \\
&= \int d\mathbf{p}\, \delta(\frac{1}{2}\mathbf{p}^{\mathrm{T}} \boldsymbol{M}^{-1} \mathbf{p} - T) \\
&= \left[\prod_{n=1}^{N} (2m_n)^{3/2}\right] \int d\mathbf{x}\, \delta(\mathbf{x}^T \mathbf{x} - T) \\
&= \left[\prod_{n=1}^{N} (2m_n)^{3/2}\right] S_N(T)
\end{aligned} \tag{1.7}$$

where we have made use of the transformation of variables $\mathbf{x} = (2\boldsymbol{M})^{-1/2}\mathbf{p}$ to simplify the integral of the delta function of all of Cartesian momenta space to a surface integral over a hypersphere. Here, m_n denotes the mass of particle n. $S_N(r)$ denotes the surface area of the N-sphere of radius r, and is given by

$$S_N(r) \quad = \quad \frac{2\pi^{N/2} r^{N-1}}{\Gamma(N/2)} \tag{1.8}$$

where $\Gamma(x)$ is the well-known Gamma function. Note that the kinetic energy is always non-negative, *i.e.* $T \geq 0$, Because the potential function $V(\mathbf{q})$ is often much more complicated, we can generally only write

$$\Omega_V(V) \quad = \quad \int d\mathbf{q}\, \delta(V(\mathbf{q}) - V) \tag{1.9}$$

The total energy density of states $\Omega(E)$ can therefore be computed from $\Omega_T(T)$ and $\Omega_V(V)$ by integrating over all combinations of T and V that yield $T + V = E$:

$$\Omega(E) \quad = \quad \int_0^E dT\, \Omega_T(T)\, \Omega_V(E - T) \tag{1.10}$$

If the system size is large, and hence contains a large number of solvent molecules, it can be shown that the solvated macromolecule will experience a distribution of kinetic and potential energies and volumes consistent with the temperature and pressure desired. The effective temperature and pressure the macromolecule experiences will be a complicated function of the total system energy E and volume V cannot generally be easily determined *a priori* and depends on the density of states of the system. If a big enough box can be afforded, it may be therefore appropriate to model the macromolecule in a solvated bath of constant volume and total system energy, both chosen to reproduce the desired average temperature and pressure by some equilibration process.

1.2.3 Canonical (NVT)

Because the systems typically treated in molecular mechanics simulations are finite in extent, the macromolecule may not experience a sufficiently large range of kinetic and potential energy distribution indicative of the canonical ensemble when the entire system is constrained to have constant total energy. For this reason, the NVT ensemble, where the distribution is given by

$$\begin{aligned} \rho(\mathbf{z}) &= [Z(E)]^{-1}\, e^{-\beta H(\mathbf{z})} \\ Z(\beta) &\equiv \int d\mathbf{z}\, e^{-\beta H(\mathbf{z})} \end{aligned} \tag{1.11}$$

where the inverse temperature $\beta = (k_B T)^{-1}$, where T is the absolute temperature.

The result is a system in which the temperature is fixed to some value T. The temperature is defined through the expectation of the kinetic energy

$$\frac{3}{2} N \beta^{-1} = \langle \frac{1}{2} \mathbf{p}^{\mathrm{T}} \boldsymbol{M} \mathbf{p} \rangle_{\beta} \tag{1.12}$$

Because the phase space weight, or *Boltzmann factor*, $e^{-\beta H(\mathbf{p},\mathbf{q})}$ can be decomposed into a product of individual factors factors $e^{-\beta T(\mathbf{p})}$ and $e^{-\beta V(\mathbf{p})}$ (something that is not true of non-Hamiltonian systems that employ constraints), we can deco,pose partition function $Z(\beta)$ can be decomposed into a product of the kinetic and potential energy partition functions:

$$\begin{aligned} Z(\beta) &= \int d\mathbf{z}\, e^{-\beta H(\mathbf{z})} \\ &= \int d\mathbf{p}\, d\mathbf{q}\, e^{-\beta(T(\mathbf{p})+U(\mathbf{q}))} \\ &= \left[\int d\mathbf{p}\, e^{-\beta T(\mathbf{p})}\right] \left[\int d\mathbf{q}\, e^{-\beta U(\mathbf{q}))}\right] \\ &= P(\beta)\, Q(\beta) \end{aligned} \tag{1.13}$$

where the kinetic partition function $P(\beta)$ is given by

$$\begin{aligned} P(\beta) &\equiv \int d\mathbf{p}\, e^{-\beta T(\mathbf{p})} \\ &= \end{aligned} \tag{1.14}$$

and the potential partition function $Q(\beta)$ by

$$Q(\beta) \equiv \int d\mathbf{q}\, e^{-\beta U(\mathbf{q})} \tag{1.15}$$

Both the kinetic and potential energies are therefore individually canonically distributed[7].

The volume, since it is fixed, should be chosen appropriately for the conditions of interest. Because the volume is fixed, the pressure fluctuations that result, especially if the system is small and the macromolecule experiences a large volume change as a result of some conformational change (such as folding), the pressure fluctuations that result may be artificially larger than the corresponding experimental system.

In the limit of an infinitely large system, the NVT ensemble has identical distribution function to the NVE ensemble.

1.2.4 Isothermal-isobaric (NPT)

In order

In the limit of an infinitely large system, the NPT ensemble has identical distribution function to the NVE ensemble.

1.2.5 Semi-grand canonical (μVT, μPT)

1.3 Evolution of phase space densities

Let $\mathbf{z}$ denote a point in the phase space of the system, consisting of Cartesian coordinates $\mathbf{q}$ and momenta $\mathbf{p}$. A mechanical property or *observable* A will in general be a function of the phase space of the system, and will be denoted $A(\mathbf{z})$. The expectation of A with respect to some probability distribution $\rho(\mathbf{z})$ is simply

$$\langle A \rangle_\rho = \int d\mathbf{z}\, \rho(\mathbf{z})\, A(\mathbf{z}). \tag{1.19}$$

The evolution of an ensemble of systems or, equivalently, our state of uncertainty about the microscopic state of a system, can be described by writing the density as a function of time as well as space, $\rho(\mathbf{z}, t)$. Given an initial distribution $\rho(\mathbf{z}, 0)$ that has been allowed to evolve for some time t, the expectation is given by

$$\langle A(t) \rangle_\rho = \int d\mathbf{z}\, \rho(\mathbf{z}, t)\, A(\mathbf{z}). \tag{1.20}$$

[7] Because of this decomposition, and because the form of the kinetic energy function $T(\mathbf{p})$ is known, we can write the probability distribution of kinetic energies $p(T)$ in the canonical ensemble as

$$p(T;\beta) = [P(\beta)]^{-1}\, \Omega_T(T)\, e^{-\beta T} \tag{1.16}$$

$$= \frac{\left[\prod_{n=1}^{N} (2m_n)^{3/2}\right]\, S_N(T)\, e^{-\beta T}}{\left[\prod_{n=1}^{N} (2m_n)^{3/2}\right] \int_0^\infty dT\, S_N(T)\, e^{-\beta T}} \tag{1.17}$$

$$= (2\pi)^{-3N/2}\, S_N(T)\, e^{-\beta T} \tag{1.18}$$

Comparison of observed kinetic energy distributions with this distribution function provides a good way to validate thermostat implementations in molecular dynamics simulation codes.

Instead of expressing the time dependence in the density, we can move the time-dependence to the observable A and write this expectation as a time-evolved observable $A(\mathbf{z}, t)$ integrated against the initial density $\rho(\mathbf{z}, 0)$:

$$\langle A(t) \rangle_\rho \quad = \quad \int d\mathbf{z}\, \rho(\mathbf{z}, 0)\, A(\mathbf{z}, t). \tag{1.21}$$

This is a similar story as in quantum mechanics: Eq. 1.20 represents the Schrödinger picture, where the distribution ρ evolves *forward* in time and is integrated against an observable that is time-independent, whereas Eq. 1.21 corresponds to the Heisenberg picture, where the phase function A evolves *backward* in time and is integrated against the initial distribution. For the purposes of this exposition, it is convenient to adopt the Heisenberg picture.

We presume that evolution of the phase function A is governed by the equation

$$\frac{dA}{dt} \quad = \quad i\mathcal{L}A \tag{1.22}$$

where $i\mathcal{L}$ is termed the *Liouvillean*, and is the instantaneous generator of dynamics. $\mathcal{L}$ has been defined this way so that we may also require it to be Hermitian, conferring upon it many of the same useful properties as familiar operators in quantum mechanics also enjoy. The time evolution of a phase space density $\rho(\boldsymbol{z}, t)$ can be shown by the chain rule to be governed by

$$\begin{aligned} \frac{d\rho}{dt} &= \frac{\partial \rho}{\partial \mathbf{z}} \cdot \frac{\partial \mathbf{z}}{\partial t} \\ &= \nabla \rho \cdot \dot{\mathbf{z}} \\ &= \nabla \rho \cdot iL\mathbf{z} \\ &= -i\mathcal{L}\rho \end{aligned} \tag{1.23}$$

where ∇ represents the gradient operator with respect to $\mathbf{z}$. The first step is simple application of the chain rule, while the last step is due to integration by parts[8]. We require that the Liouvillean $i\mathcal{L}$ preserves an equilibrium density $\rho_0(\mathbf{z})$, such that $\partial \rho_0 / \partial t = 0$.

With these requirements, the Liouvillean iL can represent a variety of dynamical schemes, including field-free Newtonian (Hamiltonian) mechanics, as well as thermostated (e.g. Nosé-Hoover or Andersen) or barostated (e.g. Andersen piston) dynamics. The equilibrium density $\rho_0(\mathbf{z})$ can therefore be microcanonical, canonical, or even isothermal-isobaric. Expectations over this equilibrium density are written

$$\langle A \rangle \quad = \quad \int d\mathbf{z}\, \rho_0(\mathbf{z})\, A(\mathbf{z}). \tag{1.25}$$

A *scalar product* (or *inner product*) can be defined with respect to the equilibrium density ρ_0 in the space of phase functions. The inner product of two phase functions A and B^* is given

[8] To see how the last step proceeds, note that integration of $\nabla \rho_0(\mathbf{z}) \cdot \mathbf{z}$ by parts yields

$$\int d\mathbf{z}\, \nabla \rho_0(\mathbf{z}) \cdot \mathbf{z} \quad = \quad [\rho_0(\mathbf{z}) \cdot \mathbf{z}]_S - \int d\mathbf{z}\, \rho_0(\mathbf{z}) \cdot 1 \tag{1.24}$$

where the surface term $[\rho_0(\mathbf{z}) \cdot \mathbf{z}]_S$ is zero since we require $\rho_0(\mathbf{z})$ to vanish at large $\mathbf{z}$ in order to be normalizable. Evans and Morriss [48] Section 3.3 contains an extended discussion on the relation of the Liouvillean acting on phase functions like A to acting on distribution functions like ρ.

by

$$\begin{aligned}(A, B^*) &\equiv \langle AB^* \rangle \\ &= \int d\mathbf{z}\, \rho_0(\mathbf{z})\, A(\mathbf{z})\, B^*(\mathbf{z}) \end{aligned} \tag{1.26}$$

where * denotes the complex conjugate.

We will find occasion to use the Dirac bracket notation when convenient. In this notation, the phase functions $A(\mathbf{z})$ and $B(\mathbf{z})$ are denoted by their corresponding *kets* $|A\rangle$ and $|B\rangle$, with the corresponding *bra* (e.g. $\langle A|$) corresponding to complex conjugation and the scalar product denoted

$$\begin{aligned}\langle A|B\rangle &\equiv \langle A^* B \rangle \\ &= \int d\mathbf{z}\, \rho_0(\mathbf{z})\, A(\mathbf{z})\, B^*(\mathbf{z}). \end{aligned} \tag{1.27}$$

Note that $\langle A|B\rangle = (A, B^*)$.

With the time evolution of A governed by Eq. 1.22, the explicit time dependence of A in the Heisenberg picture can be written as

$$\begin{aligned}A(\mathbf{z}, t) &= e^{i\mathcal{L}t} A(\mathbf{z}, 0) \\ &= \mathcal{P}_t A(\mathbf{z}, 0) \end{aligned} \tag{1.28}$$

where the operator $\mathcal{P}_t \equiv e^{i\mathcal{L}t}$ is termed the *propagator*, since it evolves[9] the phase function A by time t. Since L is Hermitian, the propagator e^{iLt} is unitary, and time evolution corresponds to a rotation in the space in which the phase functions live. The exponential is simply a formal shorthand for the infinite power series expansion

$$\mathcal{P}_t \equiv e^{i\mathcal{L}t} \equiv \sum_{n=0}^{\infty} \frac{(i\mathcal{L}t)^n}{n!}. \tag{1.29}$$

1.4 Dynamics in the canonical ensemble

For the majority of the remainder of this dissertation, we consider the canonical ensemble almost exclusively due to its widespread popularity in molecular simulations. Many of the results that follow can easily be adapted to other ensembles, but this may be nontrivial in some cases.

Because the canonical ensemble is purely an equilibrium construct, there is no definitive specification of how equilibrium dynamics is to be modeled. A number of choices are possible, each with advantages (theoretical, computational, or conceptual) that make it difficult to decide which is most appropriate. Here, we enumerate a number of these models, all of which produce the canonical *equilibrium* distribution (though perhaps in the limit of long sampling times), but treat the issue of dynamics in different manners.

[9] The sign difference between the time evolution operator for an observable (Eq. 1.23) and a density (Eq. 1.23) is sometimes a source of confusion. One way to think of this is that, in order for the time-evolved phase function to give the proper expectation when integrated against the initial density, the value of the phase function must be evolved *backward* in time at every point in phase space for the two pictures to coincide.

1.4.1 Canonical average over constant-energy trajectories

In the simplest model of dynamics in the canonical ensemble, the system evolves according to Hamilton's equations of motion (that is, motion is purely Newtonian). Because the system is confined to a single energy shell E, we must therefore consider a canonical distribution of energy shells, weighting the dynamical trajectories from each energy shell appropriately:

$$p(E) \quad = \quad [Z(\beta)]^{-1} \, \Omega(E) \, e^{-\beta E} \tag{1.30}$$

where $\Omega(E)$ is the total energy density of states defined above.

1.4.2 Stochastic thermostats

In this section, we briefly consider some of the more popular stochastic thermostats which have been employed.

Smoluchowski dynamics

Langevin dynamics

The Andersen thermostat

Andersen demonstrated that a canonical ensemble can be recovered by having molecules of the system undergo stochastic collisions with fictitious bath particles, where upon collision, the momenta of the atoms in the molecule undergoing collision would be reassigned from the Maxwell distribution [2]. A constant collision rate ν per unit time can be employed, or collisions could occur at predetermined intervals. Alternatively, each atom may undergo collisions independently[10], or even the entire system could undergo "massive collision" at random (or regular) intervals, though such massive collisions are obviously less suitable for modeling realistic kinetics. Andersen demonstrated that this procedure preserves the canonical distribution, as it consists of two components:

1.4.3 Deterministic thermostats

The Berendsen weak-coupling algorithm

SImulations sometimes employ the Berendsen weak-coupling algorithm for thermal control [11], in which a the *instantaneous kinetic temperature*, $K(\mathbf{z}) \equiv \mathbf{p}^{\mathrm{T}} \boldsymbol{M} \mathbf{p} / 3Nk_B$, derived by analogy with Eq. 1.12 above, is restrained to fluctuate the desired temperature. However, it is well known that while this method may produce correct average temperatures computed by Eq. 1.12 above, the distribution of kinetic energies differs from that in the canonical ensemble, which leads to an incorrect phase space distribution [103]. In very large solvated systems, however, the effect of the thermostat may be minimal, and a near-correct ensemble may be obtained.

The Nosé-Hoover thermostat

[10] It should be note that the Andersen thermostat with each atom independently undergoing collision only produces correct temperatures and distributions when no constraints are employed. It has been observed anecdotally that, due to the removal of velocities in the direction of constraints, the kinetic energy distribution is disrupted.

2 Macromolecular dynamics as a discrete-state Markov process

3 Validating master equation models of macromolecular dynamics

4 An automatic state decomposition algorithm

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 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.

4.1 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 [38]. Enzymatic catalysis may involve transitions between multiple conformational substates, only some of which may allow substrate access or catalysis [14, 44, 169]. 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 [19, 54, 98]. 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 ex-

plicit 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) [104], *milestoning* [49], and methods based on commitment probability distributions [12, 120] 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 [158], can be constructed by identifying local potential energy minima as states and estimating interstate transition rates by transition state theory [9, 30, 47, 86, 90, 105, 106]. 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 [131, 132, 155], 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 [4, 32, 46, 59, 117, 126, 134, 135, 137, 139, 146]. 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 [5, 8, 10, 90, 91]. 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. [135]) or a continuous-time master equation model with coarse-grained time (e.g. Ref. [139]). 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 [145], relaxation from experimentally inaccessible prepared states [23], mean first-passage times [135], the existence of hidden intermediates [114], and P_{fold} values or transmission coefficients [88], 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 [23, 46, 137, 139, 146]. In the absence of this knowledge, others have turned to conformational clustering techniques to identify conformationally distinct regions which may also be kinetically distinct [4, 32, 77, 135].

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* [128]. 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 [129]. 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 4.2, 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 4.3, we provide a detailed description of the automatic state decomposition algorithm and its implementation. In Section 4.4, 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

4.5, we discuss the advantages and shortcomings of our algorithm, with the hope that future state decomposition algorithms can address the remaining challenges.

4.2 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 4.2.1, we review Markov chain and master equation models of conformational dynamics. Next, in Section 4.2.2, we describe their construction from equilibrium molecular dynamics trajectories given any state partitioning. Section 4.2.3 enumerates a number of requirements for a useful state partitioning. Finally, Section 4.2.4 discusses possible methods for validating a given state decomposition. The actual implementation of the algorithm used here is described in detail in Section 4.3.

4.2.1 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*[1] $\boldsymbol{T}(\tau)$ 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 [158]) we additionally have the requirement

$$T_{ji} p_{\mathrm{eq},i} = T_{ij} p_{\mathrm{eq},j} \tag{4.1}$$

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

$$\boldsymbol{p}(n\tau) \;=\; \boldsymbol{T}(n\tau)\boldsymbol{p}(0) \;=\; [\boldsymbol{T}(\tau)]^n \boldsymbol{p}(0). \tag{4.2}$$

[1] 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.

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

$$\dot{\boldsymbol{p}}(t) \quad = \quad \boldsymbol{K}\boldsymbol{p}(t) \tag{4.3}$$

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

$$\boldsymbol{p}(t) \quad = \quad e^{\boldsymbol{K}t}\boldsymbol{p}(0), \tag{4.4}$$

where the exponential denotes the formal matrix exponential. Eq. 4.3 is often referred to as a *master equation* [112, 158] 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. 4.4 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. 4.2 and 4.4:

$$\boldsymbol{T}(\tau) \quad = \quad e^{\boldsymbol{K}\tau}. \tag{4.5}$$

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 [59, 139].

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

$$\mu_k(\tau) \quad = \quad e^{\lambda_k \tau}. \tag{4.6}$$

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

$$\tau_k = -\lambda_k^{-1} = -\tau[\ln \mu_k(\tau)]^{-1} \tag{4.7}$$

and the associated eigenvector gives information about the aggregate conformational transitions that are associated with this timescale [67, 127, 128, 130]. 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. 4.2), except for use of the timescales implied by the transition matrix, as described above.

4.2.2 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.* [145] show that the transition probability $T_{ji}(\tau)$ can be written as the following ratio of canonical ensemble averages:

$$T_{ji}(\tau) = \frac{\int d\boldsymbol{z}(0)\, e^{-\beta H(\boldsymbol{z}(0))}\, \chi_j(\boldsymbol{z}(\tau))\, \chi_i(\boldsymbol{z}(0))}{\int d\boldsymbol{z}(0)\, e^{-\beta H(\boldsymbol{z}(0))}\, \chi_i(\boldsymbol{z}(0))} \tag{4.8}$$

$$= \frac{\langle \chi_j(\tau) \chi_i(0) \rangle}{\langle \chi_i \rangle} \tag{4.9}$$

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. 4.9 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 [145, 146] or *partition of unity* restraints [162] can also be used to compute transition matrix elements efficiently.

4.2.3 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[2] 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 [17], 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, 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[3].

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 [17].

[2]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.

[3]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.

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.

4.2.4 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 [101]. The most rigorous test for Markovian 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 [22].

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. 4.2) 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. [23, 145].

Another approach, from Park, *et al.*, [117] 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.*, [145] 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. 4.7, as a function of increasing lag time τ [22]. 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 obser-

vation time intervals (Eq. 4.2). 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 [117].

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

4.3 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.

4.3.1 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 [52, 58], distributed computing platforms such as Folding@Home [116, 133], 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.

4.3.2 Sketch of the method.

A state decomposition algorithm intended to produce the most *useful* models, as discussed in Section 4.2.3 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. [68], 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:

$$Q \equiv \sum_{i=1}^{L} T_{ii}(\tau) \tag{4.10}$$

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 [68].

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 [152], 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. [128], 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 4.1.

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 4.1. 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 4.1). The refined set of all microstates is then lumped to form refined macrostates (the second *lump* step in Figure 4.1). 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 [129]. 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.

4.3.3 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 [64] 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 [97] 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 [148]. 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 [142], 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 [36, 67, 127, 129] 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. 4.9) 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.* [36], an initial guess for the optimal lumping of microstates to macrostates is generated using the *left* eigenvectors[4] 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 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. [36], 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. There-

[4]The left eigenvector $\boldsymbol{v}_k$ is simply related to the right eigenvector $\boldsymbol{u}_k$ by $(\boldsymbol{v}_k)_i = p_{\mathrm{eq},i}^{-1} (\boldsymbol{u}_k)_i$ [112].

fore, 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.

4.3.4 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. 4.7. 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 4.2.4). To estimate the statistical uncertainty of these implied timescales, we perform a bootstrapping procedure [43] 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.

4.4 Applications

4.4.1 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 4.2, left) are known *a priori*[5], it is 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 [6, 15, 21, 23, 69, 106]. Previously, a master equation model constructed from a set of six manually identified states (Figure 4.2, 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 [23]. 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. [23], and consisted of an equilibrium pool of 1000 constant-energy[6], 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[7]. The peptide was modeled by the AMBER parm96 forcefield [81], and solvated in TIP3P water [74]. The previous study [23] 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* [145] 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 [24, 85] by discretizing each degree of freedom into 10° bins (Figure 4.2). Because

[5] Simulations of alanine dipeptide examining the committor distribution have implicated solvent coordinates as the next-slowest degree of freedom [15, 95], 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 [23].

[6] 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.

[7] Note that only 10 ns/replica were used in Ref. [23] — 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.

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 [23], identified as dark lines in Figure 4.2, 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.

First, the automatic state decomposition method described in Section 4.3 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 4.3.3. A single sampling interval — 0.1 ps — was used for the calculation of the metastability metric Q used in lumping, as described in Section 4.3.2. 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 4.3, in comparison with the "gold standard" manual state decomposition from Ref. [23] and a "poor" manual decomposition that is expected to fail to reproduce kinetics well because its states include internal kinetic barriers[8]. 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 4.3, 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 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 4.3, right) . In the other automatic decomposition with nearly equal metastability, states 3 and 4 of the manual decomposition (numbering given in Figure 4.2) have been merged into a single state, and state 5 of the manual decomposition

[8]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.

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 4.4) 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 4.4), as verified by examination of the converged timescales (data not shown). This suggests the method yields partitionings that are relatively stable and optimal.

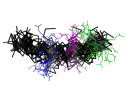
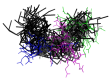
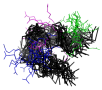
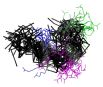
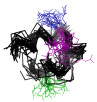
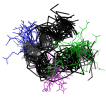
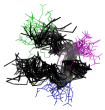
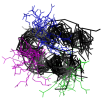
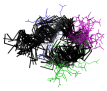
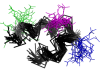
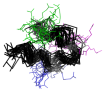
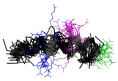
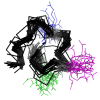
From the "poor" manual decomposition, however, a number of conformations in manual states 5 and 6 are incorrectly grouped with state 2, though this did not significantly affect the timescales. Further investigation showed that the algorithm never split these conformations from state 2, partly due to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

4.4.2 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 [87, 93, 94, 149, 164] and computationally [57, 136, 137, 170]. Since helix formation occurs on the nanosecond timescale, Sorin *et al.* were able to reach 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 [137]. 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.* [137]; 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 [137, 160] and solvated in TIP3P water [74]. Though the Berendsen weak-coupling scheme [11] was employed for thermal and pressure control[9], 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.

[9] 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.

Table 4.1: 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

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 4.1), 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 4.1), 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 [137]. 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 [137]. 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 [40] identified similar configurations to those noted by the automatic state 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) [170].

We then examined the implied timescales as a function of lag time (Figure 4.5). 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 4.2.1, 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 [71, 144] $\tau_{\mathrm{ac},i}$ 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[10], 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 4.1. 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 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 4.6. 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

[10] There is some indication that consideration of restrictions of Markov chains to these macrostates may facilitate the computation of the internal equilibration time [101].

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.

4.4.3 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 [167], we considered the engineered 12-residue β-peptide trpzip2 [26]. A set of 323 10 ns constant-energy, constant-volume simulations of the unblocked peptide[11] simulated using the AMBER parm96 forcefield [81] in TIP3P water [74] was obtained from Pitera et al. [119]; 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 4.3.3.

Figure 4.7 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.* [119]. (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. [119]) we found less than five examples of the +2 and -2 hydrogen bonding states identified in Ref. [119], 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 4.7, A, 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 4.8 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. 4.2.1) 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. 4.4.2)

[11] Note that the peptide studied experimentally in Refs. [26] and [167] was synthesized with an amidated C-terminus, whereas the termini of the simulated peptide in the dataset considered here were left zwitterionic.

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.

4.5 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 [137]. 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 [119]. 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 [117]. 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 [**?**].

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 [23] 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 [145]. Selection of states, or regions of configuration space, from which further simulations should be initiated could be chosen based on uncertainty considerations [134].

4.6 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.

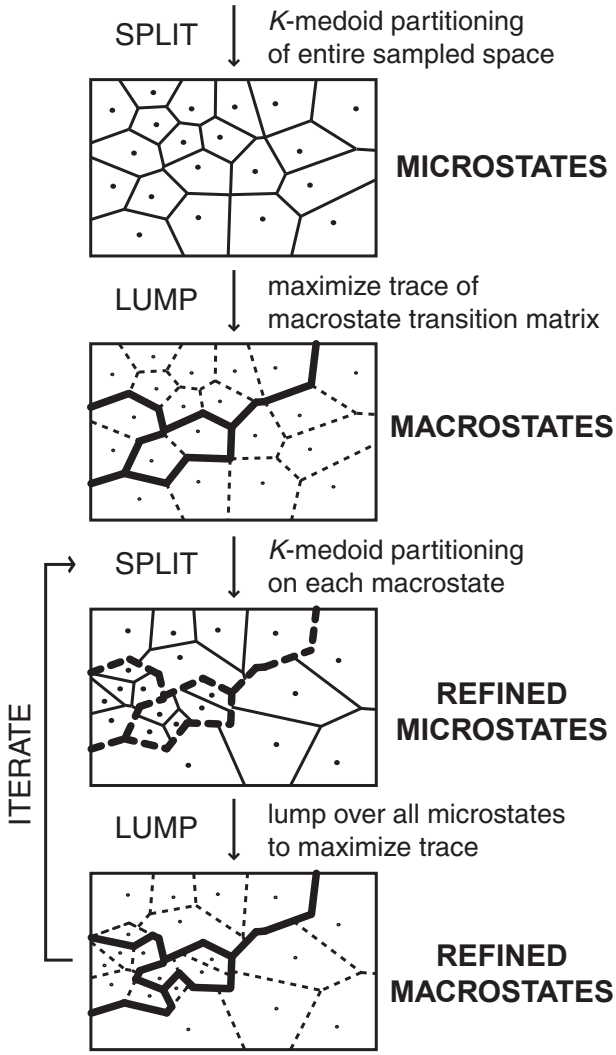


Figure 4.1: Flowchart of the automatic state decomposition algorithm.

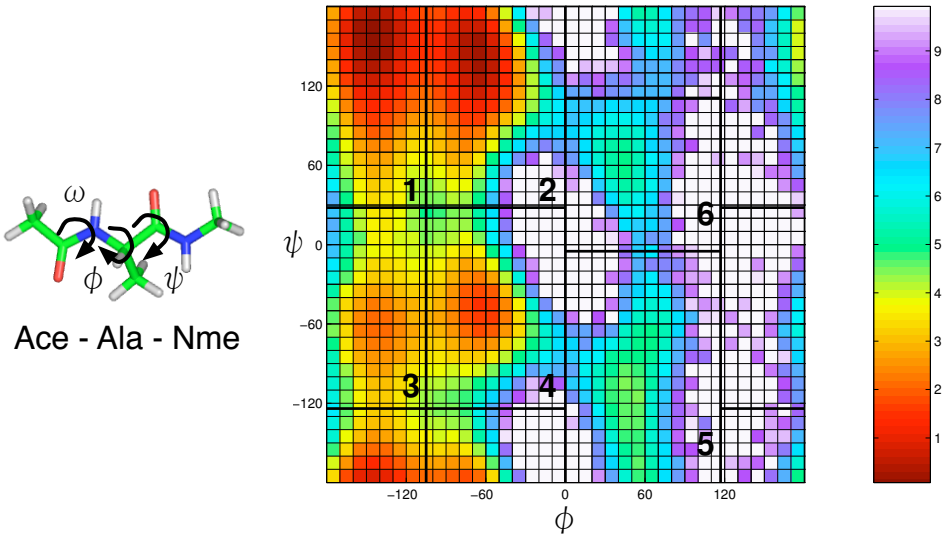


Figure 4.2: Potential of mean force and manual state decomposition for alanine dipeptide. Left: The terminally-blocked alanine peptide with ϕ, ψ, and ω backbone torsions labeled. Right: The potential of mean force in the (ϕ, ψ) torsions at 400 K estimated from the parallel tempering simulation, truncated 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. [23] from examining the 300 K PMF are superimposed, and the states labeled.

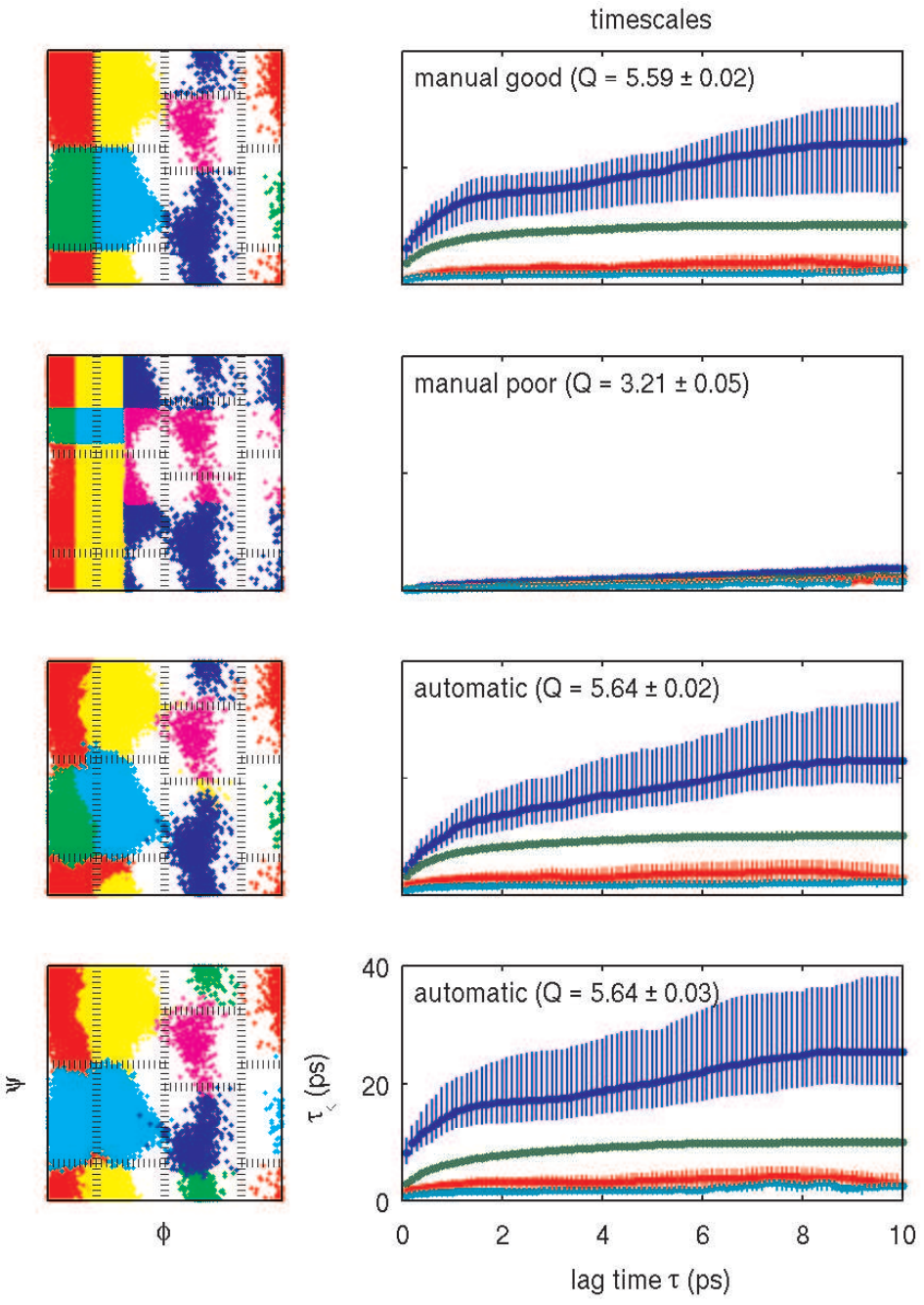


Figure 4.3: Comparison of manual and automatic state decompositions for alanine dipeptide. The left panels depict state partitionings, and the right panels the associated timescales (in picoseconds) as a function of lag time with uncertainties shown, as estimated from the procedure described in Section 4.3.4. Top two panels: Manual "good" or "gold standard" state decomposition from Ref. [23] 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.

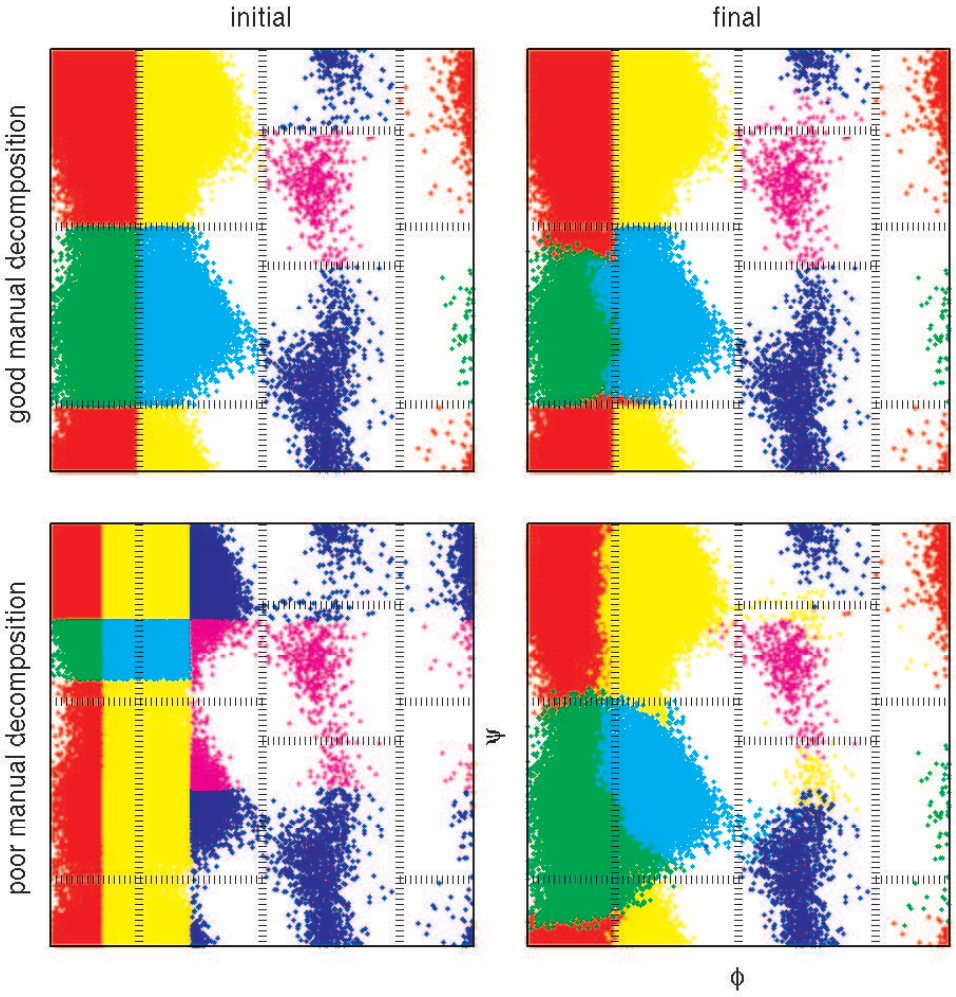


Figure 4.4: Stability and recovery of optimal state decomposition for alanine dipeptide. Top: Ten cycles of automatic state decomposition applied to a "good" manual partitioning (left) to yield an automatic partitioning (right). Bottom: Ten cycles of automatic state decomposition applied to a "poor" manual partitioning (left) to yield an automatic partitioning (right).

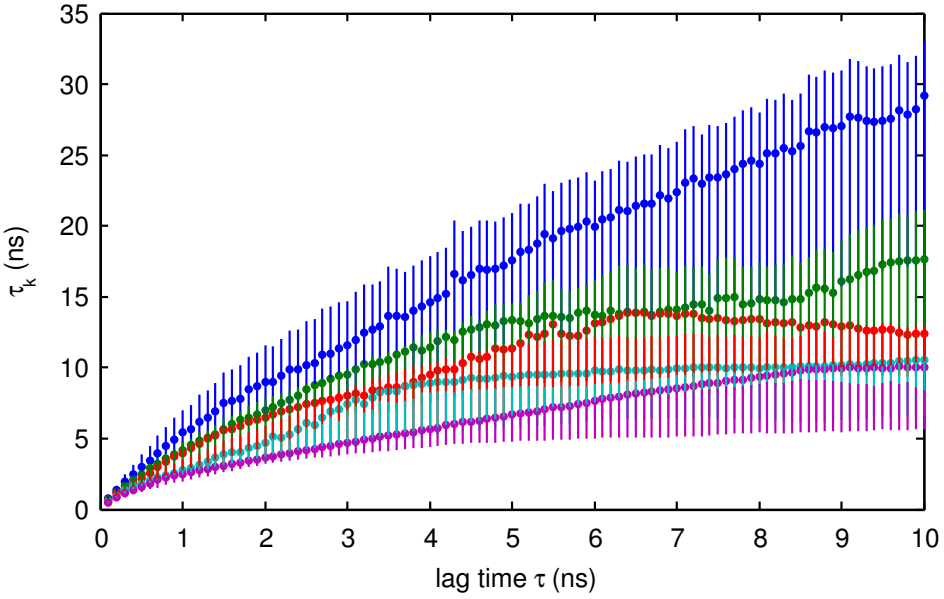


Figure 4.5: Implied timescales of the F_s peptide as a function of lag time for 20-state automatic state decomposition. The five longest timescales are shown. Circles represent the maximum likelihood estimate, and vertical bars depict 68% symmetric confidence intervals about the mean. Note the timescales 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.

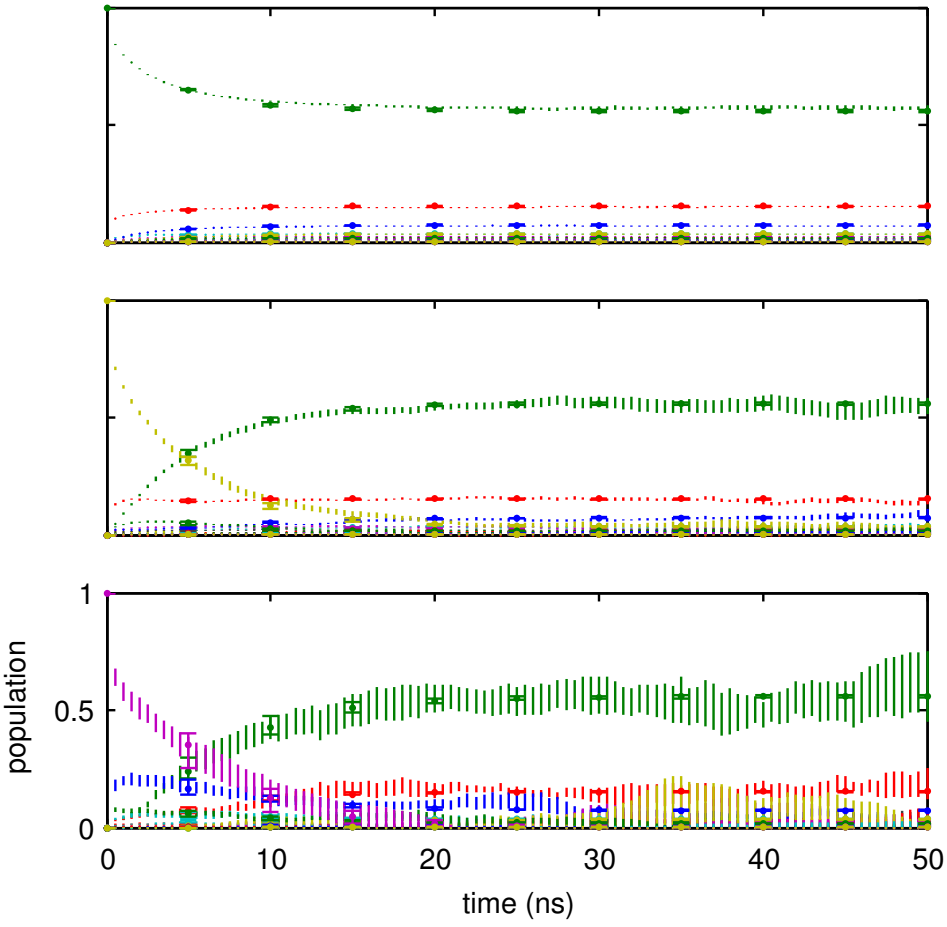


Figure 4.6: Reproduction of observed state population evolution by Markov model for the F_s peptide. The time evolution of the Markov model constructed from the 5 ns lag time transition matrix is shown by the filled circles with flat error bars, which denote the 68% confidence interval from realizations 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).

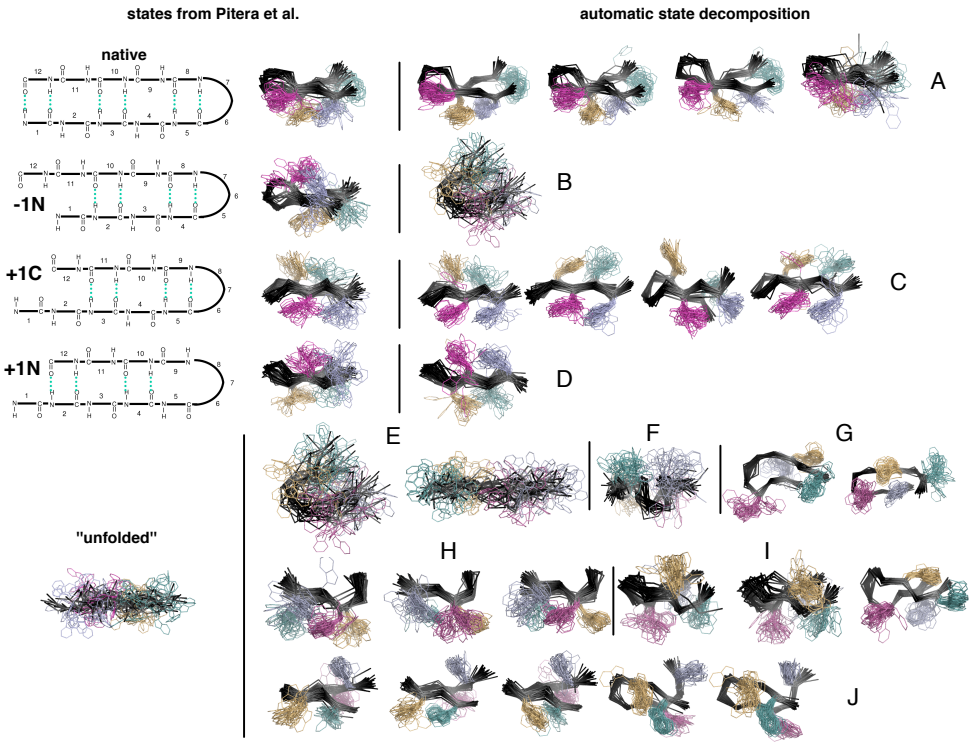


Figure 4.7: Comparison of some trpzip2 macrostates found by automatic state decomposition with misregistered hydrogen bonding states identified in a previous study. Left: The five hydrogen bonding patterns enumerated in Pitera *et al.* [119] that occurred in sufficient numbers in the subsampled trpzip2 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.

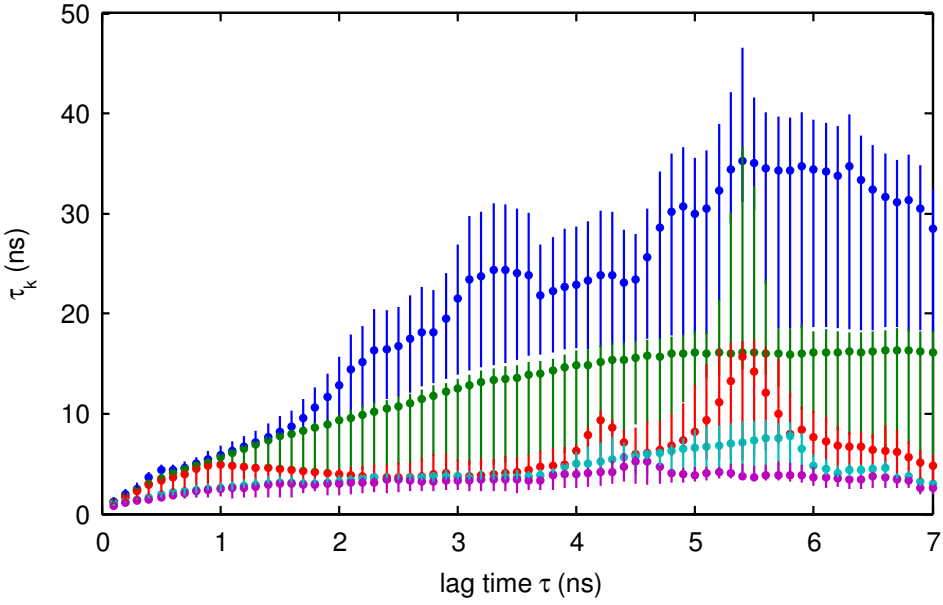


Figure 4.8: Implied timescales of trpzip2 as a function of lag time for 40-state automatic state decomposition. The five longest timescales are show. Vertical bars depict 68% confidence intervals.

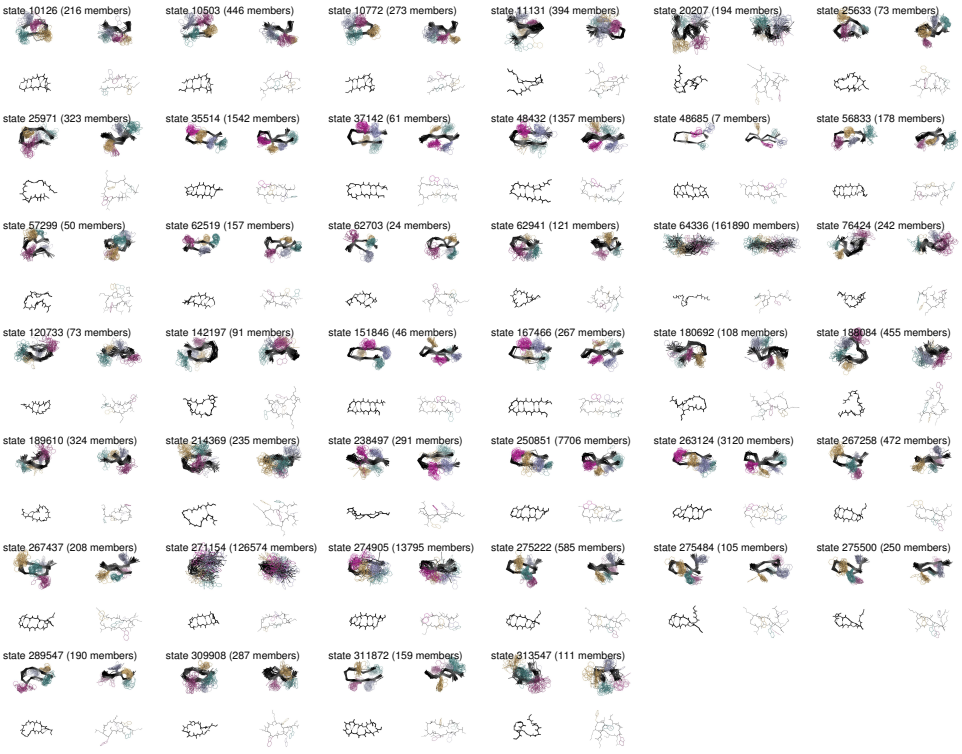


Figure 4.9: Automatic state decomposition applied to trpzip2 to produce 40 macrostates.

5 Efficient methods for computing interstate transition rates

6 Multistage and iterative methods for the efficient and automatic construction of master equation models

Conclusion

Conclusion goes here.

A The weighted histogram analysis method

B A primer on statistical uncertainties

Bibliography

[1] ALLEN, M. P., AND TILDESLEY, D. J. *Computer simulation of liquids*. Clarendon Press, Oxford, 1991.

[2] ANDERSEN, H. C. Molecular dynamics simulations at constant pressure and/or temperature. J. Chem. Phys. **72**(4):2384–2293, *1980*.

[3] ANDERSEN, H. C. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. J. Comput. Phys. **52**:24–34, *1983*.

[4] ANDREC, M., FELTS, A. K., GALLICCHIO, E., AND LEVY, R. M. Protein folding pathways from replica exchange simulations and a kinetic network model. Proc. Natl. Acad. Sci. USA **102**:6801–6806, *2005*.

[5] ANSARI, A., BERENDZEN, J., BOWNE, S. F., FRAUENFELDER, H., IBEN, I. E. T., SAUKE, T. B., SHYAMSUNDER, E., AND YOUNG, R. D. Protein states and proteinquakes. Proc. Nat. Acad. Sci. USA **82**:5000–5004, *1985*.

[6] APOSTOLAKIS, J., FERRARA, P., AND CAFLISCH, A. Calculation of conformational transitions and barriers in solvated systems: Application to the alanine dipeptide in water. J. Chem. Phys. **110**(4):2099–2108, *1999*.

[7] BAHAR, I., ATILGAN, A. R., AND ERMAN, B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold. Des. **2**:173–181, *1997*.

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

[9] BALL, K. D., AND BERRY, R. S. Realistic master equation modeling of relaxation on complete potential energy surfaces: Kinetic results. J. Chem. Phys. **109**(19):8557–8572, *1998*.

[10] BECKER, O. M., AND KARPLUS, M. The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics. J. Chem. Phys. **106**(4):1495–1517, *1997*.

[11] BERENDSEN, H. J. C., POSTMA, J. P. M., VAN GUNSTEREN, W. F., DINOLA, A., AND HAAK, J. R. Molecular dynamics with coupling to an external bath. J. Chem. Phys. **81**(8):3684–3690, *1984*.

[12] BEREZHKOVSKII, A., AND SZABO, A. One-dimensional reaction coordinates for diffusive activated rate processes in many dimensions. J. Chem. Phys. **122**:014503, *2005*.

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

[14] BOEHR, D. D., MCELHENY, D., DYSON, H. J., AND WRIGHT, P. E. The dynamic energy landscape of dihydrofolate reductase catalysis. Science **313**:1638–1642, *2006*.

[15] BOLHUIS, P. G., DELLAGO, C., AND CHANDLER, D. Reaction coordinates of biomolecular isomerization. Proc. Natl. Acad. Sci. **97**(11):5877–5882, *2000*.

[16] BORN, M., AND OPPENHEIMER, R. Zur quantentheorie der molekeln. Annalen der Physik **389**:457–484, *1927*.

[17] CHANDLER, D. Statistical mechanics of isomerization dynamics in liquids and the transition

state approximation. J. Chem. Phys. **68**(6):2959–2970, *1978*.

[18] CHANDLER, D., AND BERNE, B. J. Comment on the role of constraints on the conformational structure of n-butane in liquid solvents. J. Chem. Phys. **71**(12):5386–5387, *1979*.

[19] CHANGEUX, J., AND EDELSTEIN, S. J. Allosteric mechanisms of signal transduction. Science **308**:1424–1428, *2005*.

[20] CHEBOTAREVA, N. A., KURGANOV, B. I., AND LIVANOVA, N. B. Biochemical effects of molecular crowding. Biochemistry Moscow **69**(11):1239–1253, *2004*.

[21] CHEKMAREV, D. S., ISHIDA, T., AND LEVY, R. M. Long-time conformational transitions of alanine dipeptide in aqueous solution: Continuous and discrete-state kinetic models. J. Phys. Chem. B **108**:19487–19495, *2004*.

[22] CHODERA, J. D., SWOPE, W. C., PITERA, J. W., AND DILL, K. A. Describing protein folding kinetics by molecular dynamics simulations. 3. Validation of state space decomposition, with application to terminally-blocked alanine in explicit solvent. In preparation, *2006*.

[23] CHODERA, J. D., SWOPE, W. C., PITERA, J. W., AND DILL, K. A. Long-time protein folding dynamics from short-time molecular dynamics simulations. Multiscale Model. Simul. **Special Issue on Biomolecular Simulation, to appear**, *2006*.

[24] CHODERA, J. D., SWOPE, W. C., PITERA, J. W., SEOK, C., AND DILL, K. A. Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. Submitted to J. Chem. Theor. Comput., *2006*.

[25] CHURCH, B. W., AND SHALLOWAY, D. Top-down free-energy minimization on protein potential energy landscapes. Proc. Natl. Acad. Sci. USA **98**(11):6098–6013, *2001*.

[26] COCHRAN, A. G., SKELTON, N. J., AND STAROVASNIK, M. A. Tryptophan zippers: Stable, monomeric β-hairpins. Proc. Natl. Acad. Sci. **98**(10):5578–5583, *2001*.

[27] CORDERO-MORALES, J. F., CUELLO, L. G., AND PEROZO, E. Voltage-dependent gating at the KcsA selectivity filter. Nat. Struct. Mol. Biol. **13**(4):319–322, *2006*.

[28] CORDERO-MORALES, J. F., CUELLO, L. G., ZHAO, Y., JOGINI, V., CORTES, D. M., ROUX, B., AND PEROZO, E. Molecular determinants of gating at the potassium-channel selectivity filter. Nat. Struct. Mol. Biol. **13**(4):311–318, *2006*.

[29] CORNELL, W. D., CIEPLAK, P., BAYLY, C. I., GOULD, I. R., MERZ, K. M., FERGUSON, D. M., SPELLMEYER, D. C., FOX, T., CALDWELL, J. W., AND KOLLMAN, P. A. J. Am. Chem. Soc. **117**:5179, *1995*.

[30] CZERMINSKI, R., AND ELBER, R. Reaction path study of conformational transitions in flexible systems: Application to peptides. J. Chem. Phys. **92**(9):5580–5601, *1990*.

[31] DABLY-BROWN, W., HANSEN, H. H., KORSGAARD, M. P. G., MIRZA, N., AND OLESEN, S. K_v7 channels: Function, pharmacology and channel modulators. Curr. Topics Med. Chem. **6**:999–1023, *2006*.

[32] DE GROOT, B. L., DAURA, X., MARK, A. E., AND GRUBMÜLLER, H. Essential dynamics of reversible peptide folding: Memory-free conformational dynamics governed by internal hydrogen bonds. J. Mol. Biol. **309**:299–313, *2001*.

[33] DEMO, S. D., AND YELLEN, G. The inactivation gate of the *Shaker* k^+ channel behaves like an open-channel blocker. Neuron **7**:743–753, *1991*.

[34] DESPA, F., AND BERRY, R. S. Inter-basin dynamics on multidimensional potential surfaces. i. escape rates on complex basin surfaces. J. Chem. Phys. **115**(18):8274–8278, *2001*.

[35] DESPA, F., FERNÁNDEZ, A., AND BERRY, R. S. Interbasin motion approach to dynamics of conformationally constrained peptides. J. Chem. Phys. **118**(12):5673–5682, *2003*.

[36] DEUFLHARD, P., HUISINGA, W., FISCHER, A., AND SCHÜTTE, C. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. Linear Algebra Appl. **315**(1-3):39–59, *August 2000*.

[37] DITTRICH, M., AND SCHULTEN, K. PcrA helicase, a prototype ATP-driven molecular motor. Structure **14**:1345–1353, *2006*.

[38] DOBSON, C. M. Protein folding and misfolding. Nature **426**:884–890, *2003*.

[39] DU, R., PANDE, V. S., GROSBERG, A. Y., TANAKA, T., AND SHAKHNOVICH, E. S. On the transition coordinate for protein folding. J. Chem. Phys. **108**(1):334–350, *1998*.

[40] DUAN, Y., WU, C., CHOWDHURY, S., LEE, M. C., XIONG, G., ZHANG, W., YANG, R., CIEPLAK, P., LUO, R., LEE, T., CALDWELL, J., WANG, J., AND KOLLMAN, P. 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*.

[41] DYSON, H. J., AND WRIGHT, P. E. Intrinsically unstructured proteins and their functions. Nat. Rev. Mol. Cell Biol. **6**(3):197–208, *2005*.

[42] ECKERT, B., MARTIN, A., BALBACK, J., AND SCHMID, F. X. Prolyl isomerization as a molecular timer in phage infection. Nat. Struct. Mol. Biol. **12**(7):619–623, *2005*.

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

[44] EISENMESSER, E. Z., BOSCO, D. A., AKKE, M., AND KERN, D. Enzyme dynamics during catalysis. Science **295**:1520–1523, *2002*.

[45] ELCOCK, A. H. Atomic-level observation of macromolecular crowding effects: Escape of a protein from the GroEL cage. Proc. Nat. Acad. Sci. USA **100**:2340–2344, *2003*.

[46] ELMER, S. P., PARK, S., AND PANDE, V. S. Foldamer dynamics expressed via Markov state models. II. State space decomposition. J. Chem. Phys. **123**:114903, *2005*.

[47] EVANS, D. A., AND WALES, D. J. Folding of the GB1 hairpin peptide from discrete path sampling. J. Chem. Phys. **112**(2):1080, *2004*.

[48] EVANS, D. J., AND MORRISS, G. P. *Statistical mechanics of nonequilibrium liquids*. Academic Press, London; San Diego, CA, 1990.

[49] FARADJIAN, A. K., AND ELBER, R. Computing time scales from reaction coordinates by milestoning. J. Chem. Phys. **120**(23):10880–10889, *2004*.

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

[51] FISCHER, A. *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.

[52] FITCH, B. G., GERMAIN, R. S., MENDELL, M., PITERA, J., PITMAN, M., RAYSHUBSKIY, A., SHAM, Y., SUITS, F., SWOPE, W., WARD, T. J. C., ZHESTKOV, Y., AND ZHOU, R. Blue Matter, an application framework for molecular simulation on Blue Gene. J. Parallel Distrib. Comput. **63**:759–773, *2003*.

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

[54] FRAUENFELDER, H., MCMAHON, B. H., AUSTIN, R. H., CHU, K., AND GROVES, J. T. 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*.

[55] GALLIAT, T. *Adaptive Multilevel Cluster Analysis by Self-Organizing Box Maps*. PhD thesis, FU Berlin, 2002.

[56] GALLIAT, T., HUISINGA, W., AND DEUFLHARD, P. Self-organizing maps combined with eigenmode analysis for automated cluster identification. In *Proceeding of the ICSC Symposia on Neural Computation* (Berlin, Germany, May 2000).

[57] GARCÍA, A. E., AND SANBONMATSU, K. Y. α-helical stabilization by side chain shielding of backbone hydrogen bonds. Proc. Nat. Acad. Sci. USA **99**:2782–2787, *2002*.

[58] GERMAIN, R. S., FITCH, B., RAYSHUBSKIY, A., ELEFTHERIOU, M., PITMAN, M. C., SUITS, F., GIAMPAPA, M., AND WARD, T. C. 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* (New York, NY, USA, 2005), ACM Press, pp. 207–212.

[59] GRUBMÜLLER, H., AND TAVAN, P. Molecular dynamics of conformational substates for a simplfied protein model. J. Chem. Phys. **101**(6):5047–5057, *September 1994*.

[60] GRUEBELE, M., SABELKO, J., BALLEW, R., AND ERVIN, J. Laser temperature jump induced protein refolding. Acc. Chem. Res. **31**:699–707, *1998*.

[61] HA, T., ZHUANG, X., KIM, H. D., ORR, J. W., WILLIAMSON, J. R., AND CHU, S. Ligand-induced conformational changes observed in single rna molecules. Proc. Natl. Acad. Sci. USA **96**:9077–9082, *1999*.

[62] HAMMES-SCHIFFER, S. Hydrogen tunneling and protein motion in enzyme reactions. Acc. Chem. Res. **39**:93–100, *2006*.

[63] HARTMANN, C., AND SCHÜTTE, C. Free energy calculations in many dimensions. Submitted for publication, *2004*.

[64] HASTIE, T., TIBSHIRANI, R., AND FRIEDMAN, J. *The Elements of Statistical Learning*. Springer, 2001.

[65] HIROSE, K., AKIMARU, E., AKIBA, T., ENDOW, S. A., AND AMOS, L. A. Large conformational changes in a kinesin motor catalyzed by interaction with microtubules. Mol. Cell **23**:913–923, *2006*.

[66] HORN, H. W., SWOPE, W. C., PITERA, J. W., MADURA, J. D., DICK, T. J., HURA, G. L., AND HEAD-GORDON, T. Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew. J. Chem. Phys. **120**(20):9665–9678, *2004*.

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

[68] HUISINGA, W., AND SCHMIDT, B. *Advances in Algorithms for Macromolecular Simulation*. Lecture Notes in Computational Science and Engineering. Springer, 2005, ch. Metastability and dominant eigenvalues of transfer operators.

[69] HUMMER, G., AND KEVREKIDIS, I. G. Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. J. Chem. Phys. **118**(23):10762–10773, *June 2003*.

[70] IGUMENOVA, T. I., FREDERICK, K. K., AND WAND, A. J. Characterization of the fast dynamics of protein amino acid side chains using NMR relaxation in solution. Chem. Rev. **106**:1672–1699, *2006*.

[71] JANKE, W. Statistical analysis of simulations: Data correlations and error estimation. In *Quantum Simulations of Complex Many-Body Systems: From Theory to Algorithms*, J. Grotendorst, D. Marx, and A. Murmatsu, Eds., vol. 10. John von Neumann Institute for Computing, 2002, pp. 423–445.

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

[73] JASWAL, S. S., TRUHLAR, S. M. E., DILL, K. A., AND AGARD, D. A. Comprehensive analysis of protein folding activation thermodynamics reveals a universal behavior violated by kinetically stable proteases. J. Mol. Biol. **347**:355–366, *2005*.

[74] JORGENSEN, W. L., CHANDRASEKHAR, J., MADURA, J. D., IMPEY, R. W., AND KLEIN, M. L. Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. **79**:926, *1983*.

[75] JR., A. D. M., BASHFORD, D., BELLOTT, M., JR., R. L. D., EVANSECK, J. D., FIELD, M. J., FISCHER, S., GAO, J., GUO, H., HA, S., JOSEPH-MCCARTHY, D., KUCHNIR, L., KUCZERA, K., LAU, F. T. K., MATTOS, C., MICHNIK, S., NGO, T., NGUYEN, D. T., PRODHOM, B., III, W. E. R., ROUX, B., SCHLENKRICH, M., SMITH, J. C., STOTE, R., STRAUB, J., WATANABE, M., WIÓRKIEWICZ-KUCZERA, J., YIN, D., AND KARPLUS, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J. Phys. Chem. B **102**:3587–3616, *1998*.

[76] KAMBARA, T., AND IKEBE, M. A unique ATP hydrolysis mechanism of single-headed processive myosin, myosin IX. J. Biol. Chem. **281**(8):4949–4957, *2006*.

[77] KARPEN, M. E., TOBIAS, D. J., AND BROOKS III, C. L. Statistical clustering techniques for the analysis of long molecular dynamics trajectories: Analysis of 2.2-ns trajectories of YPGDV. Biochemistry **32**:412–420, *1993*.

[78] KE, A., ZHOU, K., DING, F., CATE, J. H. D., AND DOUDNA, J. A. A conformational switch controls hepatitis delta virus ribozyme catalysis. Nature **429**:201–205, *2004*.

[79] KHALIL, M., DEMIRDÖVEN, N., AND TOKMAKOFF, A. Coherent 2D IR spectroscopy: Molecular structure and dynamics in solution. J. Phys. Chem. A **107**:5258–5279, *2003*.

[80] KINCHIN, A. I. *Mathematical Foundations of Statistical Mechanics*. Dover, 1949.

[81] KOLLMAN, P. A., DIXON, R., CORNELL, W., VOX, T., CHIPOT, C., AND POHORILLE, A. The development/application of a 'minimalist' organic/biochemical molecular mechanic force field using a combination of *a*b initio calculations and experimental data. In *Computer Simulation of Biomolecular Systems*, A. Wilkinson, P. Weiner, and W. F. van Gunsteren, Eds., vol. 3. Kluwer/Escom, 1997, pp. 83–96.

[82] KONDRASHOV, D. A., CUI, Q., AND PHILLIPS JR., G. N. Optimization and evaluation of a coarse-grained model of protein motion using x-ray crystal data. Biophys. J. **91**:2760–2767, *2006*.

[83] KRIMINSKI, S., KAZMIERCZAK, M., AND THORNE, R. E. Heat transfer from protein crystals: Implications for flash-cooling and X-ray beam heating. Acta Cryst. **D59**:697–708, *2003*.

[84] KUBE, S., AND WEBER, M. Identification of metastabilities in monomolecular conformation kinetics. ZIB-Report, *2006*.

[85] KUMAR, S., BOUZIDA, D., SWENDSEN, R. H., KOLLMAN, P. A., AND ROSENBERG, J. M. The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J. Comput. Chem. **13**(8):1011–1021, *1992*.

[86] KUNZ, R. E., AND BERRY, R. S. Statistical interpretation of topographies and dynamics of multidimensional potentials. J. Chem. Phys. **103**(5):1904–1912, *1995*.

[87] LEDNEV, I. K., KARNOUP, A. S., SPARROW, M. C., AND ASHER, S. A. 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*.

[88] LENZ, P., ZAGROVIC, B., SHAPIRO, J., AND PANDE, V. S. Folding probabilities: A novel approach to folding transitions and the two-dimensional ising-model. J. Chem. Phys. **120**(14):6769–6778, *April 2004*.

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

[90] LEVY, Y., JORTNER, J., AND BECKER, O. M. Dynamics of hierarchical folding on energy landscapes of hexapeptides. J. Chem. Phys. **115**(22):10533–10547, *2001*.

[91] LEVY, Y., JORTNER, J., AND BERRY, R. S. Eigenvalue spectrum of the master equation for hierarchical dynamics of complex systems. Phys. Chem. Chem. Phys. **4**:5052–5058, *2002*.

[92] LINDORFF-LARSEN, K., BEST, R. B., DEPRISTO, M. A., DOBSON, C. M., AND VENDRUSCOLO, M. Simultaneous determination of protein structure and dynamics. Nature **433**:128–132, *2005*.

[93] LOCKHART, D. J., AND KIM, P. S. Internal Stark effect measurement of the electric field of the amino terminus of an α helix. Science **257**:947–951, *1992*.

[94] LOCKHART, D. J., AND KIM, P. S. Electrostatic screening of charge and dipole interactions with the helix backbone. Science **260**:198–202, *1993*.

[95] MA, A., AND DINNER, A. R. Automatic method for identifying reaction coordinates in complex systems. J. Phys. Chem. B **109**:6769–6779, *2005*.

[96] MA, H., AND GRUEBELE, M. Kinetics are probe-dependent during downhill folding of an engineered λ_{6-85} protein. Proc. Nat. Acad. Sci. USA **102**:2283–2287, *2005*.

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

[98] MAKI, N., MOITRA, K., GHOSH, P., AND DEY, S. 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*.

[99] MANNUZZU, L. M., MORONNE, M. M., AND ISACOFF, E. Y. Direct physical measure of conformational rearrangement underlying potassium channel gating. Science **271**(5246):213–216, *1996*.

[100] MARIANAYAGAM, N. J., FAWZI, N. L., AND HEAD-GORDON, T. Protein folding by distributed computing and the denatured state ensemble. Proc. Natl. Acad. Sci. **102**(46):16684–16689, *2005*.

[101] MEERBACK, E., SCHÜTTE, C., AND FISCHER, A. Eigenvalue bounds on restrictions of reversible nearly uncoupled Markov chains. Lin. Alg. Appl. **398**:141–160, *2005*.

[102] MINTON, A. P. How can biochemical reactions within cells differ from those in test tubes? J. Cell Sci. **119**(14):2863–2869, *2006*.

[103] MORISHITA, T. Fluctuation formulas in molecular-dynamics simulations with the weak coupling heat bath. J. Chem. Phys. **113**(8):2976, *2000*.

[104] MORONI, D., VAN ERP, T. S., AND BOLHUIS, P. G. Investigating rare events by transition interface sampling. Physica A **340**:395–401, *2004*.

[105] MORTENSON, P. N., EVANS, D. A., AND WALES, D. J. Energy landscapes of model polyalanines. J. Chem. Phys. **117**(3):1363–1376, *2002*.

[106] MORTENSON, P. N., AND WALES, D. J. Energy landscapes, global optimization and dynamics of the polyalanine $Ac(ala)_8NHMe$. J. Chem. Phys. **114**(14):6443–6454, *2001*.

[107] MUÑOZ, V., THOMPSON, P. A., HOFRICHTER, J., AND EATON, W. A. Folding dynamics and mechanism of β-hairpin formation. Nature **390**:196–199, *1997*.

[108] NILGES, M., GRONENBORN, A. M., BRÜNGER, A. T., AND CLORE, G. M. 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*.

[109] NOJI, H., YASUDA, R., YOSHIDA, M., AND KINOSITA JR, K. Direct observation of the rotation of F_1-ATPase. Nature **386**:299–302, *1997*.

[110] OGLE, J. M., BRODERSEN, D. E., CLEMONS JR., W. M., TARRY, M. J., CARTER, A. P., AND RAMAKRISHNAN, V. Recognition of cognate transfer RNA by the 30S ribosomal subunit. Science **292**:897–902, *2001*.

[111] OOSTENBRINK, C., VILLA, A., MARK, A. E., AND VAN GUNSTEREN, W. F. A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. J. Comput. Chem. **25**:1656–1676, *2004*.

[112] OPPENHEIM, I., SHULER, K. E., AND WEISS, G. H., Eds. *Stochastic processes in chemical physics: The master equation*. MIT Press, 1977.

[113] OZKAN, S. B., BAHAR, I., AND DILL, K. A. Transition states and the meaning of ϕ-values in protein folding kinetics. Nature Struct. Biol. **8**(9):765–769, *2001*.

[114] OZKAN, S. B., DILL, K. A., AND BAHAR, I. Fast-folding protein kinetics, hidden intermediates, and the sequential stabilization model. Protein Science **11**:1958–1970, *2002*.

[115] PANDE, V. S., BAKER, I., CHAPMAN, J., ELMER, S. P., KHALIQ, S., LARSON, S. M., RHEE, Y. M., SHIRTS, M. R., SNOW, C. D., SORIN, E. J., AND ZAGROVIC, B. Atomistic protein folding simulations on the submillisecond time scale using worldwide distributed computing. Biopolymers **68**:91–109, *2003*.

[116] PANDE, V. S., BAKER, I., CHAPMAN, J., ELMER, S. P., KHALIQ, S., LARSON, S. M., RHEE, Y. M., SHIRTS, M. R., SNOW, C. D., SORIN, E. J., AND ZAGROVIC, B. Atomistic protein

folding simulations on the submillisecond time scale using worldwide distributed computing. Biopolymers **68**:91–109, *2003*.

[117] PARK, S., AND PANDE, V. S. Validation of Markov state models using Shannon's entropy. J. Chem. Phys. **124**:054118, *2006*.

[118] PEROZO, E. Gating prokaryotic mechanosensitive channels. Nat. Rev. Mol. Cell Biol. **7**:109–119, *2006*.

[119] PITERA, J. W., HAQUE, I., AND SWOPE, W. C. Absence of reptation in the high-temperature folding of the trpzip2 β-hairpin peptide. J. Chem. Phys. **124**:141102, *2006*.

[120] RHEE, Y. M., AND PANDE, V. S. One-dimensional reaction coordinate and the corresponding potential of mean force from commitment probability distribution. J. Phys. Chem. B **109**:6780–6786, *2005*.

[121] RHOADES, E., COHEN, M., SCHULER, B., AND HARAN, G. Two-state folding observed in individual protein molecules. J. Am. Chem. Soc. **126**:14686–14687, *2004*.

[122] RIEPING, W., HABECK, M., AND NILGES, M. Inferential structure determination. Science **309**:303–306, *2005*.

[123] RYCKAERT, J., CICCOTTI, G., AND BERENDSEN, H. J. C. 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*.

[124] SALI, A., GLAESER, R., EARNEST, T., AND BAUMEISTER, W. From words to literature in structural proteomics. Nature **422**:216–225, *2003*.

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

[126] SCHULTHEIS, V., HIRSCHBERGER, T., CARSTENS, H., AND TAVAN, P. Extracting markov models of peptide conformational dynamics from simulation data. J. Chem. Theor. Comput., *2005*.

[127] SCHÜTTE, C. *Conformational dynamics: Modelling, theory, algorithm, and application to biomolecules*. PhD thesis, Konrad Zuse Zentrum Berlin, Berlin, Germany, 1999.

[128] SCHÜTTE, C., FISCHER, A., HUISINGA, W., AND DEUFLHARD, P. A direct approach to conformational dynamics based on Hybrid Monte Carlo. J, Comput. Phys. **151**:146–168, *1999*.

[129] SCHÜTTE, C., AND HUISINGA, W. Biomolecular conformations can be identified as metastable states of molecular dynamics. In *Handbook of Numerical Analysis - special volume on computational chemistry*, P. G. Ciaret and J.-L. Lions, Eds., vol. X. Elsevier, 2002.

[130] SCHÜTTE, C., AND HUISINGA, W. Biomolecular conformations can be identified as metastable sets of molecular dynamics. In *Handbook of Numerical Analysis - special volume on computational chemistry*, P. G. Ciaret and J.-L. Lions, Eds. Elsevier, in press.

[131] SHALLOWAY, D. Macrostates of classical stochastic systems. J. Chem. Phys. **105**(22):9986–10007, *1996*.

[132] SHEN, M., AND FREED, K. F. Long time dynamics of met-enkephalin: Tests of mode-coupling theory and implicit solvent models. J. Chem. Phys. **118**(11):5143–5156, *2003*.

[133] SHIRTS, M., AND PANDE, V. S. Screen savers of the world unite! Science **290**(5498):1903–1904, *December 2000*.

[134] SINGHAL, N., AND PANDE, V. S. Error analysis and efficient sampling in Markovian state models for molecular dynamics. J. Chem. Phys. **123**:204909, *2005*.

[135] SINGHAL, N., SNOW, C. D., AND PANDE, V. S. 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*.

[136] SORIN, E. J., AND PANDE, V. S. Empirical force-field assessment: The interplay between backbone torsions and noncovalent term scaling. J. Comput. Chem. **26**:682–690, *2005*.

[137] SORIN, E. J., AND PANDE, V. S. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. Biophys. J. **88**:2472–2493, *2005*.

[138] SPRONK, C. A. E. M., NABUURS, S. B., BONVIN, A. M. J. J., KRIEGER, E., VUISTER, G. W., AND VRIEND, G. The precision of NMR structure ensembles revisited. J. Biomol. NMR **25**:225–234, *2003*.

[139] SRIRAMAN, S., KEVREKIDIS, I. G., AND HUMMER, G. Coarse master equation from Bayesian analysis of replica molecular dynamics simulations. J. Phys. Chem. B **109**:6479–6484, *2005*.

[140] SRIVASTAVA, J., BARBER, D. L., AND JACOBSON, M. P. Intracellular pH sensors: Design principles and functional significance. Physiology **to appear**, *2006*.

[141] STEINBACH, P. J., IONESCU, R., AND MATTHEWS, C. R. Analysis of kinetics using a hybrid maximum-entropy-nonlinear-least-squares method: Application to protein folding. Biophys. J. **82**:2244–2255, *2002*.

[142] STEIPE, B. A revised proof of the metric properties of optimally superimposed vector sets. Acta Cryst. **A58**:506, *2002*.

[143] STRYER, L. *Biochemistry*. W. H. Freeman, New York, 2002.

[144] SWOPE, W. C., ANDERSEN, H. C., BERENS, P. H., AND WILSON, K. R. 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*.

[145] SWOPE, W. C., PITERA, J. W., AND SUITS, F. Describing protein folding kinetics by molecular dynamics simulations: 1. Theory. J. Phys. Chem. B **108**:6571–6581, *2004*.

[146] SWOPE, W. C., PITERA, J. W., SUITS, F., PITMAN, M., ELEFTHERIOU, M., FITCH, B. G., GERMAIN, R. S., RAYSHUBSKI, A., WARD, T. J. C., ZHESTKOV, Y., AND ZHOU, R. 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*.

[147] TANG, K. E. S., AND DILL, K. A. How experiments see fluctuations of native proteins: Perspective from an exact model. Intl. J. Quantum Chem. **75**:147–164, *1999*.

[148] THEOBALD, D. L. Rapid calculation of RMSDs using a quaternion-based characteristic polynomial. Acta Cryst. **A61**:478–480, *2005*.

[149] THOMPSON, P. A., EATON, W. A., AND HOFRICHTER, J. Laser temperature jump study of the helix⇌coil kinetics of an alanine peptide interpreted with a 'kinetic zipper' model. Biochem. **36**:9200–9210, *1997*.

[150] THOMPSON, P. A., MUÑOZ, V., JAS, G. S., HENRY, E. R., EATON, W. A., AND HOFRICHTER, J. The helix-coil kinetics of a heteropeptide. J. Phys. Chem. B **104**:378–389, *2000*.

[151] TOMINAGA, M., AND CATERINA, M. J. Thermosensation and pain. J. Nerobiol. **61**:3–12, *2004*.

[152] TRUHLAR, D. G., GARRETT, B. C., AND KLIPPENSTEIN, S. J. Current status of transition-state theory. J. Phys. Chem. **100**:12771–12800, *1996*.

[153] TUCKERMAN, M., BERNE, B. J., AND MARTYNA, G. J. Reversible multiple time scale molecular dynamics. J. Chem. Phys. **97**(3):1990–2001, *1992*.

[154] TUCKERMAN, M. E., LIU, Y., CICCOTTI, G., AND MARTYNA, G. J. Non-Hamiltonian molecular dynamics: Generalizing Hamiltonian phase space principles to non-Hamiltonian systems. J. Chem. Phys. **115**(4):1678–1702, *2001*.

[155] ULITSKY, A., AND SHALLOWAY, D. Variational calculation of macrostate transition rates. J. Chem. Phys. **109**(5):1670–1686, *August 1998*.

[156] ULITSKY, A., AND SHALLOWAY, D. Erratum on "variational calculation of macrostate transition rates". J. Chem. Phys. **110**(10):4975, *March 1999*.

[157] VAN DEN HEMEL, D., BRIGÉ, A., SAVVIDES, S. N., AND BEEUMEN, J. V. 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*.

[158] VAN KAMPEN, N. G. *Stochastic processes in physics and chemistry*, second ed. Elsevier, 1997.

[159] VÖLKER, J., AND BRESLAUER, K. J. Communication between noncontacting macromolecules.

Annu. Rev. Biophys. Biomol. Struct. **34**:21–42, *2005*.

[160] WANG, J., CIEPLAK, P., AND KOLLMAN, P. A. 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*.

[161] WANG, J., WOLF, R. M., CALDWELL, J. W., KOLLMAN, P. A., AND CASE, D. A. Development and testing of a general Amber force field. J. Comput. Chem. **25**:1157–1174, *2004*.

[162] WEBER, M. *Meshless methods in conformation dynamics*. PhD thesis, Free University of Berlin, 2006.

[163] WEIK, M., KRYGER, G., SCHREURS, A. M. M., BOUMA, B., SILMAN, I., SUSSMAN, J. L., GROS, P., AND KROON, J. Solvent behavior in flash-cooled protein crystals at cryogenic temperatures. Acta Cryst. **D57**:566–573, *2001*.

[164] WILLIAMS, S., CAUSGROVE, T. P., GILMANSHIN, R., FANG, K. S., CALLENDER, R. H., WOODRUFF, W. H., AND DYER, R. B. Fast events in protein folding: Helix melting and formation in a small peptide. Biochem. **35**:691–697, *1996*.

[165] WULF, G., FINN, G., SUIZU, F., AND LU, K. P. Phosphorylation-specific prolyl isomerization: is there an underlying theme? Nature Cell Biol. **7**(5):435–441, *2005*.

[166] YANG, W.-Y., AND GRUEBELE, M. Folding at the speed limit. Nature **423**:193–197, *May 2003*.

[167] YANG, W. Y., AND GRUEBELE, M. Detection-dependent kinetics as a probe of folding landscape microstrucure. J. Am. Chem. Soc. **126**:7758–7759, *2004*.

[168] YANG, W. Y., PITERA, J. W., SWOPE, W. C., AND GRUEBELE, M. Heterogeneous folding of the trpzip hairpin: Full atom simulation and experiment. J. Mol. Biol. **336**:241–251, *2004*.

[169] YOUNGBLOOD, B., AND REICH, N. O. Conformational transitions as determinants of specificity for the DNA methyltransferase EcoRI. J. Biol. Chem. **281**(37):26821–26831, *2006*.

[170] ZHANG, W., LEI, H., CHOWDBURY, S., AND DUAN, Y. Fs-21 peptides can form both single helix and helix-turn-helix. J. Phys. Chem. B **108**:7479–7489, *2004*.

[171] ZHOU, B.-R., LIANG, Y., DU, F., ZHOU, Z., AND CHEN, J. Mixed macromolecular crowding accelerates the oxidative refolding of reduced, denatured lysozyme. J. Biol. Chem. **279**(53):55109–55116, *2004*.

[172] ZHOU, H.-X. Protein folding and binding in confined spaces and in crowded solutions. J. Mol. Recognit. **17**:368–375, *2004*.

[173] ZHUANG, X., KIM, H., PEREIRA, M. J. B., BABCOCK, H. P., WALTER, N. G., AND CHU, S. Correlating structural dynamics and function in single ribozyme molecules. Science **296**:1473–1476, *2002*.

Acknowledgements

Acknowledgments go here.