ALGORITHMS FOR COMPUTATIONAL DESCRIPTION OF LARGE SCALE CONFORMATIONAL TRANSITIONS OF PROTEINS

by Peter Majek

ALGORITHMS FOR COMPUTATIONAL DESCRIPTION OF LARGE SCALE CONFORMATIONAL TRANSITIONS OF PROTEINS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Peter Majek

August 2010

ALGORITHMS FOR COMPUTATIONAL DESCRIPTION OF LARGE SCALE CONFORMATIONAL TRANSITIONS OF PROTEINS

Peter Majek, Ph. D.

Cornell University 2010

Conformational transitions of protein macromolecules are key elements in controlling functionality of proteins by changing structural and functional properties of protein molecules. These transitions consist of structural adjustments at spatial scales of 10-100 Å, between one to two orders of magnitude larger than a typical interatomic distance (2 Å). The difference in temporal scales of atomic motions and conformational transitions spans an even larger range. The transition time, microseconds to milliseconds, is between six to twelve orders of magnitudes larger than typical atomic oscillations (femtoseconds to picoseconds). The atomic resolution of studied systems (10 – 100 thousands of atoms) and long range inter-atomic interactions dictate the computational cost of simulations.

This dissertation discusses several strategies to overcome these scaling issues in computational studies of conformational transition: the temporal, spatial, and size scales. The presented algorithms provide thermodynamics, kinetics and structural descriptions of conformational transitions at overall computational costs several orders of magnitude lower than the straight forward Molecular Dynamics approach. The presented algorithms are based on combination of coarse-graining strategies of (i) boundary value approach by an action minimization, (ii) statistical coarse-grained potentials, and (iii) Milestoning algorithm with an extension to complex (nonlinear) reactions. All presented algorithms are implemented in MOIL molecular modeling

package and are parallelized to run effectively on high performance computing clusters.

BIOGRAPHICAL SKETCH

Peter Májek was born in Malacky, a small town in Slovakia on December 20, 1981. In June of 2000, he graduated from Gymnazium of Jur Hronec in Bratislava. He then spent five years at Comenius University, majoring in Informatics. In July 2005, Peter entered Tri Institutional graduate program in Computational Biology and Medicine at Cornell University, Weill Cornell Medical College, and Memorial Sloan Kettering Cancer Center. On June 27, 2009 Peter married Lucia Balážová who gave birth to their first son, Dominik, in March of 2010. Peter has graduated with a PhD after five years, from which he spent more time at University of Texas at Austin than in Ithaca. He "looks forward" to a life of making use of his graduate school experience in his home country.

To Lucka and Dominik.

ACKNOWLEDGMENTS

I would like to extend my sincerest gratitude towards people who supported me during my time at graduate school. This, of course, includes the members of my committee, David Shalloway, Harel Weinstein, Doug James, and my advisor Ron Elber. Without Ron's scientific and personal guidance I would get lost on this way many times. I have learned much more from friendly discussions with him than from any class or a paper.

My stay at gradschool would not have been possible without the constant support from my family and many of my friends. Thanks to my parents for their love and support during all those years. Thanks to God for through his blessings all this was possible. Big love to my wife, Lucka, who always motivated me to finish this long distance run, she sacrificed and came to Austin to be there for me.

Big thank-you goes to all my lab mates from CLSB lab in Austin. Without them, my life of a graduate student would be a miserable one. Thank to all friends from Ithaca campus, especially Becky Stewart, who did for me many favors to keep my status at Cornell while I was thousand miles away.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ORIGINAL RESEARCH

Research reported in this dissertation is based on the following publications

Chapter 2 – Peter Májek, Harel Weinstein, and Ron Elber: *Pathways of conformational transitions in proteins* in Coarse-Graining of Condensed Phase and Biomolecular systems, ed. Gregory Voth (Taylor and Francis Group, LLC, 2008)

Chapter 3 - Peter Májek and Ron Elber: *A coarse-grained potential for fold recognition and molecular dynamics simulations of proteins*, Proteins: Structure, Function, and Bioinformatics. 2009; **76**(4): 822-36.

Chapter 4 – Peter Májek and Ron Elber: *Milestoning without a reaction coordinate*, accepted to Journal of Chemical Theory and Computation

Other related publications of the author

Jas G. S., Hegefeld W., Májek P., Kuczera K., Elber R., Bax A., and Eaton W. A.: *Fastest Simplest Helix*, in preparation

Zheng Yang, Peter Májek, and Ivet Bahar: *Allosteric Transitions of Supramolecular Systems Explored by Network Models: Application to Chaperonin GroEL*, PLoS Computational Biology. 2009; **5**(4): e1000360.

Vallat B. K., Pillardy J., Májek P., Meller J., Bloom T., Cao B., and Elber R.: *Building and assessing atomic models of proteins from structural templates*, Proteins: Structure, Function, and Bioinformatics. 2009; **76**(4): 930-45

CHAPTER 1

INTRODUCTION

Proteins are organic macromolecules responsible for most of the biological functions in living cells. Proteins are formed as linear chains of chemical subunits called amino acids, where most of them are formed of twenty different types of amino acid residues. The majority of known proteins form a well-defined three dimensional (native) structure by a process called protein folding. The native structure of a protein is determined directly by the linear sequence of amino acids, although in some cases the folding process of a protein *in vivo* is assisted by chaperones.

However, it would be misleading to think about proteins as rigid rocks as it is now well established that proteins are in constant motion, sampling an ensemble of conformations. Additionally, by undergoing changes between different conformations, proteins carryout their biological functions. Understanding the mechanism of transitions between different conformations is of major importance to designing methods for controlling such transitions, and thereby modulating protein functions. However, exploring the transitions between conformations is hard, both experimentally and computationally, due to the transient nature of the intermediate high energy conformers encountered as the molecule undergoes structural changes. In many cases, only the two end structures of a biological process of interest are known from experiments.

These conformational transitions consist of structural rearrangements at spatial scales of 10-100 Å, between one to two orders of magnitude larger than a typical interatomic distance (2 Å). The difference in temporal scales of atomic motions and conformational transitions spans an even larger range. The transition time of biologically important conformation rearrangements is in microseconds to

milliseconds, what is between six to twelve orders of magnitudes larger than typical atomic oscillations (femtoseconds to picoseconds). The detailed atomic resolution of studied systems (10 – 100 thousands of atoms) and long range inter-atomic interactions result in large computational cost of simulations. Analysis of conformational transitions is further complicated by the fact that often the passage between the two end points does not involve a single unique pathway, but an ensemble of qualitatively different pathways is realized.

In this dissertation, we address the above obstacles in a computational description of conformational transitions of proteins. In Chapter 2 we examine a boundary value approach calculating an ensemble of plausible large scale conformational transitions in proteins. The plausible trajectories are found by minimizing an action as a property of the whole trajectory represented on a coarse temporal resolution. Our strategy to scale to systems of large size (hundreds to thousands of amino acids) and still maintain feasible computational cost is to perform initial (fast) calculations in a coarse-grained model and then refine resulting trajectories in an all-atomistic action minimization calculation.

In Chapter 2, a double–minima generalization of an elastic network model is used in coarse-grained calculations. Such a coarse-grained model suits well its purpose if the spatial fluctuations of the structure during a conformational transition are small. If however, the transition involves significant modifications of the protein structure, a more sophisticated coarse-grained model is needed. In Chapter 3, we describe an algorithm for the design of differentiable coarse grained force field that reproduces thermo-dynamical properties of experimental structures and at the same time performs well in native structure recognition.

Finally, in Chapter 4, we describe an extension of Milestoning algorithm that can be used for calculation of quantitative kinetical and thermodynamical entities of

complex reactions described for example by multiple possible reaction channels obtained by methods from the second chapter. We defer the further introduction of topics related to each of the above problems to the introductory subsections of particular chapters.

CHAPTER 2

CONFORMATIONAL TRANSITIONS BY A BOUNDARY VALUE SOLVER AND COARSE-GRAINING

2.1 Introduction

This chapter is about coarse graining of pathways and trajectories of proteins in action. In particular we focus on protein switches that flip between different structures; flips that modify their activity. These switches offer another layer of control and are of considerable interest in current fields of study such as System Biology (Alon 2006). It is the collective behavior and interactions of many protein molecules that is the topic of biological networks, and such switches are essential components of the functional mechanisms represented by the networks.

The way we compute and think about trajectories is quite different from the widely used Molecular Dynamics approach in which differential equations of motions are solved by an initial value method. Instead we have chosen, for reasons that will become clearer below, to use a boundary value formulation. Our choice of coarse graining and of boundary value calculations to study biological switches requires some discussion, so we start with an analogy.

A useful comparison to the way we compute and analyze molecular paths is a web-search for driving-directions. In such a search one specifies the starting and end locations of the drive and seeks a path that requires minimal time. A web engine analyzes the request and outputs a written description of the driving directions and a two-dimensional map of the roads, highlighting the chosen path. Obviously the web instructions are only guidelines that do not take into account extra traffic due to special events, road works that intervene and require a bypass, and other specific

circumstances. In short, the paths proposed are coarse-grained and averaged over many actual paths that differ in many details but agree in their overall shape. The coarse-grained paths miss many details but are nevertheless very useful when making travel plans. Part of their effectiveness is because we understand their limitations and operate accordingly.

Similarly, in molecular biophysics we are frequently provided with starting and ending configurations of a protein, captured with experimental techniques. These end points are stable for significant periods of time and can be measured with static techniques. We call the initial conformation $\mathbf{x}_i$ and the final structure $\mathbf{x}_f$. For example, X-ray crystallography or NMR can determine structures of activated and de-activated forms of a protein molecule. However, elucidating experimentally the "driving" pathway that links the two forms, and determines the time scale of the transition is considerably more difficult. Structures "in-between" exist for significantly shorter times than the stable end points, and are harder to measure. Computer simulations suggest a useful alternative which, in conjunction with partial experimental data (such as the change of a few distances as a function of time) can provide a comprehensive view of the process. The present chapter is about the computations of such transitions. Let us consider some approaches to molecular simulations and how they can be used to study biomolecular switches.

As mentioned earlier, the most widely used technique for atomically detailed simulations is Molecular Dynamics (MD). In the MD approach we consider the Newton's equations of motions, $\mathbf{M}\ddot{\mathbf{x}} = \mathbf{F}$, where $\mathbf{M}$ is the mass matrix, and $\mathbf{x}$ the coordinate vector of all the particles that we used to model the system. For proteins, the length of $\mathbf{x}$ (the number of degrees of freedom) is typically of order of a few thousands. The double dot on top of the vector $\mathbf{x}$ denotes second derivative with respect to time, and $\mathbf{F}$ is the force vector. In MD, initial values are used for the

integration ($\mathbf{x}$, the coordinate vector and $\mathbf{v} \equiv \dot{\mathbf{x}}$, are given at an initial time, say $t = 0$). The numerical solvers of the initial value problem employ numerical integrators such as Verlet algorithm (Verlet 1967):

$$\begin{aligned} \mathbf{x}(t+\Delta t) &= \mathbf{x}(t) + \mathbf{v}(t)\Delta t + \frac{\Delta t^2}{2}\mathbf{M}^{-1}\mathbf{F}(t) \\ \mathbf{v}(t+\Delta t) &= \mathbf{v}(t) + \frac{\Delta t^2}{2}\mathbf{M}^{-1}\left(\mathbf{F}(t) + \mathbf{F}(t+\Delta t)\right) \end{aligned} \tag{2.1}$$

The time step is Δt. Hence, given coordinates and velocities at a particular time we can generate the coordinates and the velocities at slightly later times. By repeating this process N times we can generate a trajectory (a path) of length $N \cdot \Delta t$. The final structure $\mathbf{x}(N \cdot \Delta t)$ is determined by the initial conditions since the algorithm described in equations (2.1) is deterministic. However, it is hard to predict or to tune with velocity variations the location of $\mathbf{x}(N \cdot \Delta t)$ before doing the actual calculation since the results are very sensitive to the initial value. The hard-to-predict final point is a significant difficulty with the application of MD to biomolecular switches. Unless "all the roads lead to Rome," it is not obvious that $\mathbf{x}(N \cdot \Delta t)$ is the final desired configuration of the switch, $\mathbf{x}_f$. Hence we may be wasting many cycles and computing unsuccessful trajectories while adjusting the initial conditions until $\mathbf{x}(N \cdot \Delta t)$ is the desired final coordinate vector. Notably, MD does not use effectively all the information at hand (i.e., the structures at the two ends) that could facilitate the study of switching mechanisms. It is therefore not surprising that the use of MD is not optimal for this problem.

Even if the two end points are strong attractors (i.e., "all the roads indeed lead to Rome"), the transition may require a large number of integration steps (the basic integration step Δt must be small) making the calculation (again) inefficient. Some techniques (Dellago, Bolhuis et al. 2002) use initial value formulation, starting

somewhere in between, to compute rare *short-time* trajectories between strong attractors. This allows for the sampling of trajectories between states separated by large and narrow barriers (activated processes) (Bolhuis, Chandler et al. 2002). However these conditions are not satisfied for conformational transitions of the type discussed here. An example for a broad barrier that leads to long transitional trajectories is the conversion in the Myosin molecule (West and Elber 2010). The time scale for Myosin post-recovery (a part of the transition that controls muscle motions) is in the milliseconds $\left(10^{-3}s\right)$ while the typical size of an integration step in MD is a femtosecond $\left(10^{-15}s\right)$. The number of integration steps with MD required to simulate the transition in myosin is hopelessly large $\left(10^{12}\right)$. Why is the step so small?

The reason is that many molecular motions are very rapid (e.g., bond vibrations, collisions) so that in order to maintain the numerical stability of an algorithm like Verlet the step has to be significantly smaller than the time scale of fastest motions in the system. Significant effort was therefore invested (Schlick, Skeel et al. 1999) into algorithms that increase the time steps, and into factoring out rapid motions. Perhaps the most widely used algorithm that eliminates certain categories of rapid motions is SHAKE (Ryckaert, Ciccotti et al. 1977), in which the fast bond vibrations are constrained to their ideal values. Typically used for bonds of heavier atoms with Hydrogen (x-H), SHAKE allows the increase of the time step by about a factor of two, but probably not more. The problem is that other rapid motions (non-bonded collisions between atoms) remain after the removal of bond vibrations and also require small time steps. The latter are much harder to factor out rigorously since their internal coordinate representation (the identity and distance between a pair of colliding atoms) is changing during the progress of the trajectory. Some atoms that are close to each other at a particular time (colliding) may be separated at a later time of the process in which other atoms may collide. While a special treatment for collision

can be worked out (Ulitsky and Elber 1993; Ulitsky and Elber 1994), the change of colliding partners requires complex bookkeeping, which is expensive computationally. Nevertheless, the overall success and wide use of the SHAKE algorithm suggests that other reductions in the details of the spatial description of the system may be useful. There are many approaches to perform spatial coarse-graining to simplify force calculations, reduce the number of degrees of freedom, and enable longer time and more comprehensive sampling with the simplified spatial description (Voth 2009). These reductions are established using a number of physical assumptions and approximations, and they do not solve the problem of double-ended trajectories. Furthermore, atomically detailed description may be necessary to understand many biophysical and biochemical processes. Giving it up with spatial coarse graining may lose critical elements of it.

The approach we discuss in the next section, which we have used for more than 10 years now, provides a coarse-grained description of the path (like driving directions at different resolution), while keeping a complete description of the atomic coordinates of the system. We are able to do it since the boundary value formulation is numerically more stable than initial value solvers. The boundary value representation makes it possible to use much larger path steps than is possible with initial value solvers. One must keep in mind however, that the trajectories so produced are approximate.

2.2 Theory of boundary value formulation of pathways and trajectories

Newton's equations of motion are usually derived in an analytical mechanics course (Landau and Lifshitz 1976) from the classical action. The classical action, S, is defined as

$$S\left[\mathbf{x}(t)\right] = \int_{0,\mathbf{x}_i}^{t,\mathbf{x}_f} L\left(\mathbf{x}(\tau), \dot{\mathbf{x}}(\tau)\right) d\tau$$
$$L = \frac{1}{2}\dot{\mathbf{x}}^{\mathrm{T}}\mathbf{M}\dot{\mathbf{x}} - U\left(\mathbf{x}\right) \qquad (2.2)$$

Then, Newton's equations of motion are derived from the condition that the action, which is a functional of the trajectory, is stationary with respect to variations in the path (Landau and Lifshitz 1976). The variations do not change the end configurations, illustrating that a solution of Equation (2.2) is indeed a result of a boundary value problem. However such a derivation will not help in the switching problem. A more straightforward approach to solve Equation (2.2) is to use a finite difference approximation to the integral. For example:

$$S = \sum_{j=1}^{N-1} 1/\left(2\Delta t\right)\left(\mathbf{x}_{j+1} - \mathbf{x}_j\right)^{\mathrm{T}} \mathbf{M}\left(\mathbf{x}_{j+1} - \mathbf{x}_j\right) - U\left(\mathbf{x}_j\right)\Delta t\,.$$

The action is now a function (not a functional!) of all the intermediate structures $\mathbf{x}_j$. When the action is stationary with respect to all these coordinates ($\partial S / \partial \mathbf{x}_j = \mathbf{0} \quad j = 2,...,N-1$) then the sequence of coordinates provides a finite difference approximation to a classical trajectory. The approximation is better for smaller Δt but it is stable for any time step. This stability distinguishes the approach from initial value solvers that "explode" at time steps that are too large. There are however good reasons why finding a stationary solution of the classical action is not a popular way of computing trajectories. First the action is not necessarily a minimum but a stationary point (e.g. a classical trajectory can be a saddle point of the action). Searches for stationary points are numerically more difficult than for a minimum. Second, the optimization is of a function of a very large number of variables. If the bio-molecular switch is described with K atoms then the number of degrees of freedom for the action optimization is $3\times K\times N$ (K and N are in the thousands for a

typical calculation). Third, time is not a good variable for parameterizing or indexing the path. The last surprising statement is in the sense that the parameterization determines *the density of points along the path*, and the time density can be very different from the *spatial* density. Let us return one more time to the driving analogy and parameterize the path according to the time of the drive. A realization of this parameterization is to draw a dot on the map at constant time intervals (say every minute) of the drive. The path may include a segment in which the car is likely to move slowly or even stop and a segment on a free road or the highway in which the car moves very quickly. If we distribute the dots that describe the path equally in time we will have high density at a traffic jam and dilute description of the path on the highway. Tracking the path visually under these circumstances is not ideal. A better path representation (at least from a visualization perspective) is to have the dots equally spaced, say every 200 meters. More precisely, the above suggestions for alternative parameterization means to replace the parameterization of the path with respect to time $\mathbf{x}(t)$ by the parameterization $\mathbf{x}(l)$, where l is the arc-length of the path in mass weighted coordinates, $dl = \left|\mathbf{M}^{1/2}\dot{\mathbf{x}}\right| dt$. The classical action as a function of length is (Landau and Lifshitz 1976)

$$\overline{S} = \int_{\mathbf{x}_i}^{\mathbf{x}_f} \sqrt{2(E-U)}\, dl \tag{2.3}$$

And in a discrete representation

$$\begin{aligned} \overline{S}\left[\{\mathbf{x}_j\}_{j=1}^{N}\right] &= \sum_{j=1}^{N-1} \frac{1}{2}\left(\sqrt{2\left(E-U(\mathbf{x}_j)\right)} + \sqrt{2\left(E-U(\mathbf{x}_{j+1})\right)}\right)\Delta l_{j,j+1}, \\ \Delta l_{j,j+1} &= \left|\sqrt{\mathbf{M}}\mathbf{x}_j - \sqrt{\mathbf{M}}\mathbf{x}_{j+1}\right| \end{aligned} \tag{2.4}$$

A classical trajectory will be obtained when for all intermediate configurations we have $\partial\overline{S}/\partial\mathbf{x}_j = \mathbf{0}$. The parameterization with respect to the arc-length is indeed more convenient numerically than with respect to time. However, we are still faced with the need to compute a stationary action instead of a minimum. Furthermore, Eq. (2.3) is always non-negative; as such it has an undesired minimum at $E = U$ in which the first-order variation is discontinuous. This is the classical turning point (zero kinetic energy) in which the trajectory may get stuck. The derivative is not continuous, nor zero for that path so it is not a true classical trajectory. However, attempts for direct minimization of $\overline{S}$ may pick it up. It is therefore better to work with the Gauss action in length (Elber 2006).

The Gauss action (written originally for the time dependent formulation as $S_{Gauss} = \int_0^t \left[\delta S/\delta\mathbf{x}(t')\right]dt' = \int_0^t \left(\mathbf{M}\ddot{\mathbf{x}} + dU/d\mathbf{x}\right)^2 dt$ (Lanczos 1970)) is trivially extended to the length formulation and the finite difference formula as $S^l_{Gauss} \approx \sum_{j=2}^{N-1} (\partial\overline{S}/\partial\mathbf{x}_j)^T (\partial\overline{S}/\partial\mathbf{x}_j)$ (for explicit formulas of the derivatives of $\overline{S}$ see Appendix B). A direct minimization of the function S^l_{Gauss} will provide an approximate classical trajectory as a function of length. We typically perform this minimization with simulated annealing (SA). Let $\mathbf{y}$ be the vector of the joint set of all coordinates $\mathbf{y} \equiv (\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N)$, then the simulated annealing procedure uses S^l_{Gauss} as the potential energy and the $\mathbf{y}$ as variables as follows. We integrate the following stochastic differential equation for the *whole trajectory*: $\ddot{\mathbf{y}} + \gamma\dot{\mathbf{y}} + \partial\mathbf{S}^l_{Gauss}/\partial\mathbf{y} = \mathbf{R}$, where γ is the friction constant, and $\mathbf{R}$ is a Gaussian random force sampled according to the conditions $\langle\mathbf{R}\rangle = \mathbf{0}$ $\langle\mathbf{R}^2\rangle = 2\gamma T\delta(t)\mathbf{I}$. The temperature, T, is reduced monotonically to zero at which point a minimum of the path is recovered. At finite temperatures it is possible to sample plausible paths to form a collection of trajectories depending on allowed variance in the value of the action (Elber, Meller et al. 1999).

A critical advantage of the boundary value calculations compared to MD is that the step can be very large while still providing correct qualitative behavior of a classical trajectory. The numerical stability of the optimization process is poorer with initial value solvers (with the exception of the Backward-Euler algorithm of Peskin and Schlick (1989) which is stable for large steps but still cannot be aimed to a desired product). For example, in the simulation of the folding of Cytochrome c (Cardenas and Elber 2003) we have used 1,000 length slices to provide a coarse grained description of the folding pathway; the resulting coarse-grained path was consistent with numerous experimental observations, but was of course approximate. Since the time scale of folding of Cytochrome C is in milliseconds, about 10^{12} steps would be required with an initial value solver (straightforward MD). This number of steps is nine orders of magnitude larger than the number of steps was used in the boundary value formulation.

To start the SA algorithm we need to specify an initial trajectory $\mathbf{y}^0$. The simplest initial guess for a trajectory is a Cartesian linear interpolation between the two end points. However, the energies of structures along linear paths are usually very high since they distort the covalent structure of the protein chain and have significant steric overlaps. These initial guesses require substantial optimization that is not always successful. The generation of the initial paths can benefit from *spatial* coarse graining which we discuss in Section 2.4.

In addition to an initial guess for the path, the total energy of the system E is needed for the calculation of the functional. An obvious try would be to use the average thermal energy, which is a sum of the average potential, $\langle U \rangle$, and average kinetic energy, $\langle K \rangle$, (if we would have computed a large number of trajectories then it would make sense to sample from the distribution of these energies instead of using the averages). However, our approximate procedure to compute trajectories introduces

a subtle complication. The trajectories with large steps in time or in the arc-length do not include motions with high frequencies (Olender and Elber 1996) (i.e., bond oscillations), and these degrees of freedom do not contribute to the thermal energy. The number of fast degrees of freedom is uncertain since the number of the (transient) collisions is not known. Since the collision between pairs of particles takes only a small fraction of time, the amount of filtering is also uncertain. If we use a lower bound for the filtering and consider only the bonds and the angles of the protein molecule (the number of bonds or angles is of order K - the number of atoms), the thermal energy of the non-filtered motions is approximately $E \approx (3K - 2K)(k_B T / 2) + \langle U \rangle = K(k_B T / 2) + \langle U \rangle$. This is the energy that we use for the functional (2.3).

2.3 Path constraints

In this section we describe a number of constraints that are imposed during the calculations of the path to ensure correctness, given the typical computing environments used for molecular modeling and simulations. During simulations of large molecules we use Cartesian coordinates for which the equations of motion are simple to write and manipulate. However, the Cartesian representation requires a reference frame. Changes in the reference frame of one structure along the path affect the distance between sequential coordinate sets $\Delta l_{j,j+1}$. Therefore, the same reference frame must be used for all the coordinate sets along the path. Fixing the reference system is achieved by applying the Eckart conditions (Elber 1990): six constraints on overall translations and rotations on each of the structures along the path.

$$\sum_k m_k \mathbf{r}_k = \mathbf{0} \quad \sum_k m_k \mathbf{r}_k^i \times \mathbf{r}_k = \mathbf{0} \quad \text{or} \quad \sigma_{i=1,\ldots,6} = 0\,, \tag{2.5}$$

where m_k, $\mathbf{r}_k$, and $\mathbf{r}_k^i$ are the mass, the position, and the initial position of the k^{th} atom respectively. We assume without loss of generality that the initial position of the center of mass is zero.

To ensure that the protein structures in the set are equally spaced along the trajectory we also use a harmonic restraint to keep the distances between all the structures the same. The variational principle of the action as a function of length provides a condition on the motion perpendicular to the path, but it does not explicitly constrain the motion along the path. Therefore the constraint has no impact on the equations of motion.

$$\eta_1 \sum_j \left(\Delta l_{j,j+1} - \langle \Delta l \rangle\right)^2 \quad \langle \Delta l \rangle = \frac{1}{N-1} \sum_{j=1}^{N-1} \Delta l_{j,j+1} \tag{2.6}$$

The strength of the restraint (2.6) is controlled by η_1, which should be chosen as high as possible. It should be kept in mind though that η_1 values that are too high would make the equations of motion for the annealing stiff and would require a much smaller and less efficient integration step. The same type of constraint was used in the calculation of approximate minimum energy paths and in the calculations of the steepest descent path (Elber and Karplus 1987; Czerminski and Elber 1990; Jonsson, Mills et al. 1998). Another way of implementing the equi-distance constraints is via the formulation of Lagrange's multipliers (E, Ren et al. 2002). The Lagrange multiplier approach for dealing with the equi-distance constraints was used also in the calculations of minimum energy and minimum free energy paths (E, Ren et al. 2002; Weinan, Ren et al. 2005).

At the beginning of the calculation, we also use the penalty function $\eta_2 \sum_{j=2}^{N-1} 1/[E - U(\mathbf{x}_j)]$, which makes the algorithm more stable by forcing the potential energy, U, along the whole trajectory to be smaller than the total energy *E*.

Without this additional term, the terms under the square roots in (2.4) may be negative and make the optimization ill-defined when the trajectory is far from optimal and the structures are highly distorted. After some annealing, the trajectory converges to a neighborhood of a true classical trajectory and the value of η_2 is gradually reduced to zero. The optimized trajectory is not sensitive to the initial value of η_2.

The final target function that is used in the algorithm is

$$T = \sum_{j=2}^{N-1}\left[\left(\frac{\partial \bar{S}}{\partial \mathbf{x}_j}\right)^{\mathrm{T}}\left(\frac{\partial \bar{S}}{\partial \mathbf{x}_j}\right)\right] + \eta_1 \sum_{j=1}^{N-1}\left(\Delta l_{j,j+1} - \langle \Delta l \rangle\right)^2 + \eta_2 \sum_{j=2}^{N-1}\frac{1}{E - U(\mathbf{x}_j)}. \qquad (2.7)$$

Stochastic optimization of T is performed similarly to the procedure described for S^l_{Gauss}, except that the constraints on translations and rotations of the system (2.5) are solved explicitly (the constraints are linear – see Appendix of (West, Elber, and Shalloway 2007)). We call the optimization of T, an SDEL calculation (Stochastic Difference Equation in Length (Elber, Ghosh, and Cardenas 2002)).

The formulation in Eq. (2.7) is applicable to any type of dynamics between two fixed end points that can be described by an action. Another choice of action implemented in our molecular dynamics simulation software MOIL (Elber et al. 1995) is an action that provides approximate most probable Brownian trajectories and the intrinsic reaction coordinate (Steepest Descent Path (Olender and Elber 1997; Elber and Shalloway 2000)). Calculating reaction coordinates with boundary value formulation and action minimization is intriguing, since local calculations of reaction coordinates suffer from similar problems as the calculations of trajectories. For example, they are difficult to direct to desired product states. Other global algorithms for path optimization are available (Ulitsky and Elber 1990; Jonsson, Mills et al. 1998; E, Ren et al. 2002), however, they are not based on a global optimization of an action, which makes their calculation less robust. In essence they are similar to the direct

optimization of the classical action, which is a saddle point, and equivalent to solving a large number of differential equations simultaneously. The advantage of having a target function to optimize is the global quality control it provides. As long as the function value is reduced, a large step can be accepted. In contrast, solving differential equations requires small steps and locally controlled accuracy. The action that we used for the approximate Brownian trajectories is

$$S = \int_{\mathbf{x}_i}^{\mathbf{x}_f} \sqrt{H_S + \left(\frac{dU}{d\mathbf{x}}\right)^T \left(\frac{dU}{d\mathbf{x}}\right)} dl \,, \tag{2.8}$$

where the constant H_S is zero for the calculation of minimum energy path (Elber and Shalloway 2000).

Another interesting feature of the boundary value trajectories is the ability to parallelize the code efficiently. This is in contrast to initial value solvers in which only the calculations of the forces can be made parallel at considerable communication cost. In the boundary value formulation every time (or arc-length) slice can be optimized on a different CPU (Zaloj and Elber 2000). For a detailed description of the parallelization of the algorithm and recent improvements of the SDEL implementation see Appendix A.

2.4 Spatial coarse graining

For proteins with several hundred amino acids, computing an atomically detailed trajectory starting from an initial guess (e.g., linear interpolation) far from the optimal trajectory can be a formidable task. With the resources available to us we are able to perform simulated annealing runs that optimize the initial trajectory locally but do not perform global search for alternative pathways. The algorithm puts

considerable effort in adjusting local positions of all atoms. It is less effective in relaxing collective variables of the transition that extend over significant length scales. Similar in spirit to multi-grid methods (Briggs 1987), it is worth separating the optimization of path to global and local length relaxations. Otherwise the relaxation of global variables will be slowed down by the "noise" of the local variables. In our car driving analogy this would correspond to an algorithm that tries to calculate the best driving directions by considering all car types and their conditions, experience of the driver and his level of knowledge of the neighborhood, and so on, before having a general appreciation of the driving route. The additional factors can slow down the speed of the calculations considerably, while their benefit is not obvious at the beginning of the calculations. Average properties of cars and drivers are simpler to use and are providing useful pathways. It makes sense to consider first pathways of average properties that will be refined later (if necessary) according to additional information available at the time of evaluation.

We can use a related idea for conformational transitions. First we determine a trajectory for a system of reduced dimensionality that we believe captures the global characteristics and relaxation of the path. The coarse-grained (CG) trajectory provides the backbone on which an atomically detailed trajectory is constructed and refined by the SDEL methodology

Obviously, there are numerous choices of how to coarse grain atomically detailed systems, and the choice is far from obvious or unique. Spatially CG models have been successfully used for several years to model behavior of complex biomolecular systems. In these CG models a molecule is represented by a reduced set of representative points, where a point corresponds to at least several atoms. A typical reduction of a protein that we use is to keep only the position of the C_α atom of each amino acid. One model potential energy of this reduced representation is (Tirion 1996;

Haliloglu, Bahar, and Erman 1997; Xu, Tobi, and Bahar 2003; Lu, Poon, and Ma 2006)

$$U = \frac{\kappa}{2} \sum_{r_{ij}<C} (r_{ij} - r_{ij}^0)^2 , \tag{2.9}$$

where κ is a force constant in Kcal mol^{-1}/m^2, C is a distance cutoff, and r_{ij}^0, and r_{ij} are the distances between the C_α atoms of residues *i* and *j* in the native and the current conformation respectively. This is the Anisotropic Network Model (ANM) (Xu, Tobi, and Bahar 2003), an extension of the simpler Gaussian Network Model (GNM) (Haliloglu, Bahar, and Erman 1997). It has been shown, that even these simple potentials provide a very good agreement with X-ray experimental B-factors (see for example (Yang et al. 2007)), and therefore may give adequate descriptions of the system dynamics in the neighborhood of the native conformation.

The quadratic functional form of Eq. (2.9) cannot describe multiple minima and the barriers separating them. Hence it is not an adequate model to represent transitions between stable states. To allow the study of transitions with simplified network models Maragakis and Karplus (2005) used the Empirical Valence Bond (EVB) theory of Warshel (Aqvist and Warshel 1993) and generalized the simple ANM. Two ANM models U_i and U_f are defined for the reactants and the products respectively. The EVB computes a new potential *U* that interpolates between the two models

$$U(\mathbf{x}) \equiv \frac{1}{2}\left(U_i(\mathbf{x}) + (U_f(\mathbf{x}) - \alpha) - \sqrt{\left(U_i(\mathbf{x}) - (U_f(\mathbf{x}) - \alpha)\right)^2 + 4\beta^2} \right). \tag{2.10}$$

The scalar α is the energy gap between the two minima and β is a coupling constant that helps tune the barrier height and smoothness.

2.5 Stochastic dynamics

For the coarse grained model we consider Brownian dynamics

$$\gamma\dot{\mathbf{x}} = -\nabla U(\mathbf{x}) + \mathbf{R}(t), \tag{2.11}$$

where γ is a friction constant and $\mathbf{R}(t)$ is a random force with normal distribution ($\langle \mathbf{R}(t) \rangle = \mathbf{0}$ and $\langle \mathbf{R}(t_i)\mathbf{R}^T(t_j) \rangle = 2\gamma k_B T \delta(t_i - t_j)\mathbf{I}$). The boundary value formulation in Brownian dynamics settings was discussed by Elber and Shalloway (2000). They showed that the *most probable* trajectory in approximate and discrete variant of Brownian dynamics minimizes the following action (see also Eq. (2.8))

$$\mathrm{S}_{BD}(\mathbf{x}_2, \ldots, \mathbf{x}_{N-1} \mid \mathbf{x}_1 = \mathbf{x}_i, \mathbf{x}_N = \mathbf{x}_f, H_s) = \sum_{j=1}^{N-1} \sqrt{H_S + \left(\partial U \big/ \partial \mathbf{x}_j \right)^2} \left| \mathbf{x}_{j+1} - \mathbf{x}_j \right|. \tag{2.12}$$

The constant H_S mimics the energy in classical mechanics and can be chosen empirically. If $H_S \to 0$, the optimal trajectory that minimizes S_{BD} is the steepest descent path. On the other hand, if $H_S \to \infty$, then the linear interpolation between $\mathbf{x}_i$ and $\mathbf{x}_f$ (the shortest trajectory) is the optimal path. Varying the parameter H_S provides a set of optimal CG paths with different thermal energies. The same simulated annealing algorithm as applied for all-atom S^l_{Gauss} minimization is used for optimization of S_{BD}. Since the CG model is much simpler than an all-atom model, path searches can be performed comprehensively.

2.6 Refinement of coarse grained trajectories to atomic scale

Once a set of most probable coarse-grained trajectories is obtained by minimization of S_{BD} for different values of H_s, we can return to the initial task of

finding approximate long time (arc-length) trajectories for the all-atom representation. First, atomically detailed structures are built based on the CG shapes along the trajectories. For a given length slice *j* and residue *k*, the following reconstruction is applied: Let C_j^k be a position of the k^{th} C_α atom in the j^{th} frame. For each residue, in each frame, the rigid body transformations $\mathrm{T}_{1\to j}^k$: of $C_1^{k-l},\ldots,C_1^{k+l}$ to $C_j^{k-l},\ldots,C_j^{k+l}$ and $\mathrm{T}_{n\to j}^k$: of $C_n^{k-l},\ldots,C_n^{k+l}$ to $C_j^{k-l},\ldots,C_j^{k+l}$ are calculated. The parameter *l* defines the size of the local neighborhood. If $l=0$, a single C_α atom is considered at a time and only translational transformation can be determined, for $l=1$ the local neighborhood is defined by a triplet of consecutive C_α atoms (which are not linear in proteins) and both the translation and the rotation can be determined uniquely. In the actual implementation, we have used $l=2$, which is more stable in capturing the local neighborhood. The position of a non-C_α atom *A* in the length slice *j,* belonging to the residue *k* is reconstructed as a linear interpolation of its transformed positions from the initial and the final frame:

$$A_j^k \;=\; \frac{n-j}{n-1}\mathrm{T}_{1\to j}^k A_1^k \;+\; \frac{j-1}{n-1}\mathrm{T}_{n\to j}^k A_n^k \qquad (2.13)$$

After this interpolation of non-C_α atoms the intermediate structures along the trajectory have unreasonably high potential energies, which must be reduced by minimizations before the all-atom SDEL computation can be used for the refinement of the path. The minimizations of the structures find local minima in the neighborhood of the initial structures and therefore do not change the paths significantly. There are three processes of minimization: (i) Minimization with soft (core) potential to eliminate truly bad van der Waals contacts, (ii) Minimization with regular Lennard-Jones potential, and (iii) Short molecular dynamics simulation at 10K with harmonic restraints on the positions of the C_α to escape undesired local minima. Typical

numbers of minimization steps for the Glutamate receptor problem described in Section 2.7, are 100, 200, and 1000 for each of the three minimization processes respectively. With a set of plausible trajectories in atomically detailed representation, the SDEL algorithm is executed to obtain physically and energetically sound pathways.

2.7 The allosteric transition of mGluR1 receptor

In this section we describe the calculation of the conformational transition in the extracellular (ligand binding) region of the metabotropic Glutamate receptor (mGluR). The mGluRs are membrane proteins that mediate the transmission of a signal into the cell after binding a glutamate molecule in the extracellular domain. These receptors belong to Class C of G protein-coupled receptors (GPCR). There are three different subgroups of mGluR's , termed I, II, and III, which do not differ much in their overall molecular architecture. Thus, the mGluR receptor is divided structurally into three regions: the extracellular region, the transmembrane region composed of a seven helix bundle, and the cytoplasmatic region. The extracellular region consists of the ligand-binding region (LBR) and the cysteine-rich domain (CRD) (Muto et al. 2007, Pin and Acher 2002; Pin et al. 2004, 2005)

Experimental structures for the two states of the LBR of the extracellular part of the mGluR1 (belonging to subgroup I of mGluR) are available (PDB entries 1ewk, 1ewt) (Kunishima et al. 2000). The receptor functions as a homodimer; consisting of 980 residues (490 amino acids per monomer). However, not all residues of the LBR were resolved by X-ray crystallography and in each monomer a loop of approximately 30 residues is missing. We found, however, that MD simulations of the PDB construct maintain a stable structure in nanosecond-length simulations, indicating that the

missing loop segments can be ignored for the present study. The protocol of generating the path from the PDB structures is as follows:

(i) The energy of the two PDB structures is minimized.

(ii) The C_α representation of the molecules is kept and coarse-grained reactive trajectories between the two end points are computed with simulated annealing as described in section 2.2 for different values of H_S (by minimizing formula (2.12). We have used 100 structures distributed along the path to represent the transition. The action S_{BD}, is minimized with 100K steps of simulated annealing. This requires about 10 hours on a typical CPU for about 1000 residues. The steepest descent path (a minimum of Eq. (2.12) with $H_S = 0$) deviates approximately by a 1.25A/frame root mean squared distance (RMSD) from the linearly interpolated (LI) trajectory. Figure 2.1 shows optimal coarse-grained trajectories for different values of H_S. The trajectories are projected onto a two-dimensional space with a multidimensional scaling technique (Cox and Cox 1994). As predicted by theory, the higher the value of H_S, the closer the LI path to the optimal trajectory. The potential energy profiles of the optimal trajectories as a function of their arc-lengths are shown in Figure 2.2. Figure 2.3 shows that the length of the optimal trajectory varies from 15.5 Å to 13 Å for different amounts of thermal energy (H_S) of the system.

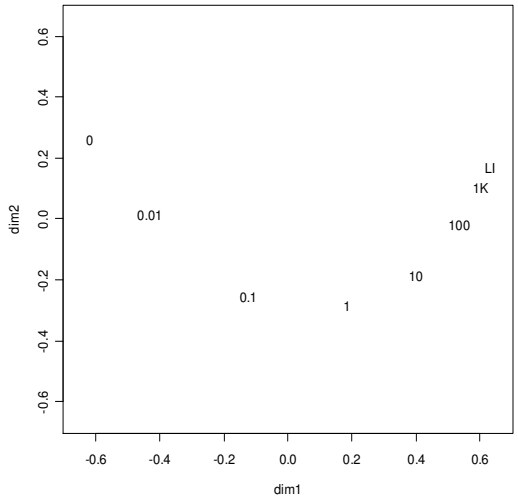


Figure 2.1: The distances between optimal coarse-grained trajectories for the transition of the extra-cellular component of mGluRI. The distances are projected onto a two dimensional space for better visualization. Each number in the plot corresponds to the trajectory with the given value of H_S in multiples of $K\kappa$ (K is the number of residues and κ is the force constant from the formula (2.9)). *LI* represents the linearly interpolated path between the two known conformations. The distance metric, upon which the projection is defined, is a sum of pairwise C_α-RMSD distances between corresponding path structures.

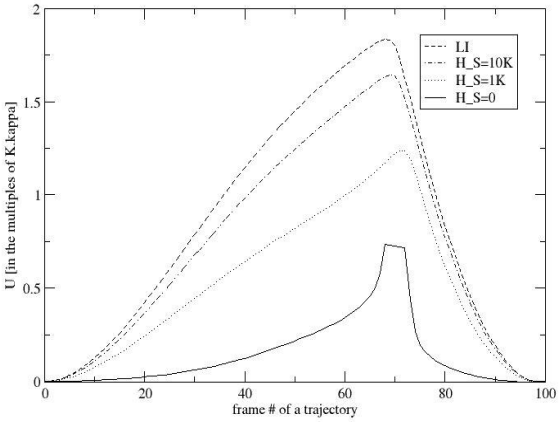


Figure 2.2: The potential energy profile of optimal Brownian trajectories of a coarse grained model for different values of H_S. The transition is of mGluRI. The energy-increasing curves correspond to the optimal trajectories with H_S equal to 0, 1K, 10K, and linear interpolation respectively. The potential and the values of H_S are in the multiples of $K\kappa$ (K is the number of residues and κ is the force constant from the formula (2.9)).

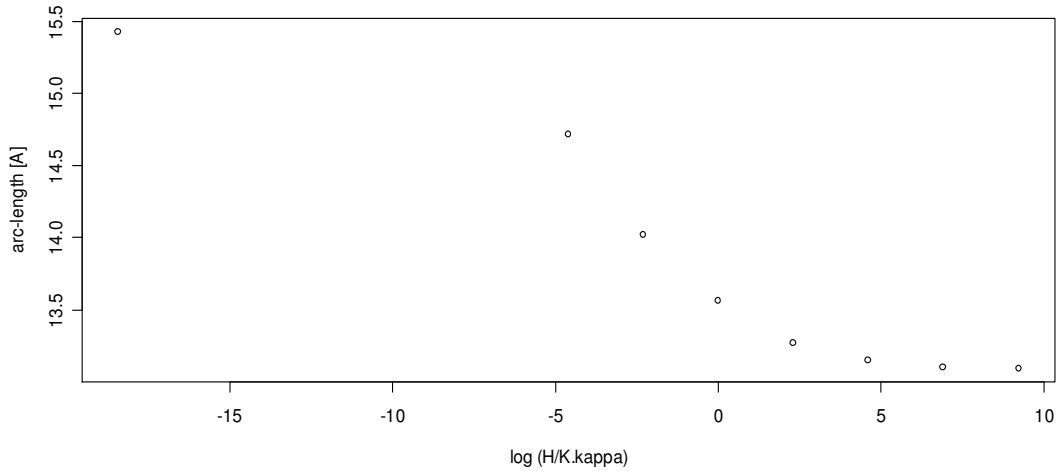


Figure 2.3: Arc-length of the optimal trajectory as a function of H_S. The RMSD between the two endpoints is approximately 13Å.

(iii) Once the optimal CG paths for different values of H_S are found, they are refined to atomically detailed trajectories and are locally minimized as discussed in Section 2.5. A three-dimensional projection (not shown) indicates that the atomically detailed refinement moves the trajectories in a direction perpendicular to the manifold defined by H_S (physically it suggests that the refinement focuses on the side chain positions, while the locations of the C_α atoms are not affected appreciably). To minimize S^l_{Gauss}, at least 10K steps of SDEL optimization are required.

The SDEL calculation takes approximately 100 hours of parallel computation on 100 CPUs, thus the SDEL part of the overall calculation is approximately 1000 times more expensive than the coarse-grained pre-processing part (nevertheless, we do believe that the resulting atomically detailed description of the system is important and worth the investment).

The most expensive part of the SDEL calculations for systems of this size in our code MOIL (Elber et al. 1995) is the Generalized Born implicit solvation energy. It takes approximately 95% of the SDEL's computation time. The complexity of this calculation is likely to be reduced in the future since there is significant room for

improving the GB implementation of MOIL - a project that we intend to pursue. The SDEL protocol tries to find a minimum in $3\times N\times K$ (for this system ≈ $3x10^6$) dimensional space, which also adds to the complexity of the calculations. Note, however, that the CG dimensionality is smaller by only a factor of 10-100, significantly smaller than the factor of 1000 mentioned above, between the calculations of the atomically detailed and coarse-grained models. The dominant factor in the latter is the much simpler (smoother) ANM potential compared to atomically detailed potentials.

Even when employing simulated annealing, the path search is in a neighborhood of the initial guess trajectory. The 10K SDEL steps modify the trajectory of mGluR transition by no more than 0.1Å/frame[1], compared to 1.25Å/frame obtained by the CG preprocessing which is clearly more significant.

Figure 2.4 shows the potential energy profiles of optimal paths selected by SDEL. Only profiles for trajectories that were optimized from the LI path and SDP ($H_S = 0$) are shown. The energy profiles of other trajectories refined by SDEL have comparable values. The SDEL adds considerable thermal kinetic energy to the SDP path, making the SDEL potential energies higher than the SDP potential energies, and the SDEL path more appropriate for describing the thermal processes. The energy barrier for a trajectory starting from the LI path is somewhat higher than the barrier obtained from a path derived from the SDP, suggesting an improvement in the SDEL trajectory produced the Steepest Descent Path CG trajectory.

[1] The 0.1Å RMSD per frame is based on C_α atoms only. The all atom RMSD is approximately twice as large.

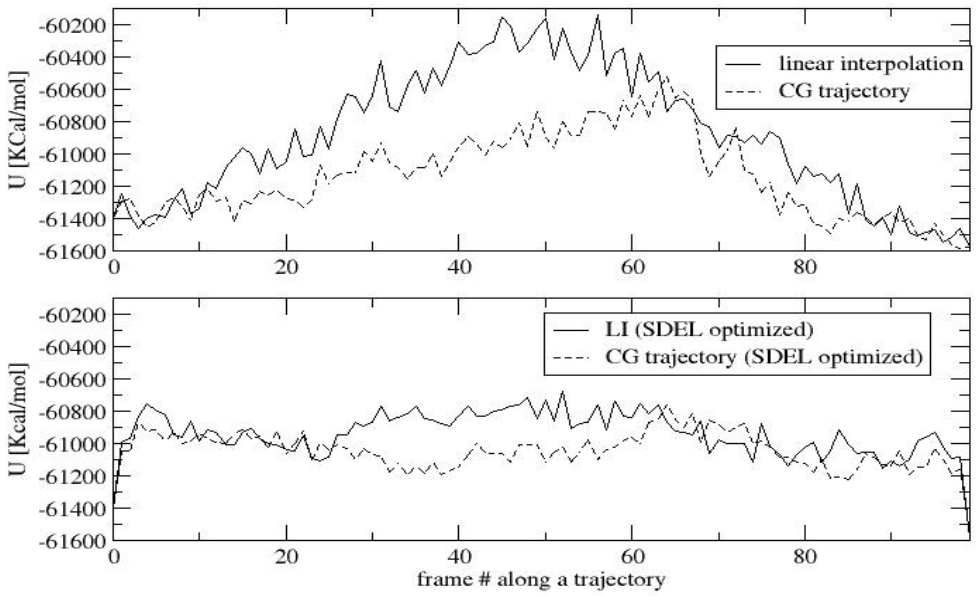


Figure 2.4: The energy profile of optimized SDEL trajectories for atomically detailed model of mGluR1. The horizontal axis shows frame index along the trajectory. The vertical axis is the potential energy. The top part shows potential energy profiles for trajectories from which SDEL optimization was started. In the bottom part the potential energy profiles after the SDEL optimization are shown. Solid lines correspond to trajectories optimized from the linear interpolation and dashed lines correspond to trajectories starting from the optimal CG trajectory for $H_S = 0$.

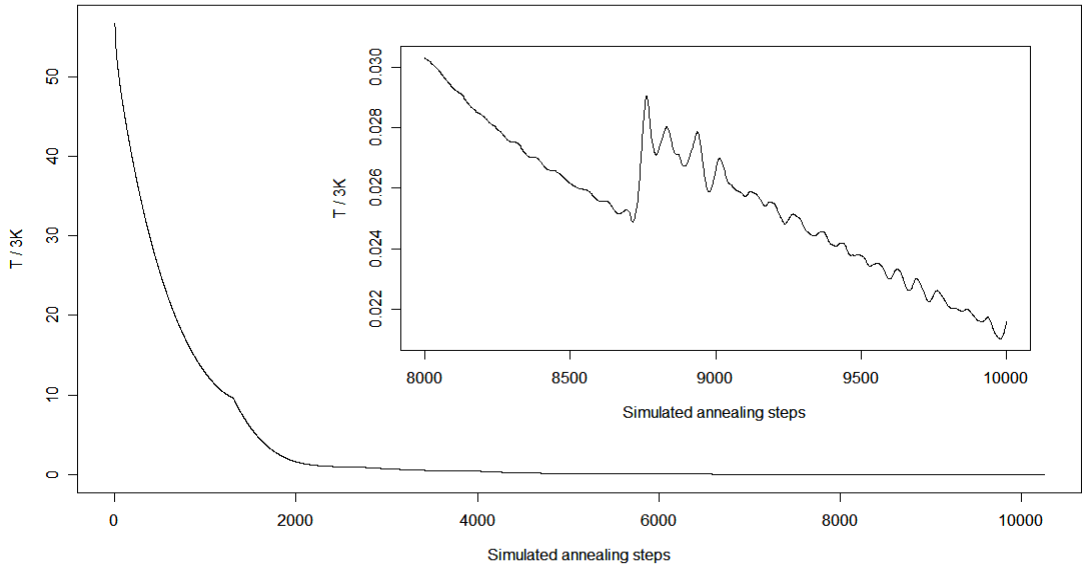


Figure 2.5: The simulated annealing profile in an SDEL minimization of the target function *T* as a function of the number of minimization steps. The example is for a CG trajectory with $H_S = 0$. In the right insert we expand the view of the last 2,000 steps. The decrease in the target function is rapid at the beginning, but in the last 2,000 steps the target function is decreased by only 33%.

Figure 2.5 shows the simulated annealing history of an SDEL run. The target function, T, is rapidly decreasing in the first thousand steps; however, further reduction with more steps is considerably slower. The RMSD changes of the trajectory in the last thousand of steps of the minimization are small (order of 10^{-2} Å/frame) and thus it might not be so important to locate the exact global minimum of T if the structural changes are of prime interest.

Any substantial differences in the inferences obtained from the SDP and LI based paths can be revealed by examining the contacts between the two monomers that substantially change during the transition. In Figure 2.6 we examine the distance between **GLU A 60 (atom OE2) and ARG B 448 (atom NH2)**. These two atoms are not in contact in either of the two end conformations, but are brought together during the transition in the optimal path based on SDP, but not in the path based on LI. Notably, the formation of an intermediate salt bridge may reduce the barrier height for the transition.

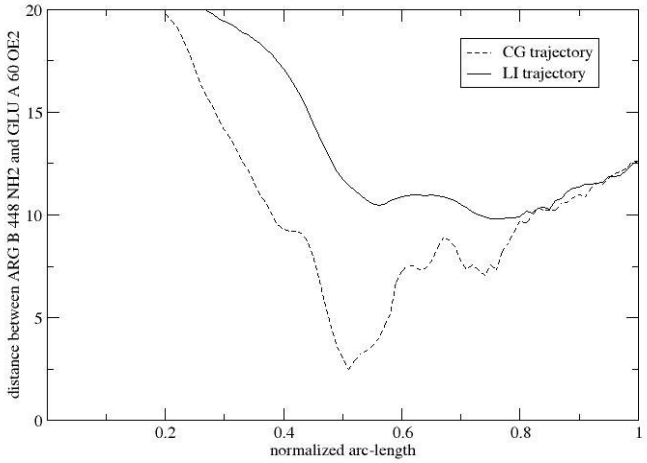


Figure 2.6: The distance between GLU A 60 (atom OE2) and ARG B 448 (atom NH2) during the transition between the inactive and active conformation of mGluR1.

This evolution is represented in the sequence of structures shown in Figure 2.7 that are taken from points along the path. This sequence illustrates the formation and the breaking of the salt bridge along the path.

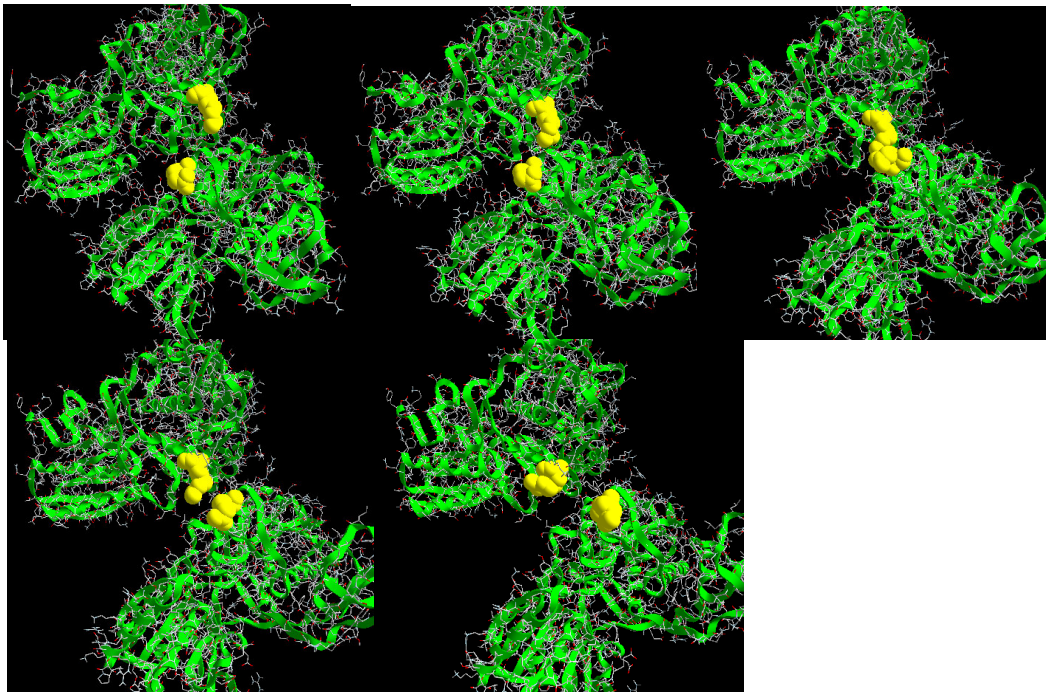


Figure 2.7: An illustration of the strong coupling between atomically detailed motion and large-scale domain opening. A sequence of events along the transitional pathway is shown starting from the upper left corner (structure 1) continuing to the right (structure 20) and then down. The length slices are shown from an atomically detailed path of 100 slices that was constructed from a coarse-grained model. Only slices 1,20,50,60 and 100 are shown. The atomically detailed event is the *transient formation of a salt bridge* between a gluamtic acid (Glu60 in chain A of the dimer) and an arginine (Arg448 in chain B) (yellow space filling model). There is also a large-scale motion that causes a visible separation between the two lobes. The salt bridge is not present at the end points. It assists in lowering the transition barrier. Notably, it is not present in the linearly interpolated path (see text for more details).

2.8 Conclusions

The first chapter was about spatial and temporal coarse-graining of pathways and trajectories of proteins. We have shown how the two coarsening strategies (temporal and spatial) can be applied together to obtain a qualitative computational description of large scale conformational transitions of biomolecular systems. The proposed method scales to systems of size of several hundreds to thousands aminoacids with large scale (12 Å RMSD) spatial rearrangements of structural domains.

The presented method was tested on a system of conformational transition of the extracellular domain of mGluR receptor upon ligand binding. The overall cost of coarse-grained part of the algorithm is negligible (about 0.1%) compared to all atomistic refinement. Moreover the algorithm in the coarse-grained mode is responsible for about 90% of adjustments to the resulting trajectories (as measured by RMSD).

There are however some limitations of the proposed method. One of them is that the computed results are only of qualitative nature with hard to interpret statistical properties of each resulting trajectory. Another issue is that the underlying coarse-grained force field might not be appropriate for large scale conformational rearrangements. We address both of these issues in the following chapters.

In Chapter 3, we discus an approach to systematically design coarse-grained potential consistent with experimental measurements (Xray structures) and in Chapter 4 we propose a way to calculate accurate quantitative analysis of a transition process. Results calculated in this chapter, a set of physically plausible trajectories, enters as input for quantitative calculations of thermodynamics and kinetics of the system.

REFERENCES

Alon, U. (2006). An introduction to systems biology: Design principles of biological circuits, CRC.

Aqvist, J. and A. Warshel (1993). "Simulation of enzyme reactions using valence bond force fields and other hybrid quantum/classical approaches." Chemical Reviews **93**: 2523-2544.

Bolhuis, P. G., D. Chandler, C. Dellago, and P. L. Geissler (2002). "Transition path sampling: Throwing ropes over rough mountain passes, in the dark." Annual Review of Physical Chemistry **53**: 291-318.

Briggs, W. L. (1987). A multigrid Tutorial. Philadelphia, SIAM.

Cardenas, A. E. and R. Elber (2003). "Kinetics of cytochrome C folding: Atomically detailed simulations." Proteins-Structure Function and Genetics **51**(2): 245-257.

Cox, T. F. and M. A. A. Cox (1994). Multidimensional scaling, Chapman Hill.

Czerminski, R. and R. Elber (1990). "Self avoiding walk between 2 fixed end points as a tool to calculate reaction paths in large molecular systems." International Journal of Quantum Chemistry: 167-186.

Dellago, C., P. G. Bolhuis, P. L. Geissler (2002). Transition path sampling. Advances in Chemical Physics, Vol 123. **123:** 1-78.

E, W. N., W. Q. Ren, and E. Vanden-Eijden (2002). "String method for the study of rare events." Physical Review B **66**(5).

Elber, R. (1990). "Calculation of the potential of mean force using molecular dynamics with linear constraints -An application to a conformational transition in a solvated dipeptide." Journal of Chemical Physics **93**(6): 4312-4321.

Elber, R. (2006). Calculations of classical trajectopries with boundary value formulation. Computer simulations in condensed matter: From materials to chemical biology. M. Ferrario, G. Ciccotti and K. Binder. Berlin, Springer. **704**.

Elber, R., A. Ghosh, and A. Cardenas (2002). "Long time dynamics of complex systems." Accounts of Chemical Research **35**(6): 396-403.

Elber, R. and M. Karplus (1987). "A method for determining reaction paths in large molecules - application to myoglobin." Chemical Physics Letters **139**(5): 375-380.

Elber, R., J. Meller, and R. Olender (1999). "Stochastic path approach to compute atomically detailed trajectories: Application to the folding of C peptide." Journal of Physical Chemistry B **103**(6): 899-911.

Elber, R., A. Roitberg, C. Simmerling, R. Goldstein, H. Y. Li, G. Verkhivker, C. Keasar, J. Zhang, and A. Utilsky (1995). "MOIL: A program for simulations of macrmolecules." Computer Physics Communications **91**(1-3): 159-189.

Elber, R. and D. Shalloway (2000). "Temperature dependent reaction coordinates." Journal of Chemical Physics **112**(13): 5539-5545.

Haliloglu, T., I. Bahar, et al. (1997). "Gaussian dynamics of folded proteins." Physical Review Letters **79**: 3090-3093.

Jonsson, H., G. Mills, and K. W. Jacobsen (1998). Nudged elastic band method for finding minimum energy paths of transitions. Classical and quantum dynamics in condensed phase simulations. B. J. Berne, G. Ciccotti and D. F. Coker. Singapore, World Scientific**:** 385-403.

Kunishima, N., Y. Shimada, Y. Tsuji, T. Sato, M. Yamamoto, T. Kumasaka, S. Nakanishi, H. Jingami, and K. Morikawa (2000). "Structural basis of glutamate

recognition by a dimeric metabotropic glutamate receptor." Nature **407**(6807): 971-977.

Lanczos, C. (1970). The variational principles of mechanics, University of Toronto press.

Landau, L. D. and E. M. Lifshitz (1976). Mechanics. Oxford, Pergamon

Lu, M. Y., B. Poon, and J. P. Ma (2006). "A new method for coarse-grained elastic normal-mode analysis." Journal of Chemical Theory and Computation **2**(3): 464-471.

Maragakis, P. and M. Karplus (2005). "Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase." Journal of Molecular Biology **352**(4): 807-822.

Muto, T., D. Tsuchiya, K. Morikawa, and H. Jingami (2007). "Structures of the extracellular regions of the group II/III metabotropic glutamate receptors." Proc Natl Acad Sci U S A **104**(10): 3759-64.

Olender, R. and R. Elber (1996). "Calculation of classical trajectories with a very large time step: Formalism and numerical examples." Journal of Chemical Physics **105**(20): 9299-9315.

Olender, R. and R. Elber (1997). "Yet another look at the steepest descent path." Theochem-Journal of Molecular Structure **398**: 63-71.

Peskin, C. S. and T. Schlick (1989). "Molecular dynamics by the backward-Euler method." Comm Pure Appl Math **42**: 1001-1031.

Pin, J. P. and F. Acher (2002). "The metabotropic glutamate receptors: structure, activation mechanism and pharmacology." Curr Drug Targets CNS Neurol Disord **1**(3): 297-317.

Pin, J. P., J. Kniazeff, C. Goudet, A. S. Bessis, J. Liu, T. Galvez, F. Acher, P. Rondard, and L. Prezeau (2004). "The activation mechanism of class-C G-protein coupled receptors." Biol Cell **96**(5): 335-42.

Pin, J. P., J. Kniazeff, J. Liu, V. Binet, C. Goudet, P. Rondard, and L. Prezeau (2005). "Allosteric functioning of dimeric class C G-protein-coupled receptors." Febs J **272**(12): 2947-55.

Ryckaert, J. P., G. Ciccotti, and H. J. C. Berendsen (1977). "Numerical integration of cartesian equations of m otion of a system with constraints - molecular dynamics of N-alkanes." Journal of Computational Physics **23**(3): 327-341.

Schlick, T., R. D. Skeel, A. T. Brunger, L. V. Kale, J. A. Board, J. Hermans, and K. Schulten (1999). "Algorithmic challenges in computational molecular biophysics." Journal of Computational Physics **151**(1): 9-48.

Tirion, M. M. (1996). "Large amplitude elastic motions in proteins from a single-parameter, atomic analysis." Physical Review Letters **77**(9): 1905-1908.

Ulitsky, A. and R. Elber (1990). "A new technique to calculate steepest descent paths in flexible polyatomic systems." Journal of Chemical Physics **92**(2): 1510-1511.

Ulitsky, A. and R. Elber (1993). "The thermal equilibrium aspects of the time-dependent hartree and the locally enhanced sampling approximations - formal proeprties, a correction, and computational examples for rare gas clusters." Journal of Chemical Physics **98**(4): 3380-3388.

Ulitsky, A. and R. Elber (1994). "Application of the locally enhanced sampling (LES) and a mean field witha binary collision correction (cLES) to the simulation of Ar diffusion and NO recombination in myoglobin." Journal of Physical Chemistry **98**(3): 1034-1043.

Verlet, L. (1967). "Computer "Experiments" on Classical Fluids. I. Thermodynamical Properties of Lennard-Jones Molecules." Physical Review **159**(1): 98-103.

Voth, G. A. ed. (2009). "Coarse-Graining of Condensed Phase and Biomolecular Systems" CRC Press,

Weinan, E., W. Q. Ren, and E. Vanden-Eijnden (2005). "Finite temperature string method for the study of rare events." Journal of Physical Chemistry B **109**(14): 6688-6693.

West, A. M. A., R. Elber, and D. Shalloway (2007). "Extending molecular dynamics time scales with milestoning: Example of complex kinetics in a solvated peptide." Journal of Chemical Physics **126**(14): 145104-14.

West, A. M. A., R. Elber (2010). "Atomically detailed simulation of the recovery stroke in myosin by Milestoning." Proceedings of the National Academy of Sciences of the United States of America **107**(11): 5001-5005.

Xu, C. Y., D. Tobi, and I. Bahar (2003). "Allosteric changes in protein structure computed by a simple mechanical model: Hemoglobin T <-> R2 transition." Journal of Molecular Biology **333**(1): 153-168.

Yang, L. W., E. Eyal, C. Chennubhotla, J. Jee, A. M. Gronenborn, and I. Bahar (2007). "Insights into equilibrium dynamics of proteins from comparison of NMR and X-ray data with computational predictions." Structure **15**(6): 741-749.

Zaloj, V. and R. Elber (2000). "Parallel computations of molecular dynamics trajectories using the stochastic path approach." Computer Physics Communications **128**(1-2): 118-127.

CHAPTER 3

A COARSE-GRAINED POTENTIAL FOR FOLD RECOGNITION AND MOLECULAR DYNAMICS SIMULATIONS OF PROTEINS

3.1 Introduction

Hierarchical description of complex systems motivates the creation of coarse grained or reduced models with two goals in mind: (i) capture essential features of the system with simplified models that can be solved exactly (or almost exactly), and (ii) describe quantitatively properties of complex systems with a reduced representation computed from detailed experiment or theory. Examples for coarse grained models of type (i) are the HP model on a square lattice (Dill 1985), or the Elastic Network Model for protein flexibility (Tirion 1996; Haliloglu, Bahar et al. 1997). Examples for type (ii) models are detailed folding simulations on lattices (Kolinski and Skolnick 1996), or coarse description of membranes (Marrink, Risselada et al. 2007). Approaches of type (ii) attempt to significantly reduce the computational cost and at the same time maintain a high level of accuracy that approaches the results of more detailed models.

The potential we describe in here belongs to class (ii). Our aim was to develop an empirical force field with a reduced set of variables for physical simulations of proteins in the neighborhood of the native states. Simulations at the coarse level can be done more efficiently than atomically detailed calculations. Indeed, we illustrate here test simulations with accumulated time length of tens of microseconds that require only 12 hours on 500 computer cores. A nanosecond simulation of a medium size solvated protein (200 amino acids) can take a few days. The computational saving for simulations is about 3 orders of magnitude. We expect that equilibrium distributions generated by simulations with the designed potential will show characteristics of

atomically detailed simulations. In parallel we require that the potential will recognize native folds of proteins as the lowest energy minimum when compared with an extensive set of "decoy" structures.

Our potential is purely empirical and the experimental observables which we use to fit the potential parameters are native structures of proteins determined by experimental techniques and deposited in the Protein Data Bank (PDB) (Berman, Westbrook et al. 2000). These observables are clearly incomplete and a correct energy function should reproduce also the thermodynamics and kinetics of the system.

In the last twenty years many energy functions were estimated from empirical structures of proteins using the methodologies initiated by the following studies: inverse Boltzmann formula (statistical potentials) (Miyazawa and Jernigan 1985), memory associated Hamiltonians (Goldstein, Lutheyschulten et al. 1992), Z score optimization (Luthy, Bowie et al. 1992), and Mathematical Programming (Maiorov and Crippen 1992). Learning potentials from empirical structures should be contrasted with physically based energy functions. The usual design of a physical energy relies on experiments (and/or ab-initio calculations) on small model systems (Rizzo and Jorgensen 1999; Wang and Kollman 2001; Lagant, Nolde et al. 2004). From a learning view-point, an advantage of physical potentials is the separation of types of input (the data to learn)) from types of output (the data to predict). On the other hand, potentials that are learned from empirical structures recognize correct folds with significantly less computational resources compared to physical energies, allowing for more extensive exploration of conformation space. The number of degrees of freedom is smaller by a factor between three and ten even without explicit solvent.

The approach described in this chapter is an extension of the usual implementation of statistical potentials. We therefore start with a brief discussion of statistical potentials. After the introduction of statistical potentials by Miyazawa and

Jernigan (1985), a number of groups, including for example, Sippl (1990), Skolnick et al. (1997), Betancourt and Thirumalai (1999), Bryant and Lawrence (1993), Hinds and Levitt (1994) and others more recently (Xia, Huang et al. 2000; Lu and Skolnick 2001; Buchete, Straub et al. 2003; Lu, Dousis et al. 2008) continue to develop this concept and to examine the basic algorithm, functional form, and the data sets.

The basic concept of statistical potentials is similar in spirit to that of the potential of mean force (Hill 1956) but important differences remain. Let the complete coordinate vector in continuous space representing the system be X, and the subset of coordinates that we use to describe the protein be $y_{i=1,\ldots,n}$, for example the set of backbone torsions or distances between amino acids. The number of reduced degrees of freedom is n, while the number of total number of degrees of freedom in the system is N. If the probability of a conformation, $p(X)$, is known we can determine the probability of a variable of interest, y_i, by direct integration $p(y_i) = \int P(X)\delta(y_i - \overline{y}_i(X))dX$. The delta function matches the value y_i with the function of the canonical coordinates $\overline{y}_i(X)$. If the probability $P(X)$ obeys Boltzmann statistics ($P(X) \propto \exp(-\beta U(X))$, $U(X)$ is the potential energy, and β is the inverse temperature) then the probability $p(y_i)$ is related to a potential of mean force (PMF), $V_i(y_i) = -(1/\beta)\log(p(y_i))$.

The first assumption made in the derivation of Statistical Potentials (SP) is that the Protein Data Bank (PDB) provides a Boltzmann sample of conformations, therefore a PMF can be estimated from the observed frequencies of certain degrees of freedom $V_i(y_i) = -(1/\beta)\log(f(y_i))$ (Miyazawa and Jernigan 1985).

The second assumption made in the calculations of SP is the representation of the total potential as a sum of PMF terms. An "energy" of the system is written as $U(y_1, y_2, \ldots, y_n) = V_1(y_1) + V_2(y_2) + \ldots + V_n(y_n)$.

The problem with this assumption is easy to illustrate using the definition of the PMF. The "energy" in the subspace of $y_{i=1,\ldots,n}$ is used to sample conformations in the full coordinate space of the protein X. The sampling is in the canonical ensemble with β for inverse temperature and for all degrees of freedom X:

$$p(y_i) = \int \exp\left[-\beta\left(V_1(y_1)+\ldots+V_i(y_i)+\ldots+V_n(y_n)\right)\right] \cdot \delta\left(y_i - \bar{y}_i(X)\right) J(Y,X)\, d\Gamma \prod_j dy_j ,$$

where we plugged in the integral the usual form of the statistical potential, $d\Gamma$ is a volume element of the remaining coordinates not in y_i's, and $J(Y,X)$ is the Jacobian of the transformation from X to Y. Note that X and Y are not of the same dimension and Γ denotes the remaining degrees of freedom.

Instead of the statistical potential we can write a new effective energy that is used in the sampling $V_{eff}(y_1,\ldots,y_n) = \sum_i V_i(y_i) - (1/\beta)\log\left(J(X,Y)\right)$. If the Jacobian was a constant then we would trivially recover the probability $p(y_i) \propto \exp\left(-\beta V(y_i)\right)$ that we started with. However, for most degrees of freedom used in statistical potentials (e.g. distances) this is not the case. We can still seek an effective potential $V_i^*(y_i)$ that will make the desired definition of the mean force potential to hold, i.e., $\bar{p}(y_i) \propto \int \exp\left(-\beta \sum_i V_i^*(y_i) + \log\left(J(X,Y)\right)\right) \delta\left(y_i - \bar{y}_i(X)\right) d\Gamma \prod_j dy_j$ and at the same time $\bar{p}(y_i)$ is equal to the PDB distribution $p(y_i)$. A statistical potential used "as is" will not reproduce the PDB distribution if it is implemented in an algorithm that generates the canonical distribution. Note that the potential $V^*(y_i)$ and the distribution $p(y_i)$ are no longer related by the inverse Boltzmann relation. The algorithm proposed in this chapter attempts to generate such a $V^*(y_i)$.

Besides the basic difference between PMF and SP pointed above, writing the overall potential as a sum of PMFs introduces additional approximations. The first is the factorization of the overall probability to a product of probabilities. It suggests lack of correlations between the y_i's. The use of multiple internal coordinate probabilities

(Buchete, Straub et al. 2004; Ngan, Inouye et al. 2006; Feng, Kloczkowski et al. 2007) $p(y_i y_j)$ addresses some of the concerns. However, the choice of correlations to focus on is not trivial and acquiring appropriate statistics for these higher order interaction terms is another challenge. The second approximation is the use of types. It is not obvious that probability distribution of type α (e.g., a contact between phenylalanine and valine) will be the same in a different environment (e.g., hydrophobic or polar medium).

SP most frequently aim at the fold recognition problem; i.e., given a set of plausible structures that are all protein-like, how to choose a configuration that is the closest to the native fold. It typically does not address the problem of direct and extensive sampling of configuration space with a potential according to a pre-determined weight (e.g. canonical). We generate a potential that is consistent with both (MD simulations and fold recognition). Not surprisingly new problems emerge. One practical problem is that the sampling of coordinate space in the PDB is incomplete. As a result MD simulations with straightforward statistical potentials do not produce protein-like conformations.

The problem of generating a single potential, which is optimal for the task of fold recognition and of MD simulations, can be solved by additional potential terms that take care of interactions poorly sampled in the PDB. The combination of the statistical potential and the new terms is not obvious. Once these terms are added to "traditional" statistical potentials the simulations with the adjusted energy function no longer (necessarily) reproduce the distributions of the y_i's extracted from the PDB. We address this particular problem by adopting an algorithm from condensed phase simulations which is a variant of the generalized ensemble approach (Kinnear, Jarrold et al. 2004). It generates iteratively a potential consistent with the PDB distributions of internal coordinates and the supplements discussed above.

The resulting potential is significantly more complex than the usual form of statistical potentials. It is also continuous and differentiable. We emphasize that even with these advances we do not address the two basic approximations of statistical potentials (factorization of the probability and transferability of parameters). It is therefore not surprising that significant deviations from native folds are still observed in simulations for a significant number of proteins, even if the design requirements are satisfied. Despite the drawbacks, the performance we obtain with the final form of the potential is adequate for the usual fold recognition (and it was used in CASP8 http://predictioncenter.org/casp8/index.cgi), and also for Molecular Dynamics simulations. Another continuous and differentiable potential that learns its parameters from the PDB with a different technique and can be used for energy minimization and simulations was introduced recently (Amir, Kalisman et al. 2008). Bridging potential parameters from small molecule data to macromolecular modeling was also pursued recently by Z score optimization (Jagielska, Wroblewska et al. 2008). These potentials are however designed for all atom models.

3.2 Potential functional form

In this section we present the functional form and the parameterization of a coarse grained potential which we call FREADY (a potential for Fold REcognition And DYnamics). The starting functional form and parameterization of the potential were motivated by the simple physical model of the group of Thirumalai (J. D. Honeycutt and Thirumalai 1992) and its enhancements by the group of Head-Gordon (Brown and Head-Gordon 2004; Yap, Fawzi et al. 2008). However, as we look in more detail into the conformation data available in the Protein Data Bank and examine structures generated by Molecular Dynamics (MD) simulations (using coarse grained potentials), a significantly more complex form becomes necessary.

The number of degrees of freedom in the complex form remains relatively small, only two points per amino acid are used - the position of the Cα atom and the side chain center of mass (CM). It was also decided to keep the functional form independent of any information about the native structure (e.g. secondary structure or native contacts); thus enabling unbiased dynamical studies of biophysical processes where the information about the native conformation is not available or well defined (e.g. large conformational transitions).

The potential employs the functional form (3.1) that includes bond, angular, and torsional terms as well as non-bonded interaction and explicit hydrogen bonding. Solvent is treated implicitly since the parameters of the potential are learnt from statistics of solvated protein. By insisting that solvent induced structures (most structures in the PDB are reasonably well solvated) are reproduced in the simulations we incorporate some solvent effects.

$$U(X)=\sum_{i\in bonds}U_B(r_i,\tau_{Bi})+\sum_{i\in angles}U_A(\theta_i,\tau_{Ai})+\sum_{i\in torsions}U_T(\phi_i,\tau_{\phi i})+\sum_{i,j>i}U_{NB}(r_{ij},\tau_i,\tau_j)+\sum_{i\in dipole\ centers}U_{HB}(i) \tag{3.1}$$

We denote by τ the type of interactions (for example atom type, or the type of a bond between two atoms). Typically, bond and angle interactions in other force fields (atomic or coarse-grained) are modeled by quadratic terms with a single minimum; however these functions do not give acceptable fits to the statistics of bond lengths and angles we extract from the PDB structures (Figure 3.1) and later from MD. The reason is that the internal degrees of freedom of side chains and backbone that are removed in the coarse representation have internal structure with multiple stable states that is reflected in multiple minima of the coarse variables. This observation is especially true for covalent terms that include a side chain atom but is also correct for angles of three sequential backbone atoms (Cα). Therefore, the bond energy as well as the angle energy terms of FREADY, are described with a single, a double, or a triple well potential (see Eq. (3.2) and (3.3)). The multiple well potentials we consider in this work are

$$U_{B/A}(x,\tau)=\begin{cases}k_\tau(x-x_\tau)^2 & \text{if } \tau\in \text{ terms with a single well}\\ C\left(k_{\tau1}(x-x_{\tau1})^2,k_{\tau2}(x-x_{\tau2})^2+\alpha_\tau,\beta_\tau\right) & \text{if } \tau\in \text{ terms with a double well}\\ C\left\{C\left(k_{\tau1}(x-x_{\tau1})^2,k_{\tau2}(x-x_{\tau2})^2+\alpha_\tau,\beta_\tau\right),k_{\tau3}(x-x_{\tau3})^2+\alpha'_\tau,\beta'_\tau\right\} & \text{if } \tau\in \text{ terms with a triple well}\end{cases} \tag{3.2}$$

$$C(U_1,U_2,\beta)=\frac{1}{2}\left(U_1+U_2-\sqrt{(U_1-U_2)^2+\beta^2}\right), \tag{3.3}$$

where x denotes a bond length or an angle size and all variables with τ in the subscript are potential parameters to be determined. The parameters x_τ are equilibrium positions, k_τ are force constants, α_τ are relative energy differences between the different minima, and β_τ are determining the barrier height between two wells. The coupling function $C\left(U_1,U_2,\beta\right)$ joins the two energy functions U_1 and U_2

as in empirical valence bond theory (Aaqvist and Warshel 1993), a form that was used in another coarse-grained model (Maragakis and Karplus 2005; Okazaki, Koga et al. 2006). Triple well terms require multiple parameters α and β.

The current model has 22 different types of bonds and 58 different types of angles. There are 19 different bonds between Cα and CM particles for each of the different amino acid (GLY does not have a CM particle), one bond type for the typical Cα-Cα backbone bond, one for a bond between Cα of a proline in a cis-isomer and a preceding Cα atom. The last bond type is for modeling the disulfide bridges between cysteine residues.

The 58 angle types are built from the following three templates $C\alpha_{i-1}$-$C\alpha_i$-$C\alpha_{i+1}$, CM_i-$C\alpha_i$-$C\alpha_{i-1}$, and CM_i-$C\alpha_i$-$C\alpha_{i+1}$ for each different type of a central ($C\alpha_i$) atom with the exception of GLY. The 20 types of angle templates $C\alpha_{i-1}$-$C\alpha_i$-$C\alpha_{i+1}$ are all very similar and could be reduced to a single backbone angle type. Since subtle differences may have remained we did not merge all these terms in the first version of the potential.

The torsional terms $U_T\left(\phi,\tau_\phi\right)$ take as input an angle ϕ and a type of the torsional angle τ_ϕ. The torsional term is modeled as the following sum of cosine and sine terms:

$$U_T\left(\phi,\tau_\phi\right)=\sum_{n=1}^{5} C_{\tau,n}\cos(n\phi)+S_{\tau,n}\sin(n\phi) \tag{3.4}$$

We have used five expansion terms for the periodic function. This number of terms is probably unnecessary, however, in the present version of the potential they do not harm. It is still possible that subtle effects are captured by the high order terms and therefore we left these terms "as are" and did not attempt to simplify them further. There are almost $4\cdot 20^2$ different types of torsional/dihedral angles: A torsion (the angle between two planes) is defined by four points. All torsions in our model are

along $C\alpha_i$-$C\alpha_{i+1}$ backbone bonds (we do not consider torsions related to CYS-CYS bonds). The type of a torsional interaction, τ_ϕ, is determined by the residue types of the central Cα pair and by the particle types (Cα or CM) of the two remaining points. For a given Cα pair there can be up to four different dihedral angles present ($C\alpha_{i-1}$-$C\alpha_i$-$C\alpha_{i+1}$-$C\alpha_{i+2}$, CM_i-$C\alpha_i$-$C\alpha_{i+1}$-$C\alpha_{i+2}$, $C\alpha_{i-1}$-$C\alpha_i$-$C\alpha_{i+1}$-CM_{i+1}, and CM_i-$C\alpha_i$-$C\alpha_{i+1}$-CM_{i+1}). The number of different torsional types is not exactly $4 \cdot 20^2$ since glycine does not have a side chain.

The function $U_{NB}(r, \tau_1, \tau_2)$, describes non-bonded interactions where τ_1, τ_2 are the types of the interacting particles and r is the distance between them. There are 39 different particles considered for non-bonded interactions (20 Cα atoms and 19 CM particles). Thus we have $39 \cdot 40/2$ types of non-bonded interactions in the system. The function $U_{NB}(r, \tau_1, \tau_2)$ is continuous and differentiable to the first order and is defined below.

$$U_{NB}(r,\tau_1,\tau_2) = \begin{cases} U^0_{NB}(r) + A_{\tau_1\tau_2} r^{-6} + B_{\tau_1\tau_2} r^{-2} + C_{\tau_1\tau_2} & \text{if } r < 4.2 \text{ Å} \\ U^0_{NB}(r) + \sum_{i=0}^{9} a_{\tau_1\tau_2 i} r^i & \text{if } r \in \langle 4.2 \text{ Å}, 13.5 \text{ Å} \rangle \\ 0 & \text{if } r > 13.5 \text{ Å} \end{cases} \quad (3.5)$$

$$U^0_{NB}(r) = \begin{cases} 0.6 \cdot 10^7 r^{-12} - 3 \cdot 10^3 r^{-6} & \text{between CM-CM particles} \\ 0.6 \cdot 10^6 r^{-12} & \text{otherwise} \end{cases} \quad (3.6)$$

We do not consider a pair of particles for non-bonded interactions if they are separated by one or two bonds; if they are separated by three bonds (1-4 interaction) we scale the non-bonded interaction down by a factor f_{14}. S-S bonds between CYS residues are not considered for these exceptions. If a scaling factor $f_{14} = 1$ is used the non-bonded energy distorts the local geometry when CM_i and CM_{i+1} are a strongly repulsive pair. At the other limit, if $f_{14} = 0$, some pairs of neighboring sidechains may

overlap. The value of f_{14} was set to 0.3 after some experimentation and was found to reproduce well the local structure.

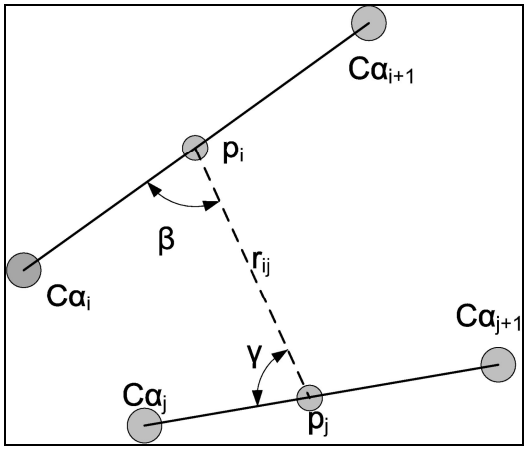


Figure 3.1: Description of terms entering the calculation of the backbone hydrogen bonding term $U_{HB}(i,j)$. The angle α_{ij} is defined as an angle between the bonds $C_{\alpha i} - C_{\alpha i+1}$ and $C_{\alpha j} - C_{\alpha j+1}$.

Backbone hydrogen bonding potential between residues *i* and *j*, $U_{HB}(i,j)$, is based on the model developed by Liwo and coworkers (Liwo, Pincus et al. 1993; Liwo, Oldziej et al. 2004). These hydrogen bonds are modeled by dipole interactions between the peptide centers which are implicitly assumed to be located in the centers of Cα-Cα bonds. The explicit functional form of $U_{HB}(i,j)$ is given below

$$U_{HB}(i,j) = \frac{A_{\tau_i\tau_j}}{r_{ij}^3} f_{ij} - \frac{B_{\tau_i\tau_j}}{r_{ij}^6}\left[4 + f_{ij}^{\,2} - g_{ij}\right] + \varepsilon_{\tau_i\tau_j}\left[q_{ij}^{12} - 2q_{ij}^6\right]$$
$$q_{ij} = \frac{r^0_{\tau_i\tau_j}}{r_{ij}} \qquad f_{ij} = \cos\alpha_{ij} - 3\cos\beta_{ij}\cos\gamma_{ij} \qquad g_{ij} = 3(\cos^2\beta_{ij} + \cos^2\gamma_{ij}) \qquad (3.7)$$

where r_{ij}, α_{ij}, β_{ij}, and γ_{ij} are the coordinates that determine the geometry of a hydrogen bond (Figure 3.1). There are two types of peptide centers $(\tau_i \in \{1,2\})$

defined in this work similarly to reference (Liwo, Pincus et al. 1993): a usual peptide bond and a proline-type peptide bond. The interaction parameters to be determined are $r^0_{\tau_i\tau_j}$, $A_{\tau_i\tau_j}$, $B_{\tau_i\tau_j}$, and $\varepsilon_{\tau_i\tau_j}$. Eq. (3.7) is derived in (Liwo, Pincus et al. 1993) by Boltzmann averaging over torsional degrees of freedom of the two interacting dipoles. Our initial attempt to model backbone hydrogen bonding by $\sum_{i,j} U_{HB}(i,j)$ follows UNRES (Liwo, Pincus et al. 1993; Liwo, Oldziej et al. 2004). However, with other terms at hand, simulations with the UNRES potential generate conformations that are often too compact and contain unnatural hydrogen bonding patterns. Another observation was that typically each residue contributed to the sum $\sum_{i,j} U_{HB}(i,j)$ by 1 to 5 partners. Five hydrogen bonds per residue are too many compared to the typical saturation number of about two that we observed in the PDB. To reduce over bonding of the hydrogen bonds within the context of FREADY potential, we retain at most the two strongest interactions described by Eq. (3.7) per amino acid. The hydrogen bond energy of a site i is determined as follows. The energies of all the candidates j for a hydrogen bond with i, $U_{HB}(i,j)$, are sorted and the lowest energy, $Hb_{ij}^{(\min)}$ is kept. We then examine the possibility of having two (lowest energy) hydrogen bonds to the site i. The energy of the two hydrogen bonds depends on their relative orientation ϕ_{jik}, $Hb_{ijk}^{(\min)} = \min\left(U_{HB}(i,j)+U_{HB}(i,k)\right)\cdot F\left(-\cos\left(\phi_{jik}\right)\right)$, where ϕ_{jik} is the angle between the dipole centers *j*, *i*, and *k*.

$$F(x)=\begin{cases} 1 & \text{if } x>0.9 \\ (x-0.3)/0.6 & \text{if } x\in[0.3,0.9] \\ 0 & \text{if } x<0.3 \end{cases}$$

The optimal single bond energy is then compared to the optimal two-hydrogen-bond energy and the option with the lowest energy is used

$$U_{HB}(i)=\min\left[Hb_{ij}^{(\min)}, Hb_{ijk}^{(\min)}\right]. \qquad (3.8)$$

3.3 Learning the potential parameters

As discussed in the introduction to this chapter the most common approach to derive parameters of a statistical potential is based on the assumption of mutual independence of different interactions in the protein. Based on statistics collected from experimental structures the potential function along a degree of freedom q is obtained by Boltzmann inversion formula

$$U(q) = -k_B T \ln\left(\frac{P_{native}(q)}{P_{reference}(q)} \right), \tag{3.9}$$

where k_B is the Boltzmann constant, T is the temperature (300 K), and $P_{native}(q)$, $P_{reference}(q)$ are probability distributions of a variable q in the experimentally solved dataset and an expected probability distribution of q (also known as the reference state). Examples for reference states are (i) a state of no interactions between amino acids (unfolded protein), and (ii) a state of random interaction between the amino acids. A proper choice of the reference state was a topic of much discussion in the literature (Betancourt and Thirumalai 1999; Zhou and Zhou 2002). The complete potential for a particular protein is given by a sum of $U(q)$ terms: $\bar{U}_{total}(q_1,...,q_l) = \sum_{i=1}^{l} U(q_i)$. This functional form assumes that the total probability of finding these variables factors into a product of probabilities of individual terms.

We bridge the learning of potentials for fold recognition and potentials for Molecular Dynamics simulations by an iterative procedure to recover the native distributions of relevant degrees of freedom $P_{native}(q_j)$, where j is an index that goes through types considered in Eq. (3.1) (e.g. distance between Cα particles of ALA and THR residues). Before the first iteration, the training set of native structures is used to calculate $P_{native}(q_j)$ and a zero-order potential $\bar{U}_0(q_1,\ldots,q_l)$ is chosen. The particular choice of $\bar{U}_0(\mathbf{q})$ is not important and any reasonable initial guess is corrected in the following learning iterations. The potential $\bar{U}_i(\mathbf{q})$ is then used to initiate long

Molecular Dynamics trajectories in the CG model producing canonical distribution of structures at room temperature (300 K) consistent with $\bar{U}_i(\mathbf{q})$. These simulations are run for 600 picoseconds (with a time step of 3 fs) and for all proteins (4867) in the training set. Probability distributions $P_i(q_j)$ of bond lengths, angles, torsions, pairwise particle distances and hydrogen bond lengths are collected from the final structures of simulated trajectories. However, as discussed in the introduction section, canonical sampling with statistical potentials does not reproduce the PDB distributions because of the Jacobian coupling. An attempt to fix this problem is to consider the ratio of the sampled and of the native distributions. The logarithm of the ratio of these probabilities will be added to the potential to initiate a new iteration (new Molecular Dynamics trajectories with the fixed potential). The formula for the adjustment (following Reith and co-workers (Reith, Pütz et al. 2003) and (Sun, Ghosh et al. 2008)) is

$$U_{i+1}(q) = U_i(q) + k_B T \ln\left(\frac{P_i(q)}{P_{native}(q)}\right). \tag{3.10}$$

We reiterated the calculations of the potential and Molecular Dynamics simulations a number of times until the correction to the potential parameters was negligible, in practice this happens in about 20 iterations. It is similar in spirit to a generalized ensemble approach that was used extensively by others (see for instance (Hansmann, Okamoto et al. 1996)). Reith and co-workers proposed this procedure to derive coarse grained potentials for polymers. Atomically detailed simulations were used in their work to define $P_{native}(q_j)$. Instead of running expensive all-atom MD simulations on the whole training set we infer $P_{native}(q_j)$ from the structures deposited in PDB.

It is important to emphasize the difference of equation (3.10) from the usual statistical potential approach (Miyazawa and Jernigan 1985) which is a one step

calculation from probability to potential. The iterative form of equation (3.10) allows us to add external terms (external to the probabilities determined from the PDB) and use the iterations to merge the different terms such that the original probabilities will be recovered in the canonical sampling. Such a potential refinement scheme is new and is not part of the "traditional" statistical potential approach. The final distributions $P(q_j)$ that we obtain are not identically equal to the native PDB distributions. However, the deviations are within the usual statistical errors of this type of calculation (Figures 3.2 and 3.4) and are due to the discrete representation of the distributions and the finite size of the training set.

Nevertheless, one must keep in mind that even with the iterations the potential is approximate. First (as discussed above) the factorization is an approximate procedure and only a general $P\left(q_1,...,q_l\right)$ is exact. Second, it is assumed that the potential is transferable, i.e. that we can have one coarse-grained potential to describe many proteins. Third, we assume that the iterative process of running Molecular Dynamics trajectories and adjusting the potential as described above converges to a stable solution (there is no proof of convergence). With the above mentioned approximations, it is perhaps no surprise that the procedure we finally adopt to compute all the potential parameters involved considerable heuristic, and that the resulting potential is not perfect: (i) it does not recognize native folds as the lowest energy in all cases, and (ii) MD simulations sampled with significant probability (for some proteins) structures that are far from the native fold.

As a training set, we used a set of PDB protein structures that forms the prediction database for our modeling program LOOPP (http://www.loopp.org, for a recent publication see (Vallat, Pillardy et al. 2008)). It includes 9513 native structures that have at most 70% sequence identity between any two proteins in the set. This is a higher sequence similarity than the similarity used in other studies of statistical

potential (about 20%). Our data provides reasonably dense sampling in sequence space. At least for fold recognition (after all, we wish to predict protein structure from a sequence) we argue that folds with larger sequence capacity (the number of sequences that are compatible with a given fold (Meyerguz, Grasso et al. 2004)) should have a higher weight than folds that capture only a few sequences. This weight might be lost if the selection emphasizes structural diversity instead of sequence variations. Another (pragmatic) reason that led us to broaden the set of structures and sequences is that of statistics. We need more proteins in order to obtain reliable statistics to fit our complex differentiable interaction terms (e.g. we need to sample at least 100 times every pair of neighboring residues along the backbone to fit reliably each torsional interaction).

The training set is further refined by removing membrane proteins (Jayasinghe, Hristova et al. 2001; Tusnady, Dosztanyi et al. 2004) and proteins complexed with polynucleotides (Spirin, Titov et al. 2007). All occurrences of selenomethionines (MSE) were replaced by regular MET residues and pyroglutamic acids (PCAs) were removed from the C-terminals. Proteins that contain other non-standard amino acids were removed from the training set. We used structures that correspond to the biological molecules (remarks BIOMT 350 in the PDB files) rather than the units determined by crystallography. In the training process we limited ourselves to globular proteins, therefore proteins with radius of gyration 15% larger than expected were not considered. The formula for expected radius of gyration of globular proteins $Rg = 0.395N^{3/5} + 7.257$ was taken from (Narang, Bhushan et al. 2005; Jayaram, Bhushan et al. 2006). Lastly, since MD simulations for larger proteins take longer time only proteins with at most 750 residues are used in the training process. The final training set contains 4867 proteins. All MD simulations were performed in the MOIL molecular modeling package (Elber, Roitberg et al. 1995)

(http://clsb.ices.utexas.edu/prebuilt/) and the final version of FREADY is fully integrated with other functionalities of the package such as energy minimization or visualization. MD calculations conducted with FREADY potential are about 10^3 faster than an all-atom simulation in explicit solvation. The converged set of FREADY potential parameters can be found in the file moil.mop/CG.PROP of the MOIL distribution package or is also available in an extended form in the tar file http://clsb.ices.utexas.edu/research/group/fready.tgz.

In practice, distributions $P_i(q)$ and $P_{native}(q)$ are represented as discrete sets of bins. Bin sizes used in this work are 0.1 Å, 1°, 3°, 0.3 Å, and 0.1 Å for bond, angle, torsion, non-bonded, and hydrogen-bonding terms respectively. The discrete descriptions of $U_{i+1}(q)$ are then fitted by continuous functions described in Eq. (3.2) - (3.8). Fitting of bond and angle parameters has been performed manually, since the convergence is reached after one or two iterations. Torsional terms are fitted in a straightforward manner by the Discrete Fourier Transform algorithm.

The parameters $A_{\tau_i\tau_j}$, $B_{\tau_i\tau_j}$, and $\varepsilon_{\tau_i\tau_j}$ of the backbone hydrogen bonding term $U_{HB}(i,j)$ are not optimized independently in this work, but their ratios are taken from (Liwo, Pincus et al. 1993) where they were optimized by fitting restricted free energy surfaces of UNRES model to those obtained from all atom simulations. Only the overall multiplicative factor of these energy constants and the parameters $r^0_{\tau_i\tau_j}$ are optimized so that the distribution of hydrogen bond lengths seen in MD simulation in the FREADY model matches those seen in the experimental native structures. The resulting distributions of angles describing the geometry of hydrogen bonds (α, β, γ) agree with corresponding native distributions (even the parameters $A_{\tau_i\tau_j}$, $B_{\tau_i\tau_j}$ were optimized only relatively based on the hydrogen bonds length distribution).

We can use the hydrogen bonding functional form developed for UNRES since the coarsening in FREADY is similar to that in UNRES model. UNRES, same as

FREADY, represents each residue by two beads. A difference is that in UNRES positions of the peptide centers are considered explicitly and positions of Cα atoms are implicitly reconstructed. In FREADY, we explicitly model the Cα particles and the centers of the hydrogen bonding groups are assumed to be in the center of the Cα-Cα bonds. Conceptually UNRES relies on chemical physics principles, while the main drive of the FREADY model is the requirement that hydrogen bond distribution of MD simulations will mimic the hydrogen bond distribution observed in statistics of experimentally determined protein structures. The use of a hydrogen bond term is also a nice illustration of mixing different potential terms (from different sources) with the iterative sampling.

Fitting of $U_{NB}(r,\tau_1,\tau_2)$ is more complex and has been fully automated. In order to speed up convergence of our iterative algorithm it is a good idea to obtain a reasonable zero order guess for non bonded interactions. The zero order guess we have used is a Lennard Jones like potential between all pairs of CM particles and a repulsion r^{-12} term between all other particles which are described by $U_{NB}^{0}(r)$ in Eq. (3.6). For sake of simplicity, $U_{NB}^{0}(r)$ does not depend on interacting residues' types and residue dependent features of the non-bonded term are recruited throughout the iterative learning process. The three adjustable parameters of $U_{NB}^{0}(r)$ were selected such that the average radius of gyration is preserved after *600 ps* long MD simulation for the structures in the training set.

For numerical reasons the functions $U_{NB}(r,\tau_1,\tau_2)$ are not fitted along the whole range of distances at once. The non bonded interactions are constructed as piecewise continuous and differentiable (to the first order) terms. The distances in range $r\in\langle 4.2\text{ Å},13.5\text{ Å}\rangle$ are fitted by least squares (LS) algorithm to nine degree polynomials. The optimization is constrained such that the function $U_{NB}(r,\tau_1,\tau_2)$ and its first derivative vanish at $r=13.5$ Å. The parameters $A_{\tau_1\tau_2}$, $B_{\tau_1\tau_2}$, $C_{\tau_1\tau_2}$ (from Eq.

(3.5)) of the target functions are fitted against the distributions at distances smaller than 4.2 Å with the constraints that $U_{NB}(r,\tau_1,\tau_2)$ has continuous first derivative at $r = 4.2$ Å. The function splitting at 4.2 Å was motivated by steep characteristics of $U_{NB}(r,\tau_1,\tau_2)$ at shorter distances and by rather smooth behavior of the non-bonded potential at larger separation.

3.4 Results

The iterative algorithm described in the previous section converged to a fixed set of parameters for the FREADY potential after about 20 iterations. Covalent local interaction terms such as bond lengths converge more rapidly and stabilize after a few (up to three) iterations. Figure 3.2 shows typical converged angular and torsional interactions. Comparisons of the native distributions to those obtained from the final training iteration are also shown. In Figure 3.3 we illustrate how a non-bonded interaction term evolves during the training process and Figure 3.4 illustrates how the radial distribution functions between these pairs of residues evolved from the initial to the final iteration. Overall individual distributions of the variables extracted from the PDB are accurately represented by the converged distributions of the final iteration. The small deviations from the PDB distribution that are observed in Figure 3.4 are typical.

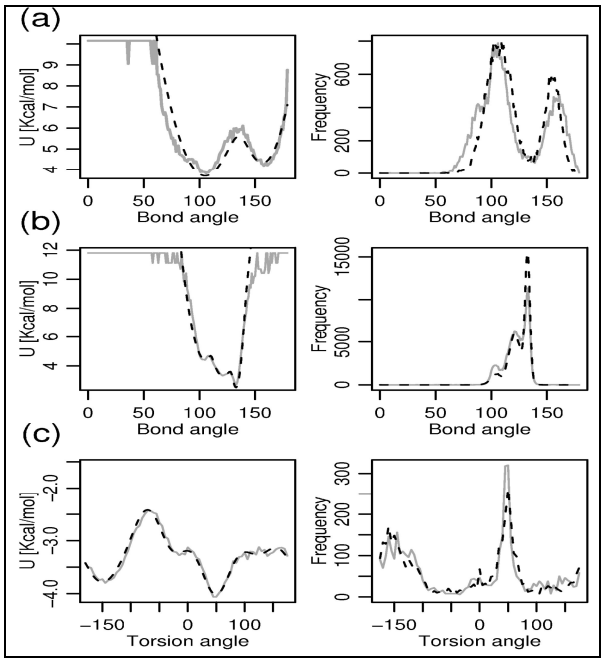


Figure 3.2: (a) left: Fit of the angle interaction term defined by $C\alpha_{i-1}$, $C\alpha_i$,CM_i for i-th residue being a TRP obtained by Boltzmann's inversion of the native distribution (gray) and the analytical fit to a double-well function (black, dashed). **right**: Comparison of distributions for this type of angles seen in the native structures (gray) and in the MD simulations driven by FREADY (black, dashed). **(b)** same as in (a), only for the central residue being VAL. The angle is of triple-well character in this case. **(c) left**: Fit by Discrete Fourier Transform (black, dashed) to the final version of the torsion potential (gray) defined by four consecutive Cα particles (for central two residues being TYR, ASN) **right**: Comparison of this torsion angle distribution in the native structures (gray) and in the MD simulations (black, dashed) for this dihedral angle type.

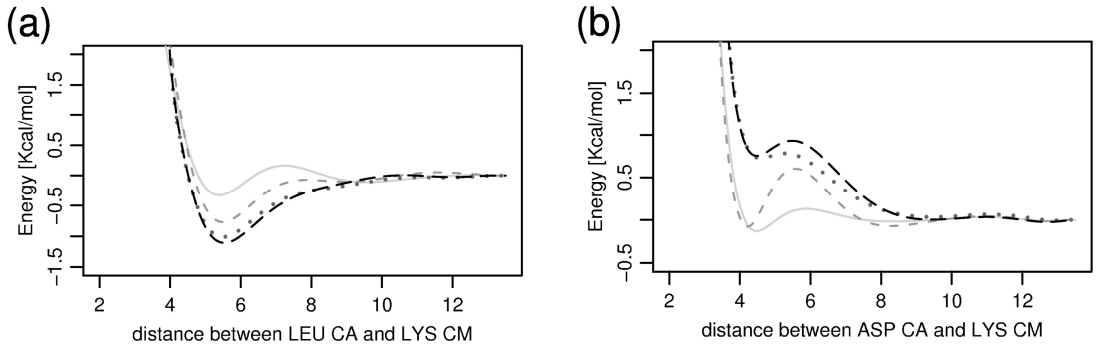


Figure 3.3: Iterative adjustments to the non-bonded interaction term between **(a)** LEU particle Cα and LYS particle CM; (**b)** ASP particle Cα and LYS particle CM. The interactions are evolving during the training in the order gray-solid (1. iteration), gray-dashed (3.iteration), dotted (11. iteration), and black-dashed (the final, 20^{th}, iteration).

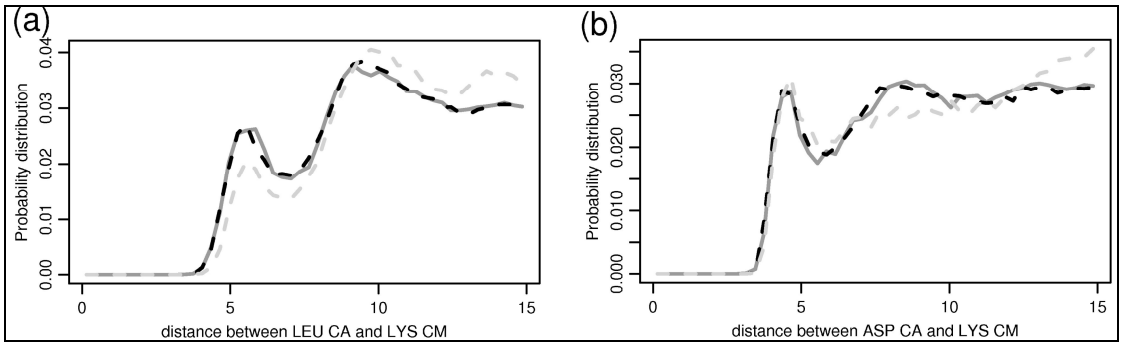


Figure 3.4: Radial distribution functions between pair of particles (**a)** LEU-Cα and LYS-CM (**b)** ASP-Cα and LYS-CM. The solid line corresponds to the distribution in the native structures, gray-dashed line depicts the distribution obtained after the first iteration of the training, and the black-dashed one stands for the distribution seen in the structures simulated by the final version of FREADY.

The quality of the final set of FREADY parameters was verified by two different tests: a) a stability test of the native protein conformations during MD simulations and b) a decoy recognition task. The stability of native conformations in FREADY potential was tested on native structures of proteins independent of the training set. The set used for the iterative training was based on the non-redundant set of protein structures covering the shapes available in PDB as of 6/28/2005. The test set for FREADY potential includes non-redundant representation of the protein structures deposited to the PDB between 6/28/2005 and 6/13/2006 (Vallat, Pillardy et al. 2008). The test set was filtered, as was done for the training set. We remove membrane proteins, RNA/DNA complexes, and PCAs (pyroglutamic acids). Group type MSEs (selenomethionines) are replaced by MET. Proteins with other non-standard amino acids were removed. Only proteins with a typical radius of gyration were kept. Further on, we reduced the test set to single chain proteins without any breaks in the backbone and limited the size of each protein to up to 500 residues. After all these constraints are met the test set consists of 956 native structures. A 21 ns MD simulation of each structure from the test set (driven by FREADY potential function) was performed. Every simulation begins from the native conformation by a short (200 steps) conjugate

gradient minimization. The simulations are initiated with 300 ps linear heating from 1 K to 300 K followed by 20.7 ns constant temperature simulation (controlled by velocity scaling).

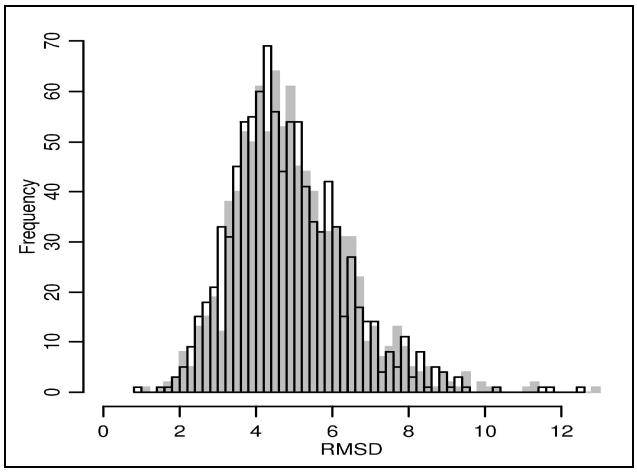


Figure 3.5: The distribution of RMSDs from the native fold after 10 ns (gray) or 21 ns (black, transparent) long MD simulation initiated from the native conformation.

Figure 3.5 shows a distribution of the RMSDs of the final structure of each MD simulation and the corresponding native conformation. Similarly Figure 3.6 shows distribution of the TM-score (Zhang and Skolnick 2004), which is measure of structural similarity that scales between 0 and 1. It is calculated as

$$\text{TM-score} = \max\left[\frac{1}{L}\sum_{i=1}^{L}\frac{1}{1+\left(\frac{d_i}{d_0}\right)^2}\right], \tag{3.11}$$

where L is the protein length, d_i is the distance between i-th pair of residues, $d_0 = 1.24\sqrt[3]{L-15} - 1.8$ is a distance scale, and maximum is taken over all structural superpositions.

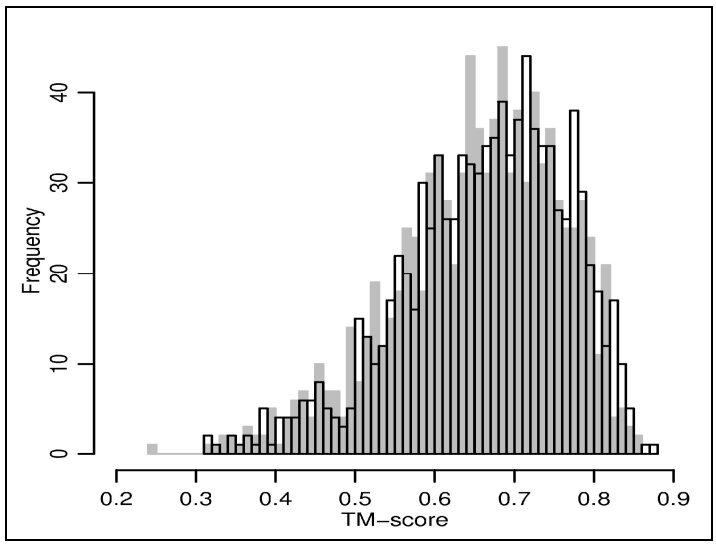


Figure 3.6: The distribution of TM-score between the native fold and structures obtained by 10 ns (gray) or 21 ns long (black, transparent) long MD simulation starting from the native conformation.

In contrast to RMSD the TM score can capture local similarities while the RMSD is sensitive to overall changes and to outliers. TM-score is calculated by an algorithm described in (Zhang and Skolnick 2004) and available from http://zhang.bioinformatics.ku.edu/TM-score/. The mean RMSD and TM-score against the native structures after 21 ns MD simulation are 4.95Å or 0.65, respectively. Figures 3.5 and 3.6 also show the distributions after 10 ns of MD. Only minor differences between the final distributions are observed. This observation suggests that most of the structures in the test set reach equilibrium after 10 ns.

The equilibrated distributions of internal degrees of freedom after 21 ns of MD are in good agreement with the distributions obtained from the native folds. Nevertheless, as shown on Figure 3.5 and 3.6, even when the target distributions of internal coordinates are preserved there are structures that diverge significantly from the native fold (RMSD larger than 10 Å or TM-score less than 0.4). This implies that the functional form of the potential chosen here (i.e. sum of local, pairwise terms and

backbone HB) is not sufficient to fix the average structure in the neighborhood of the native fold during room temperature simulations.

In Figure 3.7 we show results for three representative medium sized structures. Two of these proteins (1ido, 1a3k) remain relatively close to the native structure (RMSD of 2.33 Å and 3.42 Å). The third protein shown (1ge6) is an example in which the MD simulation drives the structure away from the native structure (9.87 Å). Figure 3.8 shows a comparison of mean square displacements of Cα particles during the last 10 ns of the test simulation with experimental crystallographic B-factors. The mean square displacements are in weak agreement with the experimental values. The location of the large fluctuations along the sequence seems to agree with experiment, but not the amplitudes. There are several residues in loop regions and close to either N or C terminals that have significantly higher displacements than those implied by B-factors. The same figure also shows that many of these overly-flexible parts of the structures are predicted as flexible also by Anisotropic Network Model (Eyal, Yang et al. 2006). Crystal packing might influence the reduced flexibility in some of these regions. Hence, the B factor may not represent the properties of an isolated protein molecule in solution.

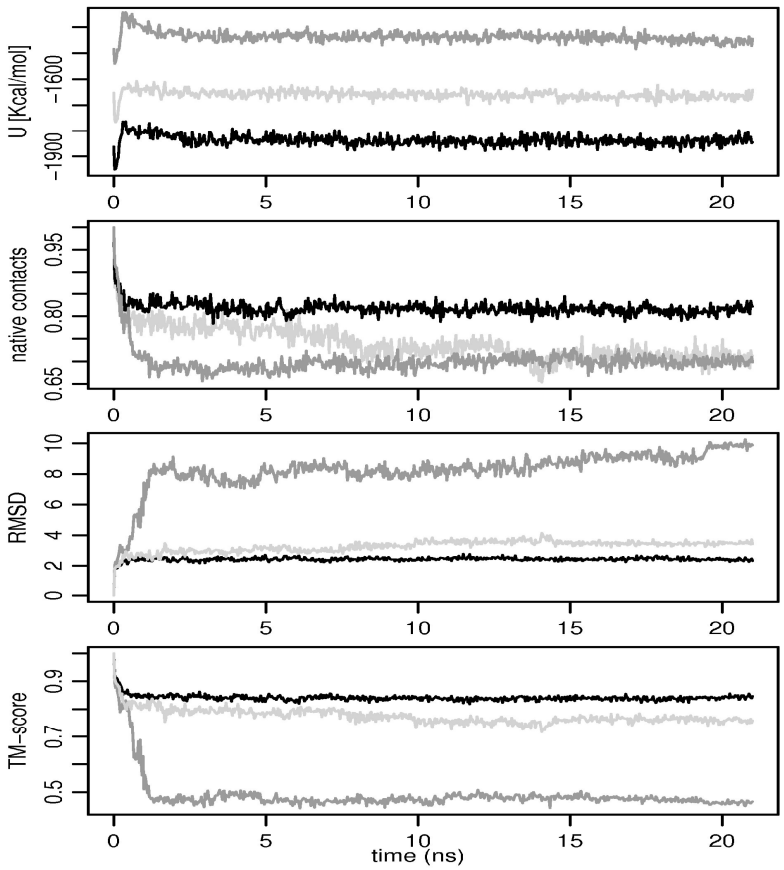


Figure 3.7: Behavior of three proteins 1a3k (an α/β protein, light gray), 1ido (an β protein, black), and 1ge6 (an α protein, dark gray) during the testing MD simulation driven by FREADY (21 ns). The figure shows from the top to the bottom the potential energy, the percentage of native contacts, the RMSD, and the TM-score.

Structural alignments of the final MD structures with the native conformations for these three proteins are given in Figure 3.9 - Figure 3.11. We have not found any correlation between stability of the native conformations in FREADY potential and the secondary structure content or composition (data not shown). We initially attempted to train FREADY without an explicit hydrogen bonding term. However, MD simulations of the training set driven by a potential trained without hydrogen bonding term resulted in the average deviation of 6.37 Å RMSD from the native structures compared to 4.95 Å obtained with a potential trained with explicit

backbone hydrogen bonding term. The reduced accuracy in our initial attempt was caused mainly by weak stability of native β sheets elements.

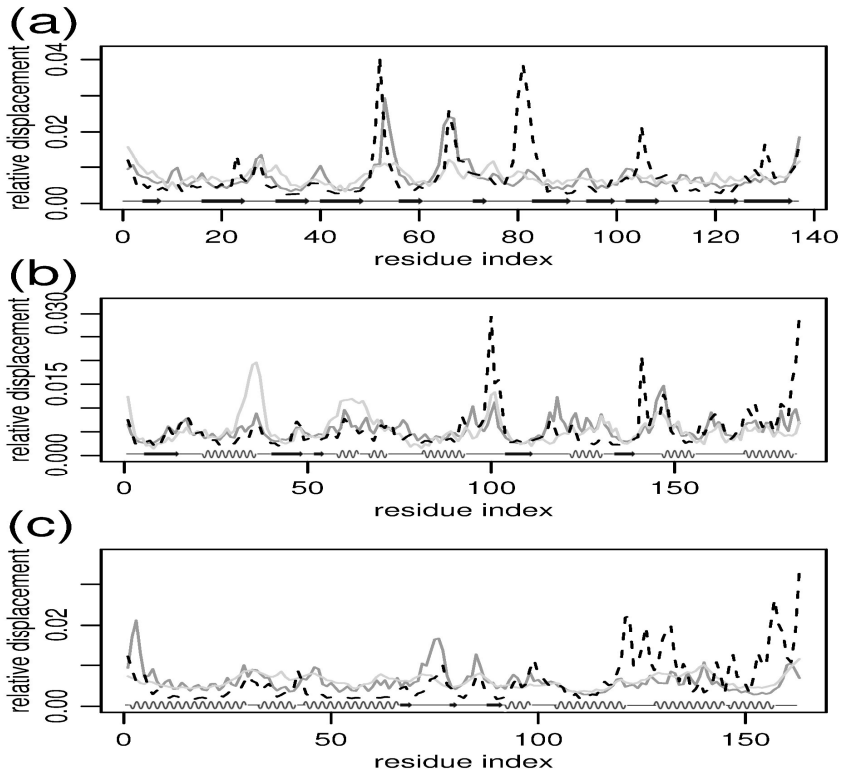


Figure 3.8: Comparison of experimental B-factors (light gray) of Cα atoms with mean square displacement in FREADY 21 ns MD simulations (black-dashed) and mean square displacements as predicted by ANM (Eyal, Yang et al. 2006) from the native conformation (dark gray). The values of all methods were scaled to have equal average displacements, so only relative displacements are meaningful. The graphs correspond from top to bottom to proteins 1a3k, 1ido and 1ge6. The correlation coefficients between experimental B-factors and simulation displacements are 0.4, 0.33, and 0.3 respectively. Secondary structure elements are shown at the lower part of the figure.

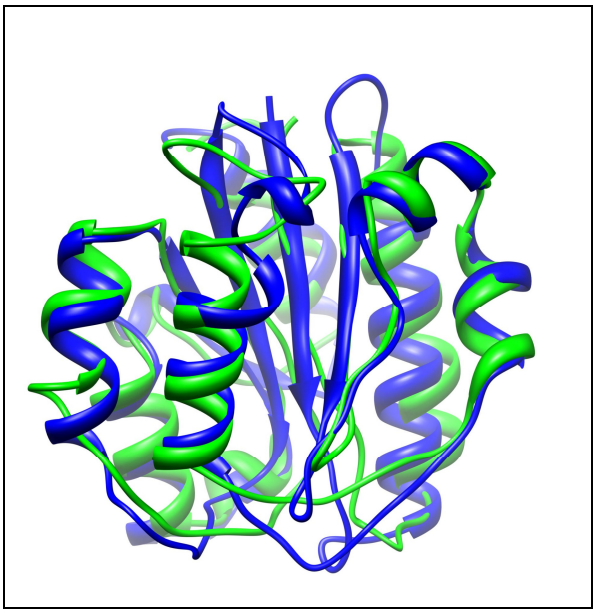


Figure 3.9: Alignment of native structure (blue) of 1ido (an α/β protein) and the conformation obtained after 21 ns of MD simulation (green). The RMSD is 2.33Å. Protein structures were aligned and visualized with UCSF Chimera tool (Pettersen, Goddard et al. 2004).

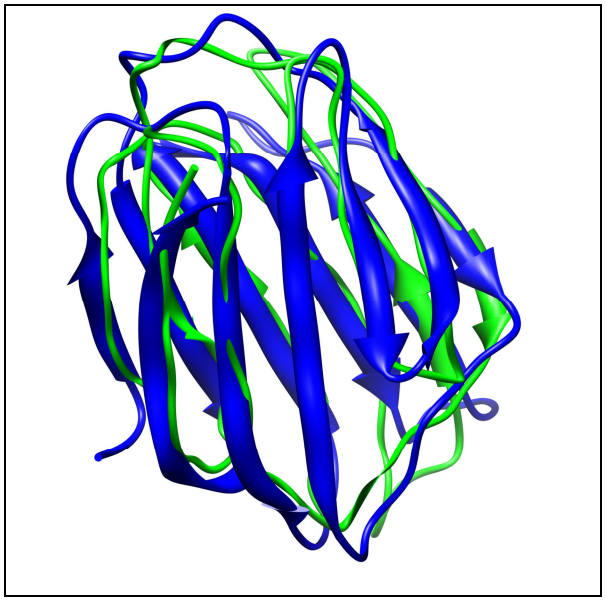


Figure 3.10: Alignment of native structure (blue) of 1a3k (a β protein) and the conformation obtained after 21 ns of MD simulation (green). The RMSD is 3.42Å.

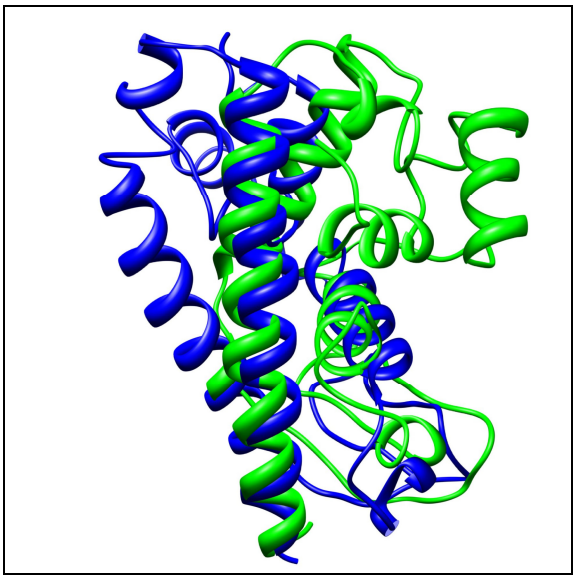


Figure 3.11: Alignment of native structure (blue) of 1ge6 (an α protein) and the conformation obtained after 21 ns of MD simulation (green). The RMSD is 9.87Å.

Better stability of native folds (3.92 Å from native in average) was reported recently by Minary and Levitt (Minary and Levitt 2008). They used a 3-bead model based on an all-atomistic statistical potential (Summa and Levitt 2007). There are two major differences between their approach and the results presented here. More extensive conformational search with a combination of parallel tempering and equi-energy Monte Carlo was performed in their work, whereas we only ran long MD simulations. Another important difference is in the number of degrees of freedom. In the work of Minary and Levitt secondary structure elements are fixed and the loop torsional angles are considered as the only degrees of freedom. Fixing the secondary structures in the simulations that uses the FREADY potential reduces the distance (RMSD) between the simulated structures and the native conformations in the 21 ns MD simulations to 3.04 Å in RMSD. The similarity increases to 0.78 measured with the TM-score.

The FREADY potential was also tested on native and near-native recognition from a set of decoy structures. Two datasets of decoys used in this study are “Decoys

'R' Us" dataset (Samudrala and Levitt 2000) and the set of decoys used for the training of LOOPP (Vallat, Pillardy et al. 2008). Both sets consist of a collection of different models generated as possible conformations for protein sequences with known structures (targets). "Decoys 'R' Us" dataset includes 34 targets, each target having from 500 to 2414 different models including the native structure. In the LOOPP dataset, there are 2470 protein targets, each having from 30 to 200 models. There is no overlap between the FREADY training set and the set of targets used in the LOOPP testing dataset.

In the decoy recognition task a set of different structures with an identical sequence (i.e. the sequence of the target) is provided. The task is to score the structure closest to the native (or the native itself, if present in the input set) as the model with the lowest energy. To use FREADY for this purpose only the sum of the non-bonded interactions and the torsional energies was used. By construction, the structures of the decoys have reasonable covalent geometries. Moreover, the local interaction terms of the bond and angular stretching are quite sensitive to local modifications in the structure and do not provide significant information about the overall quality of the three-dimensional shape. Therefore bond and angle terms of FREADY are not helpful in differentiating between native and decoy shapes.

Another type of interaction with a limited contribution is the short-range repulsion. The non-bonded interaction term as learned from MD simulations has steep repulsion for short distances (see Figure 3.3) which is not desirable for a structure recognition task (a single close contact can significantly increase the energy of an overall good model), thus the non-bonded interaction term U_{NB} for short distances was reduced through a logarithmic transformation to yield an adjusted value U'_{NB}

$$U'_{NB} = 0.6 + \frac{\log(U_{NB} + 0.4)}{10} \qquad \text{if } U_{NB} > 0.6 . \tag{3.12}$$

The last remaining term, the backbone hydrogen bonding, was not useful in recognition, probably because decoys in the datasets were generated with methods that optimize backbone hydrogen bonds.

FREADY performs similarly (see Table 3.1) to other statistical potentials on "Decoys 'R' Us" dataset. Only OPUS-PSP potential (Lu, Dousis et al. 2008), which uses more elaborate representation of side chain packing, performs significantly better than FREADY. The detailed performance of FREADY on "Decoys 'R' Us" dataset is provided in Table 3.2 and the contribution of different energy terms to the recognition in threading experiments is shown in Table 3.3. Seven targets from this dataset (1ctf, 1r69, 2cro, 1nkl, 1trl, 1dtk, 1shf) were present in the FREADY training set.

On the LOOPP dataset we tested the recognition of "native like targets," since statistical potentials tend to perform well in distinguishing the native structure from non-native ones but often fail in recognition of "close to native" conformations. Thus, in the case of LOOPP dataset, we ask how well does FREADY recognize native-like models (RMSD-wise) from other structures. FREADY ranks the model with the lowest RMSD as the lowest energy structure (within the top 5 lowest energy structures) in 50% (73%) of all 2470 targets. While clearly not perfect, FREADY provides a useful signal for model selection that when combined with other signals leads to more accurate prediction. FREADY signals were used in the LOOPP server during CASP8 exercise (Vallat, Pillardy et al. 2009).

Table 3.1: The comparison of several statistical potentials on "Decoys 'R' Us" dataset. Results for all potentials (except FREADY) are taken from the reference (Lu, Dousis et al. 2008). The second column lists number of targets which a given force field ranks as the lowest energy structure versus the total number of targets evaluated by that force field. The third column shows the average Z-score, $\left(\langle U\rangle - U_{native}\right) \Big/ \sqrt{\langle U^2\rangle - \langle U\rangle^2}$, of native structures.

	Top 1/Total Number	Mean Z-sore
OPUS-PSP [21]	**31/34**	**5.37**
HPMF [56]	**29/32**	**4.18**
FREADY	**28/34**	**4.62**
DOPE [57]	**28/32**	**-**
MSE [58]	**21/23**	**5.78**
DFIRE [38]	**27/32**	**4.52**
MJ_2005 [59]	**27/34**	**5.93**
DFIRE-SCM [60]	**23/32**	**4.36**
MM-PBSA [61]	**23/24**	**1.95**
DGR [62]	**21/25**	**5.25**
DWL [63]	**21/32**	**3.66**
TE13 [64]	**14/25**	**3.53**
CALSP [65]	**15/25**	**-**
Rosetta [66]	**14/32**	**-**

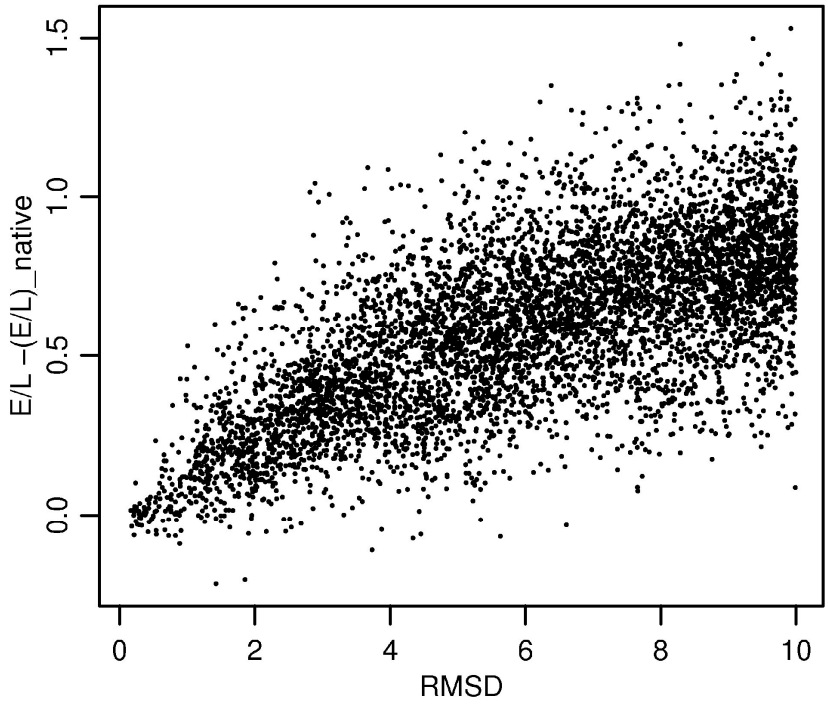


Figure 3.12: The difference of FREADY energy normalized by protein length from that of the native as a function of the RMSD from the native conformation. Each point in the figure corresponds to a model for a structure of a protein. There are 6034 models (for 338 targets) shown in the figure and only several structures score below the native conformations (negative values). On the average the energy seems a linear function of the RMS from the native suggesting a broad radius of influence for the FREADY potential.

Table 3.2: Performance of FREADY potential on "Decoys 'R' Us" dataset. The table lists for each target its PDB code, size of the decoy set, rank of the native structure in the set of decoys based on FREADY energy evaluation and Z-score of the native energy.

	PDB code	Decoy set size	Rank	Z-score
4state_reduced				
1	1cft	631	1	3.91
2	1r69	676	1	3.84
3	1sn3	661	1	3.83
4	2cro	675	1	3.29
5	3icb	654	1	2.57
6	4pti	688	1	4.34
7	4rxn	678	1	3.14
fisa				
8	1fc2	501	336	-0.27
9	1hhd-C	501	1	3.55
10	2cro	501	1	4.55
11	4icb	501	1	5.37
fisa_casp3				
12	1bg8-A	1201	1	3.91
13	1bl0	972	2	2.83
14	1eh2	2414	3	2.71
15	1jwe	1408	1	4.60
16	smd3	1201	1	6.72
lattice_ssfit				
17	1beo	2001	1	7.13
18	1cft	2001	1	8.37
19	1dkt-A	2001	1	7.71
20	1fca	2001	1	6.29
21	1nkl	2001	1	7.22
22	1pgb	2001	1	9.19
23	1trl-A	2001	1	4.98
34	4icb	2001	1	8.74
lmsd				
25	1b0n-B	498	16	1.62
26	1bba	501	493	-2.10
27	1cft	498	1	4.99
28	1dtk	216	1	3.12
29	1fc2	501	4	2.74
30	1igd	501	1	7.02
31	1shf-A	438	1	6.18
32	2cro	501	1	6.89
33	2ovo	348	1	3.57
34	4pti	344	1	4.48

Table 3.3: Contributions of different energy terms to the recognition of native structures in "Decoys 'R' Us" dataset. For each energy term the number of native structures recognize as the lowest energy structure by that term is given in the first column and the average Z-score of the native structures is given in the second column. Based on this data the sum of non-bonded and torsional energy terms was used for final prediction (the last row in the table).

	Top 1(from 34)	**Mean Z-score**
Bonds	**9**	**0.55**
Angles	**2**	**0.65**
Torsions	**14**	**2.45**
Nonbonded term	**27**	**4.17**
Hydrogen bonding	**2**	**1.19**

It turns out that FREADY performs better in recognition of structures obtained by X-ray crystallography than those obtained by NMR. The rate of best model recognition for targets solved by NMR drops to 31% (compared to 64% for structures solved by X-ray). The performance of FREADY on a subset of LOOPP dataset is shown in Figure 3.12. This set contains 338 targets that are single chain proteins, solved by X-ray crystallography, not forming biological complexes with other proteins or RNA/DNA, and are not membrane proteins. The correlation coefficient for this set between $E/L-(E/L)_{native}$ and the RMSD from the native conformation is 0.68. As seen in the figure, only several models have lower scores than the native (negative values on the figure) and most of the native-like models (low RMSD values) do not have high scores.

3.5 Final remarks

In this chapter we discussed a coarse grained potential that was learned using a mix of machine learning arguments and computational statistical mechanics. The potential was tested and illustrated to perform adequately at the two extreme limits of structural biology: (i) maintaining the structure in the neighborhood of the native fold in Molecular Dynamics simulations, and (ii) effectiveness in threading experiments. The significantly reduced number of degrees of freedom enables more comprehensive sampling for longer times. The simpler model (compared to all atom representation) is also effective in screening efficiently a large number of candidates to the correct fold. On the other hand, we do not expect the potential to work in domains it was not tuned for (e.g. protein folding).

We have addressed algorithmically two significant limitations of statistical potentials, that is, (i) how to learn a statistical potential that recovers experimental statistics in canonical simulations and (ii) how to combine statistical potentials with other energy terms that are necessary when comprehensive sampling is desired. Specifically in the present study we illustrate that the addition of hard cores and hydrogen bonding potentials is straightforward once generalized ensemble approach is applied. While hard cores could be added by statistical means (Miyazawa and Jernigan 1996), the iterative procedure allows for easy combination of different energy terms, potentially from different sources calibrated against the PDB distribution.

Perhaps the most intriguing observations made in this chapter are the limitations of the internal coordinate representation and of the assumption of potential transferability. We typically assume that a potential can be represented by pair interactions between amino acids (keeping the covalent geometry intact). The pair interaction is assumed to be transferable from a protein to a protein. Mathematical

programming studies illustrated however that the parameters of such a potential do not have a feasible solution on typical protein-like decoy sets (Michele and Eytan 1998; Tobi and Elber 2000; Tobi, Shafran et al. 2000). It is intriguing that a related conclusion is reached in the present work from a different perspective and for more general functional form.

Further studies of plausible functional forms of potentials, building on innovative work on modeling many-body potentials (Buchete, Straub et al. 2004; Ngan, Inouye et al. 2006; Feng, Kloczkowski et al. 2007), with comprehensive sampling and iterative refinement of potential parameters are of considerable interest.

REFERENCES

Aaqvist, J. and A. Warshel (1993). "Simulation of enzyme reactions using valence bond force fields and other hybrid quantum/classical approaches." Chem. Rev. **93**(7): 2523-2544.

Amir, E. A. D., N. Kalisman, et al. (2008). "Differentiable, multi-dimensional, knowledge-based energy terms for torsion angle probabilities and propensities." Proteins-Structure Function and Bioinformatics **72**(1): 62-73.

Berman, H. M., J. Westbrook, et al. (2000). "The Protein Data Bank." Nucl. Acids Res. **28**(1): 235-242.

Betancourt, M. R. and D. Thirumalai (1999). "Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes." Protein Sci **8**(2): 361-369.

Brown, S. and T. Head-Gordon (2004). "Intermediates and the folding of proteins L and G." Protein Sci **13**(4): 958-970.

Bryant, S. H. and C. E. Lawrence (1993). "An empirical energy function for threading protein-sequence through the folding motif." Proteins-Structure Function and Genetics **16**(1): 92-112.

Buchete, N. V., J. E. Straub, et al. (2003). "Anisotropic coarse-grained statistical potentials improve the ability to identify nativelike protein structures." The Journal of Chemical Physics **118**(16): 7658-7671.

Buchete, N. V., J. E. Straub, et al. (2004). "Development of novel statistical potentials for protein fold recognition." Current Opinion in Structural Biology **14**(2): 225-232.

Buchete, N. V., J. E. Straub, et al. (2004). "Orientational potentials extracted from protein structures improve native fold recognition." Protein Science **13**(4): 862-874.

Dill, K. A. (1985). "Theory for the folding and stability of globular proteins." Biochemistry **24**(6): 1501-1509.

Elber, R., A. Roitberg, et al. (1995). "MOIL: A program for simulations of macromolecules." Computer Physics Communications **91**(1-3): 159-189.

Eyal, E., L.-W. Yang, et al. (2006). "Anisotropic network model: systematic evaluation and a new web interface." Bioinformatics **22**(21): 2619-2627.

Feng, Y. P., A. Kloczkowski, et al. (2007). "Four-body contact potentials derived from two protein datasets to discriminate native structures from decoys." Proteins-Structure Function and Bioinformatics **68**(1): 57-66.

Goldstein, R. A., Z. A. Lutheyschulten, et al. (1992). "Protein tertiary structure recognition using optimized hamiltonians with local interactions." Proceedings of the National Academy of Sciences of the United States of America **89**(19): 9029-9033.

Haliloglu, T., I. Bahar, et al. (1997). "Gaussian Dynamics of Folded Proteins." Physical Review Letters **79**(16): 3090.

Hansmann, U. H. E., Y. Okamoto, et al. (1996). "Molecular dynamics, Langevin and hydrid Monte Carlo simulations in a multicanonical ensemble." Chemical Physics Letters **259**(3-4): 321-330.

Hill, T. L. (1956). Statistical Mechanics: Principles and selected applications. New York, Dover.

Hinds, D. A. and M. Levitt (1994). "Exploring conformational space with a simple lattice model for protein-structure." Journal of Molecular Biology **243**(4): 668-682.

J. D. Honeycutt and D. Thirumalai (1992). "The nature of folded states of globular proteins." Biopolymers **32**(6): 695-709.

Jagielska, A., L. Wroblewska, et al. (2008). "Protein model refinement using an optimized physics-based all-atom force field." Proceedings of the National Academy of Sciences of the United States of America **105**(24): 8268-8273.

Jayaram, B., K. Bhushan, et al. (2006). "Bhageerath: an energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins." Nucl. Acids Res. **34**(21): 6195-6204.

Jayasinghe, S., K. Hristova, et al. (2001). "MPtopo: A database of membrane protein topology." Protein Sci **10**(2): 455-458.

Kinnear, B. S., M. F. Jarrold, et al. (2004). All-atom generalized-ensemble simulations of small proteins, Elsevier Science Inc.

Kolinski, A. and J. Skolnick (1996). Lattice models of protein folding, dynamics and thermodynamics. Austin, Texas, Landes Company and Chapman Hill.

Lagant, P., D. Nolde, et al. (2004). "Increasing normal modes analysis accuracy: The SPASIBA spectroscopic force field introduced into the CHARMM program." Journal of Physical Chemistry A **108**(18): 4019-4029.

Liwo, A., S. Oldziej, et al. (2004). "Parametrization of Backbone-Electrostatic and Multibody Contributions to the UNRES Force Field for Protein-Structure Prediction from Ab Initio Energy Surfaces of Model Systems." J. Phys. Chem. B **108**(27): 9421-9438.

Liwo, A., M. R. Pincus, et al. (1993). "Calculation of protein backbone geometry from {alpha}-carbon coordinates based on peptide-group dipole alignment." Protein Sci **2**(10): 1697-1714.

Liwo, A., M. R. Pincus, et al. (1993). "Prediction of protein conformation on the basis of a search for compact structures: Test on avian pancreatic polypeptide." Protein Sci **2**(10): 1715-1731.

Lu, H. and J. Skolnick (2001). "A distance-dependent atomic knowledge-based potential for improved protein structure selection." Proteins-Structure Function and Genetics **44**(3): 223-232.

Lu, M., A. D. Dousis, et al. (2008). "OPUS-PSP: An Orientation-dependent Statistical All-atom Potential Derived from Side-chain Packing." Journal of Molecular Biology **376**(1): 288-301.

Luthy, R., J. U. Bowie, et al. (1992). "Assesments of protein models with 3-dimensional profiles." Nature **356**(6364): 83-85.

Maiorov, V. N. and G. M. Crippen (1992). "Contact potential that recognizes the correct folding of globular-proteins." Journal of Molecular Biology **227**(3): 876-888.

Maragakis, P. and M. Karplus (2005). "Large Amplitude Conformational Change in Proteins Explored with a Plastic Network Model: Adenylate Kinase." Journal of Molecular Biology **352**(4): 807-822.

Marrink, S. J., H. J. Risselada, et al. (2007). "The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations." J. Phys. Chem. B **111**(27): 7812-7824.

Meyerguz, L., C. Grasso, et al. (2004). "Computational analysis of sequence selection mechanisms." Structure **12**(4): 547-557.

Michele, V. and D. Eytan (1998). "Pairwise contact potentials are unsuitable for protein folding." The Journal of Chemical Physics **109**(24): 11101-11108.

Minary, P. and M. Levitt (2008). "Probing Protein Fold Space with a Simplified Model." Journal of Molecular Biology **375**(4): 920-933.

Miyazawa, S. and R. L. Jernigan (1985). "Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation." Macromolecules **18**(3): 534-552.

Miyazawa, S. and R. L. Jernigan (1996). "Residue - Residue Potentials with a Favorable Contact Pair Term and an Unfavorable High Packing Density Term, for Simulation and Threading." Journal of Molecular Biology **256**(3): 623-644.

Narang, P., K. Bhushan, et al. (2005). "A computational pathway for bracketing native-like structures for small alpha helical globular proteins." Physical Chemistry Chemical Physics **7**: 2364-2375.

Ngan, S. C., M. T. Inouye, et al. (2006). "A knowledge-based scoring function based on residue triplets for protein structure prediction." Protein Engineering Design & Selection **19**(5): 187-193.

Okazaki, K.-i., N. Koga, et al. (2006). "Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations." Proceedings of the National Academy of Sciences **103**(32): 11844-11849.

Pettersen, E. F., T. D. Goddard, et al. (2004). "UCSF Chimera - A visualization system for exploratory research and analysis." Journal of Computational Chemistry **25**(13): 1605-1612.

Reith, D., M. Pütz, et al. (2003). "Deriving effective mesoscale potentials from atomistic simulations." Journal of Computational Chemistry **24**(13): 1624-1636.

Rizzo, R. C. and W. L. Jorgensen (1999). "OPLS all-atom model for amines: Resolution of the amine hydration problem." Journal of the American Chemical Society **121**(20): 4827-4836.

Samudrala, R. and M. Levitt (2000). "Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction [In Process Citation]." Protein Sci **9**(7): 1399-1401.

Sippl, M. J. (1990). "Calculation of conformational ensembles from potentials of mena force: An approach to the knowledge-based prediction of local structures in globular proteins." Journal of Molecular Biology **213**(4): 859-883.

Skolnick, J., L. Jaroszewski, et al. (1997). "Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct?" Protein Science **6**(3): 676-688.

Spirin, S., M. Titov, et al. (2007). "NPIDB: a Database of Nucleic Acids Protein Interactions." Bioinformatics **23**(23): 3247-3248.

Summa, C. M. and M. Levitt (2007). "Near-native structure refinement using in vacuo energy minimization." Proceedings of the National Academy of Sciences **104**(9): 3177-3182.

Sun, Q., J. Ghosh, et al. (2008). Coarse-Graining of Condensed Phase and Biomolecular Systems. Boca Raton FL, CRC press.

Tirion, M. M. (1996). "Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis." Physical Review Letters **77**(9): 1905.

Tobi, D. and R. Elber (2000). "Distance-dependent, pair potential for protein folding: Results from linear optimization." Proteins: Structure, Function, and Genetics **41**(1): 40-46.

Tobi, D., G. Shafran, et al. (2000). "On the design and analysis of protein folding potentials." Proteins: Structure, Function, and Genetics **40**(1): 71-85.

Tusnady, G. E., Z. Dosztanyi, et al. (2004). "Transmembrane proteins in the Protein Data Bank: identification and classification." Bioinformatics **20**(17): 2964-2972.

Vallat, B. K., J. Pillardy, et al. (2008). "A template-finding algorithm and a comprehensive benchmark for homology modeling of proteins." Proteins: Structure, Function, and Bioinformatics **72**(3): 910-928.

Vallat, B. K., J. Pillardy, et al. "Building and assessing atomic models of proteins from structural templates." Proteins: Structure, Function, and Bioinformatics **76**(4): 930-945.

Wang, J. M. and P. A. Kollman (2001). "Automatic parameterization of force field by systematic search and genetic algorithms." Journal of Computational Chemistry **22**(12): 1219-1228.

Xia, Y., E. S. Huang, et al. (2000). "Ab initio construction of protein tertiary structures using a hierarchical approach." Journal of Molecular Biology **300**(1): 171-185.

Yap, E.-H., N. L. Fawzi, et al. (2008). "A coarse-grained alpha-carbon protein model with anisotropic hydrogen-bonding." Proteins: Structure, Function, and Bioinformatics **70**(3): 626-638.

Zhang, Y. and J. Skolnick (2004). "Scoring function for automated assessment of protein structure template quality." Proteins: Structure, Function, and Bioinformatics **57**(4): 702-710.

Zhou, H. and Y. Zhou (2002). "Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction." Protein Sci **11**(11): 2714-2726.

CHAPTER 4

MILESTONING WITHOUT A REACTION COORDINATE

4.1 Introduction

Milestoning (Faradjian and Elber 2004; Shalloway and Faradjian 2006; Elber 2007; West, Elber et al. 2007; Vanden-Eijnden, Venturoli et al. 2008; Kuczera, Jas et al. 2009; Maragliano, Vanden-Eijnden et al. 2009; Vanden-Eijnden and Venturoli 2009) is a method to calculate kinetics and thermodynamics of molecular systems that evolve on long time scales typically not accessible for straightforward Molecular Dynamics (MD) simulation.

Straightforward Molecular Dynamics can be used to compute rate of reactions. In these applications coordinates and velocities are initiated in the reactant state and the equations of motion are integrated until the product state is reached. While considerably promising there are caveats: (i) the numerical integration of a typical biomolecular process is computationally demanding and may not be feasible; (ii) actual realizations of reactive trajectories are noisy, making their analysis difficult and may require significant filtering to recover useful signals.

In Milestoning, the conformational space between the reactant and the product is partitioned by a set of dividing hypersurfaces called Milestones (Fig. 4.1). An ensemble of initial conditions is prepared at each Milestone and trajectories are simulated from each initial point until another nearby Milestone is reached. These trajectories are significantly shorter and trivially parallelized compared to a reactive trajectory of the overall process. The efficiency of the algorithm is discussed in (West, Elber et al. 2007).

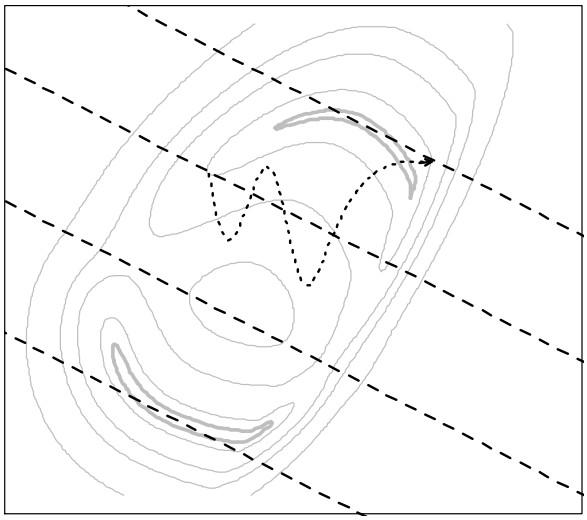


Figure 4.1: A schematic arrangement of Milestones (dashed lines) in a two-well potential. Also shown is a trajectory (dotted line) starting on a second Milestone and terminating on the first one.

In the original milestoning papers (Faradjian and Elber 2004; West, Elber et al. 2007), a theory that relates the statistical properties of the short trajectories initiated on each Milestone and the overall rate was developed. In the present work we consider a variant of the Markovian limit of Milestoning (Shalloway and Faradjian 2006; West, Elber et al. 2007), a method that uses only the first moments of local first passage time (LFPT) distributions. The advantage of the Markovian limit of Milestoning is that it is easier to implement and is statistically more stable. As we will show in Section 4.2.1 it calculates the overall mean first passage times (MFPT) accurately, given that certain assumptions are met. Milestoning in its complete settings (non-Markovian) provides a useful alternative if more detailed understanding of the reaction process is desired, for example if the reaction is non-exponential in time.

In (Vanden-Eijnden, Venturoli et al. 2008) reaction dynamics with overdamped Langevin dynamics was considered. It was shown that if Milestones are chosen as isocommittor surfaces, i.e. surfaces for which the probability of reaching the product state before the reactant is constant, then Milestoning calculation of the MFPT

using Brownian dynamics is exact. However, determination of exact isocommittor surfaces can be very difficult in practice.

Other limits in which Milestoning is expected to be accurate are available for systems near equilibrium. As outlined in the original Milestoning papers (Faradjian and Elber 2004; West, Elber et al. 2007), even when other surfaces are used (surfaces that are not isocommittors) Milestoning can still work well. If successive crossing events of Milestones are sufficiently separated in time to "lose" velocity memory Milestoning was illustrated to provide accurate results. This assumption is achieved in practice by placing Milestones sufficiently far from each other such that the average termination time of trajectories is at least a few hundred femtoseconds (West, Elber et al. 2007).

In Section 4.2 we propose a variant of Milestoning in the Markovian limit which we call Directional Milestoning (DiM) – the dividing hypersurfaces are redefined in more than one dimension to capture features of the reaction (e.g. multiple reaction channels or multiple collective variables) that at the same time maintain the concept of Milestone separation, e.g. trajectories initiated on any Milestone have time to "lose memory" before terminating on other Milestones.

The original Milestoning approach approximates the initial ensemble on each hypersurface by an equilibrium distribution. To be exact the initial distribution at a Milestone must be the first hitting point distribution (FHPD). A first hitting point is a phase space point on the Milestone crossed for the first time by a trajectory arriving from a nearby hypersurface. The distribution of these phase space points is complex and a closed form of it is known only for overdamped Langevin dynamics in low dimensions (Vanden-Eijnden, Venturoli et al. 2008).

In recent work (Vanden-Eijnden and Venturoli 2009), Vanden-Eijnden and Venturoli proposed a modification of Milestoning that avoids generation of initial

ensembles on each of the dividing surfaces. As we discuss later their approach is more accurate compared to the original Milestoning for the generation of the FHPD. Memory loss, however, is harder to control in the new approach. To improve the accuracy of the original Milestoning approach while retaining some of its advantages we propose in Section 4.2.4 another way to approximate FHPD which is better than the original Milestoning.

In Section 4.3 we illustrate the Directional Milestoning (DiM) on a calculation of MFPT of a conformational transition of alanine dipeptide, both in vacuum and in water and on a calculation of folding kinetics of a pentapeptide. We compare Directional Milestoning with exact Molecular Dynamics and with the related method Markovian Milestoning with Voronoi Tessellation (MMVT) (Vanden-Eijnden and Venturoli 2009). We illustrate that as the complexity of the underlying energy surface increases, DiM becomes more effective. Discussions and conclusions are presented in Section 4.4.

4.2 Directional Milestoning – theory

4.2.1 Definition of Milestones in higher dimensions

We discuss below an extension of Milestoning that avoids the use of a reaction coordinate. Instead of placing hypersurfaces orthogonal to a one-dimensional curve as introduced in the original papers (Faradjian and Elber 2004; West, Elber et al. 2007) we define the interfaces (Milestones) based on a set of coordinates (images) that sample the conformational space of the biophysical process under consideration. (Two of the images define the reactant and the product state.) These images may be obtained from long time simulations, high temperature trajectories, replica exchange simulations, etc., as discussed later in examples in the article. Having N images

$X_1, \ldots, X_N$ placed in the conformational space, we intuitively want to arrange Milestones as interfaces between the images, which is the approach taken in the Voronoi Tessellation of Markovian Milestoning (Vanden-Eijnden and Venturoli 2009). However, we aim to place the Milestones in conformational space in such a way that a trajectory initiated on any Milestone has time and space to "lose memory" of its starting point before terminating at a different Milestone. A formal definition of "losing memory" will be given in the following section. For each pair of images X_i and X_j we define the Milestone $M_{i\to j}$ as a set of conformational points on which a trajectory enters the region of image X_j from the region of image X_i. Formally, the above intuitive requirements on Milestone placement can be accomplished in several different ways. We define a Milestone $M_{i\to j}$ as follows

$$M_{i\to j} \equiv \left\{ X \mid d(X, X_i)^2 = d(X, X_j)^2 + \Delta_i^2 \text{ and } \forall k \; d(X, X_j) \le d(X, X_k) \right\}, \qquad (4.1)$$

where $d(X,Y)$ is a distance function of images X and Y and $\Delta_i = \min_{j\neq i} d(X_i, X_j)$. The arrangement (4.1) has a few important properties discussed in detail in Section 4.2.3. We name some of the properties here, referring the formal proofs to Section 4.2.3: A Milestone $M_{i\to j}$ is located in the region between the images X_i and X_j and is always closer to the image X_j. The Milestone $M_{i\to j}$ does not intersect any of $M_{i\to l}$ Milestones (for $l \neq j$) and there is a finite separation in conformational space between the Milestones $M_{i\to j}$ and $M_{l\to i}$. See Fig. 4.2 for an example of the proposed arrangement. As shown in the figure, the outgoing (black) Milestones bound the region of the central image and all the incoming (gray) Milestones are located within this region with a minimal distance to any of the outgoing Milestones.

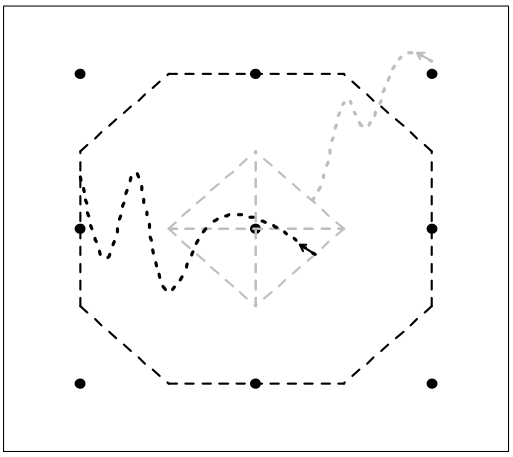


Figure 4.2: Example of Milestones according to definition (4.1). Conformational images are represented as black dots, Milestones related to the central image are displayed as dashed lines. A trajectory coming to the central region (gray, dotted) terminates on one of the gray Milestones (depending on the previously assigned region). A trajectory re-initiated on any of the gray (incoming) Milestones leaves the region through one of the black (outgoing) Milestones.

The proper selection of the conformational images $X_1, \ldots, X_N$ will be explained in more detail in Section 4.2.3; for now we assume their arbitrary placement. If Δ_i were omitted in the above definition ($\Delta_i = 0$) then the set of Milestones $M_{i\to j}$ is reduced to the Voronoi tessellation proposed in (Vanden-Eijnden and Venturoli 2009; Maragliano, Vanden-Eijnden et al. 2009); we refer to this arrangement as Markovian Milestoning with Voronoi Tessellation (MMVT) throughout this chapter. In the MMVT arrangement, the Milestone $M_{i\to j}$ is equivalent to the Milestone $M_{j\to i}$ and the only information they preserve is the identity of last-crossed Milestone, not the direction of such a crossing. (In a private communication Vanden-Eijnden disclosed an extension of MMVT to make the Milestones velocity dependent).

It is important to emphasize that the proposed placements of Milestones is not a tessellation. In accord with the definition of the original Milestoning, a trajectory is

identified by the last Milestone that it passes and not by its actual current position. A memory is carried out in time until the trajectory crosses another interface (Milestone). Trajectories from X_i to X_j can be fundamentally different from trajectories from X_j to X_i. To exploit this observation it is useful to make the Milestones dependent on the direction. We therefore call Milestones defined according to Eq. (4.1) Directional Milestones. The role of the additional flexibility offered by Δ_i is to avoid counting rapid transitions between interfaces due to spatial proximity of Milestones. As a result, the Milestones defined by Eq. (4.1) depend on more than the coordinates alone. This is consistent with the notion of a Milestone $M_{i\to j}$ ($M_{k\to j}$) as a state of a trajectory that arrives from the region X_i (X_k) to the region of image X_j. Hence the definition of a Milestone is extended to include information about the previous assignment of the trajectory. If the system is assigned to a region X_{i_0} at time 0 then by following a trajectory of the system one can deterministically identify the sequence of Milestones the trajectory has passed through $M_{i_0\to i_1}, M_{i_1\to i_2}, M_{i_2\to i_3}, \ldots, M_{i_{K-1}\to i_K}$.

4.2.2 Calculation of the mean first passage times

In the rest of this chapter we will use Roman subscripts to denote image index (as was done in the previous section) and Greek letters to denote Milestones. Consider the mean first passage time (MFPT) from any Milestone α to a given target Milestone β. We define it as follows: a trajectory is *assigned* to a Milestone α if the last Milestone it has passed through is α. *One-step transition* from a Milestone α to a Milestone β ($\beta \neq \alpha$) is a change of assignment of a trajectory from α to β. This step is clearly on a coarse Milestoning level and does not mean a single Molecular Dynamics step, which we will call a time-step. If such an event is possible we say that α connects to β. Note that by definition given in equations (4.1) if α connects to β, the second index of α (e.g. $M_{i\to j}$) must be equal to the first index of β ($M_{j\to k}$). The

first hitting point distribution on β, $\rho_\beta(p)$, is the distribution of phase space points (denoted by p) at which an equilibrium trajectory passes through β numerous times while the previous Milestone it passes through was not β. In further discussion only the relative weight of trajectories that pass through β is important so we can choose to normalize $\rho_\beta(p)$ such that $\int \rho_\beta(p)\, dp = 1$. We denote by $\left\langle \tau_{\alpha\beta}(p) \right\rangle$ the mean time of all trajectories that start from the phase space point p in α and terminate on Milestone β (possibly crossing other Milestones on the way). Integrating the last entity over p, weighting it by the probability that p is a phase space point at which an equilibrium trajectory hits α for the first time, $\int \left\langle \tau_{\alpha\beta}(p) \right\rangle \rho_\alpha(p)\, dp \equiv \left\langle \tau_{\alpha\beta} \right\rangle$, we obtain the MFPT from α to β.

Let the distribution of one-step transitions from α to β be $T_{\alpha\beta}(p,q,t)$, where p is the phase space point at which a trajectory starts in α and q is the phase space point at which the trajectory changes its assignment to β after time t. $T_{\alpha\beta}(p,q,t)$ is normalized in such a way that if we integrate over t and q we get conditional probability of a trajectory reaching β in one step given that it originates from p in α: $\int\int T_{\alpha\beta}(p,q,t)\, dqdt = P(\beta \mid \alpha, p)$, or alternatively $\sum_\beta \int\int T_{\alpha\beta}(p,q,t)\, dqdt = 1$. Note that by the definition of trajectory assignment, $T_{\alpha\alpha}(p,q,t) = 0$ for all p and q (since a trajectory cannot change its assignment from α to α).

Assuming that the phase space point $p(t+dt)$ can be determined from $p(t)$ only, as is true for most microscopic dynamics (e.g. Newtonian, or Langevin dynamics, but not Generalized Langevin dynamics) we make the following argument: The MFPT from α to β, $\left\langle \tau_{\alpha\beta} \right\rangle$, is defined as the weighted average of termination times of trajectories from α to β. Each trajectory, starting at p in α jumps in one step to some other Milestone γ ($\gamma \neq \alpha$) at phase point q and then in multiple steps (possibly 0, if $\gamma = \beta$) continues to β. Consider all the trajectories that jump in one-step from p in α to q in γ exactly in time t and then eventually reach β (in

potentially different total time). Since the microscopic dynamics is Markovian we can replace the contribution of these trajectories to $\langle \tau_{\alpha\beta} \rangle$ by $\left(t + \langle \tau_{\gamma\beta}(q) \rangle\right)$ weighted by sum of the weights of all of them (which is $\rho_\alpha(p) T_{\alpha\gamma}(p,q,t)$). By doing this for all possible combinations of γ and q we get the following equation:

$$\begin{aligned} \langle \tau_{\alpha\beta} \rangle &= \sum_\gamma \iiint \rho_\alpha(p) T_{\alpha\gamma}(p,q,t) \left(t + \langle \tau_{\gamma\beta}(q) \rangle\right) dp\,dq\,dt \\ &= \sum_\gamma \int \rho_\alpha(p) \left(\iint T_{\alpha\gamma}(p,q,t)\, t\, dq\,dt \right) dp \\ &+ \sum_\gamma \int \langle \tau_{\gamma\beta}(q) \rangle \left(\iint \rho_\alpha(p) T_{\alpha\gamma}(p,q,t)\, dp\,dt \right) dq \end{aligned} \tag{4.2}$$

The first term of the above equation can be reduced as

$$\begin{aligned} &\sum_\gamma \int \rho_\alpha(p) \left(\iint T_{\alpha\gamma}(p,q,t)\, t\, dq\, dt \right)\, dp \\ &= \sum_\gamma \int \rho_\alpha(p) \left(\frac{\iint T_{\alpha\gamma}(p,q,t)\, t\, dq\, dt}{\iint T_{\alpha\gamma}(p,q,t)\, dq\, dt} \iint T_{\alpha\gamma}(p,q,t)\, dq\, dt \right) dp \\ &= \sum_\gamma \int \rho_\alpha(p) \left(\langle t_{\alpha\gamma}(p) \rangle P(\gamma \mid \alpha, p) \right) dp \\ &= \int \rho_\alpha(p) \left(\sum_\gamma \langle t_{\alpha\gamma}(p) \rangle P(\gamma \mid \alpha, p) \right) dp \\ &= \int \rho_\alpha(p) \langle t_\alpha(p) \rangle\, dp = \langle t_\alpha \rangle, \end{aligned} \tag{4.3}$$

where $\langle t_{\alpha\gamma}(p) \rangle$, $\langle t_\alpha(p) \rangle$, and $\langle t_\alpha \rangle$ are average times of one-step transitions from $p \in \alpha$ to γ, from $p \in \alpha$ to any other Milestone, and from α to any other Milestone (averaged over p), respectively. In the second term of equation (4.2) the average time from $q \in \gamma$ to β is weighed by a factor depending on the phase space point $p \in \alpha$! To overcome this problem we use the following assumption: *The distribution at which any Milestone γ is hit does not depend on the Milestone to which the trajectory was assigned before the hit:*

$$\forall \alpha, \gamma: \quad \rho_\gamma(q) \propto \int \rho_\alpha(p) T_{\alpha\gamma}(p,q,t)\, dp\, dt\,. \tag{4.4}$$

It is easier to illustrate the properties of equation (4.4) if we consider a one-dimensional arrangement of Milestones in which the forward and the backward

Milestones occupy the same spatial coordinates. Consider a Milestone α that is pointing forward and is therefore denoted for the clarity of this discussion by $\alpha+$. There are two Milestones that initiate trajectories that may terminate at $\alpha+$. They are $(\alpha-1)+$ and $(\alpha-1)-$. Hence they occupy the same place in space but have their velocities pointing in the opposite directions. The assumption of equation (4.4) states that it does not matter if we start at $(\alpha-1)+$ or at $(\alpha-1)-$, both Milestones will generate the same hitting point distribution on $\alpha+$. If the initial direction of the velocity de-correlates quickly there should be no difference in the results from Milestone $(\alpha-1)+$ and $(\alpha-1)-$. In this case the assumption formulated in equation (4.4) will be satisfied. Indeed, we observed empirically in (West, Elber et al. 2007) that even the usual Milestoning works well when the velocity de-correlates. This empirical formulation is now formulated mathematically. In higher dimension we will also require spatial de-correlation

The multiplicative factor in the above equation is determined by the fact that if both sides of equation (4.4) are integrated over *q* the left side equals to 1 and the right side to $P(\gamma|\alpha)$; the conditional probability that if a trajectory changes its assignment from α it changes to γ. Therefore using the above assumption the second term of equation (4.2) reduces to $\sum_\gamma P(\gamma|\alpha)\langle\tau_{\gamma\beta}\rangle$ and we obtain the final form for the MFPT:

$$\langle\tau_{\alpha\beta}\rangle = \langle t_\alpha\rangle + \sum_\gamma P(\gamma|\alpha)\langle\tau_{\gamma\beta}\rangle . \qquad (4.5)$$

The set of equations (4.5) is extended by boundary conditions $\langle\tau_{\beta\beta}\rangle = 0$, $\langle t_\beta\rangle = 0$, and $\forall\gamma \quad P(\gamma|\beta) = 0$. It is a set of linear equations for all the $\langle\tau_{\alpha\beta}\rangle$ that can be solved by any standard linear solver. The size of the problem (the number of Milestones) never exceeded a few hundred in our hands. Equation (4.5) can be directly generalized for considering more than a single target Milestone (e.g. all incoming

interfaces to the folded state of a peptide). Alternative equations equivalent to equation (4.5) were derived in (West, Elber et al. 2007; Vanden-Eijnden, Venturoli et al. 2008). These equations are independent of the type of microscopic dynamics that we use (e.g. overdamped Langevin or Newtonian as long as it is microscopically Markovian). The system of linear equations (4.5) relates the overall rate (τ's) with the local kinetics information ($\langle t_\alpha \rangle$ and $P(\gamma|\alpha)$). Milestoning collects this local information in a more effective way than running an ensemble of trajectories from α to β. On each Milestone α, N_α phase space points are sampled from the FHPD ρ_α (see Section 4.2.4 for details). As a second step, each of the sampled phase space points is propagated in time until a connected Milestone is reached. The termination times of these trajectories are typically several orders of magnitude shorter than the overall MFPT of the system. Furthermore the trajectories between Milestones are independent of each other and thus can be run in parallel. For each Milestone γ connected to α we record $N_{\alpha\gamma}$- the number of trajectories that are initiated on α and terminated on γ. We also record T_α, the mean termination time of all N_α trajectories regardless of their terminal Milestone. The collected information $\{ N_{\alpha\gamma}, T_\alpha \}$ is used to estimate the required entities for equation (4.5) as

$$P(\gamma|\alpha) \cong N_{\alpha\gamma} / N_\alpha \qquad \text{and} \qquad \langle t_\alpha \rangle \cong T_\alpha . \tag{4.6}$$

In practice instead of using equation (4.6) we employ Bayesian inference on the collected data to calculate the MFPT supported by the data as well as an estimate of the statistical error due to the finite size of collected data. This procedure is described in detail in Appendix D.

4.2.3 Properties of Directional Milestones

The use of equation (4.5) for calculating MFPT depends on validity of the assumption expressed in equation (4.4). It has been shown in (Vanden-Eijnden, Venturoli et al. 2008) that the assumption formulated in equation (4.4) holds if overdamped Langevin dynamics is used and the Milestones are chosen as isocommittor surfaces. To our knowledge there is no efficient algorithm that identifies exact isocommittor surfaces and scales moderately with system size. However, there are other ways of satisfying equation (4.4). Instead we base our strategy on selecting Milestones according to equation (4.1), making sure that Milestones are sufficiently separated to allow for a memory loss of trajectories as outlined in the arguments of reference (West, Elber et al. 2007). Consider a pair of connected Milestones $M_{i\to j}$, $M_{j\to k}$ (defined by coordinate images $X_i, X_j,$ and X_k). Let S_{jk} be a hyperplane perpendicular to the line segment $X_j - X_k$ and passing through its midpoint. From equation (4.1) that defines $M_{i\to j}$ we know that each point on $M_{i\to j}$ is closer to X_j than to X_k. Thus the Milestone $M_{i\to j}$ lies on the X_j's side of S_{jk}. It follows from Lemmas C.1 and C.2 in Appendix C that S_{jk} and $M_{j\to k}$ are parallel, $M_{j\to k}$ lies on the X_k's side of S_{jk}, and that $d(S_{jk}, M_{j\to k}) = \Delta_j^2 / 2d(X_j, X_k)$. Therefore $d(M_{i\to j}, M_{j\to k}) \geq \Delta_j^2 / 2d(X_j, X_k)$. This minimal separation of connected Milestones is a property of Directional Milestoning that allows for some velocity relaxation to at least approximately satisfy the assumption described in equation (4.4). Note that the lower bound for the distance $d(M_{i\to j}, M_{j\to k})$ is a function of distances between the images that we place at will. Minimal separation of any two images places a lower bound on Δ_j's; additionally if one guarantees for each connected pair $M_{i\to j}$, $M_{j\to k}$ that $d(X_j, X_k)$ is about Δ_j then $d(M_{i\to j}, M_{j\to k}) \approx \Delta_j / 2$.

4.2.4 Sampling of the first hitting point distribution

The first step of Milestoning is to sample the initial conditions on each Milestone α from the first hitting point distribution $\rho_\alpha(p)$. An analytical expression for $\rho_\alpha(p)$ is in general unknown. In (Vanden-Eijnden, Venturoli et al. 2008) the authors provided the formula $\rho_\alpha(x) \propto e^{-\beta V(x)} \left|\nabla q(x)\right|$ for the case of overdamped Langevin dynamics with Milestones being placed as isosurfaces of the committor function $q(x)$. The last formula includes the gradient of committor function $\nabla q(x)$ which is difficult to get in high dimensions.

Instead of computing $\rho_\alpha(p)$ exactly (no exact expression is available for Newtonian dynamic), we approximate it. First, phase space points are sampled from the equilibrium distribution at Milestone $M_{i\to j}$. It can be done either by running an MD simulation constrained to the Milestone (Faradjian and Elber 2004; West, Elber et al. 2007) or by employing the Umbrella Sampling technique (see Appendix E and (Torrie and Valleau 1977)). The second step involves filtering each of the sampled phase points to determine those that are indeed first hitting events of $M_{i\to j}$. Exact verification tracks each of the sampled phase space points *p* back in time and tests termination on one of the incoming Milestones to the cell X_i ($M_{k\to i}$) before the trajectory intersects any of $M_{i\to l}$. (If $M_{i\to j}$ itself is crossed before any of $M_{k\to i}$, *p* is not the first hitting event of $M_{i\to j}$, it is at least a second hit of $M_{i\to j}$; if $M_{i\to l}$, $l \neq j$, is crossed before any of $M_{k\to i}$ then the trajectory must have entered to the cell of X_l before reaching *p* – therefore *p* cannot be the first hitting event of $M_{i\to j}$). Tracking the trajectory back in time to any of the Milestones $M_{k\to i}$ is similar in spirit to Transition Interface Sampling (Moroni, Bolhuis et al. 2004; Moroni, van Erp et al. 2004; van Erp and Bolhuis 2005) (TIS), the difference is that a TIS trajectory is propagated back in time until the reactant or the product state is hit. In DiM we perform significantly shorter backward verification, applicable only for equilibrium processes. TIS is exact,

however it is more expensive since in Milestoning we still exploit the use of trajectory fragments. Trajectory fragments are easier to parallelize and they can lead to implicit long time trajectories while in TIS long time individual trajectories need to be computed explicitly.

To retain high efficiency we track the trajectory back in time only until it reaches an empirical test boundary that is placed at a distance *d* on the X_i's side of the target Milestone $M_{i\to j}$ (*d* being smaller than or equal to the minimal distance to any of $M_{k\to i}$ from $M_{i\to j}$). If the trajectory reaches the checking boundary without re-crossing any other Milestone $M_{i\to l}$, we assume that *p* is a first hitting event. Otherwise we reject it. The procedure is schematically illustrated on Fig. 4.3.

In principle we can follow the trajectory back in time until one of the incoming Milestones to X_i ($M_{k\to i}$) or any of the outgoing Milestones from X_i ($M_{i\to l}$) is hit (a comment by Giovanni Ciccotti). By performing this complete verification the prepared ensemble on each Milestone would be the exact first hitting point distribution. However, the complete verification of each of the sampled phase points roughly doubles the overall computational cost (assuming reasonable acceptance ratio). The result of the more expensive exact verification will be reported elsewhere; in this chapter we report results and analysis of the more efficient (but approximate) checking protocol.

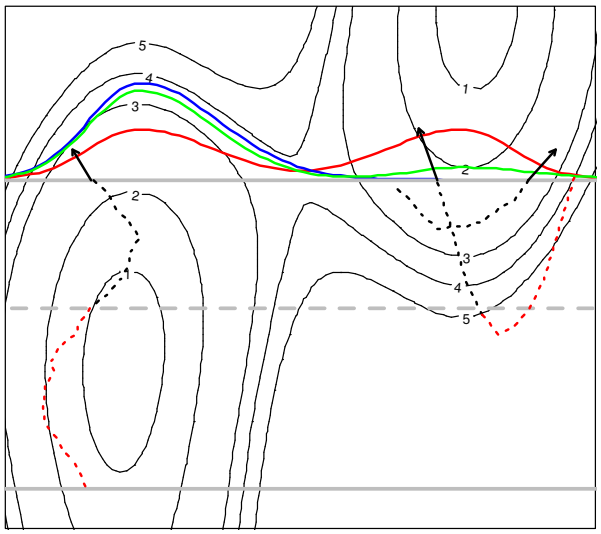


Figure 4.3: Illustration of sampling of the first hitting point distribution of trajectories initiated on the lower gray Milestone and terminating on the top (target) Milestone. The FHPD on the target Milestone (blue) is centered in the left basin, which is different from the equilibrium distribution (red). The FHPD is approximated by sampling phase space points from the equilibrium distribution and following each of them back in time until it hits the target Milestone on which it was initiated (the point is rejected) or the test boundary shown as a dashed gray line (it is accepted). Tracking of three phase space points is shown; the algorithm tracks only the black parts of the trajectories. Two of the points are accepted; one of them, however, is accepted by a mistake. The point is accepted because the test boundary was reached, however if the trajectory were checked further on (the red part) it would have been detected that the trajectory turns back and is not coming from the lower Milestone. Because of these false positive samples the resulting distribution (green) only approximates $\rho_\alpha(p)$ (blue). As the test boundary approaches the originating Milestone (lower gray) the sampled distribution approaches the true FHPD.

4.3 Applications of Directional Milestoning

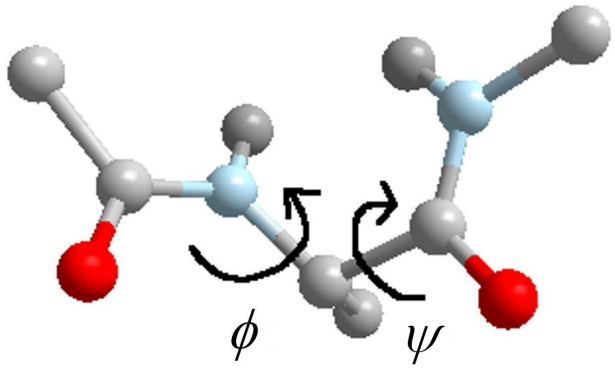


Figure 4.4: Alanine dipeptide.

4.3.1 Alanine dipeptide solvated in water

To demonstrate an application of Directional Milestoning we compute the MFPT of the transition between α helix and β sheet conformations in solvated alanine dipeptide (Fig. 4.4). The thermodynamics and kinetics of alanine dipeptide has been investigated in several studies (Ensing, De Vivo et al. 2005; Ren, Vanden-Eijnden et al. 2005; West, Elber et al. 2007; Maragliano and Vanden-Eijnden 2008; Maragliano, Vanden-Eijnden et al. 2009). In aqueous solution two dihedral angles, ϕ and ψ, shown in Fig. 4.4 are adequate coarse variables for the dynamics of the peptide. We therefore use a 2-norm distance in the reduced space of ϕ and ψ as the distance metric in the definition of Milestones (periodicity of the angles was taken into account in the calculation of a distance between two torsion angles).

The new module for Directional Milestoning was created in the program MOIL (Elber, Roitberg et al. 1995) and is available at https://wiki.ices.utexas.edu/clsb/wiki. The peptide molecule is solvated in a periodic box (20 Å)^3 of 248 TIP3P water molecules. The OPLS force field (Jorgensen and Tirado-Rives 2002) is used with electrostatics real space cutoff of 9 Å augmented with Particle Mesh Ewald summation. Van der Waals interactions are cut at a distance of

8 Å. All calculations were run in NVT ensemble at temperature of 303 K by employing a weak Andersen thermostat that acts only on the center-of-mass motion of the water molecules (Juraszek and Bolhuis 2008). The probability of velocity re-sampling was set to $5 \cdot 10^{-4}$ per fs. For a water box of this size an average of 13 water molecules had their velocities re-sampled in a 100 fs interval. This weak coupling does not change the transition rate obtained from NVE (Newtonian) simulations (with initial conditions sampled from the NVT ensemble). The free energy surface as a function of the two dihedral angles (ϕ, ψ) is shown in Fig. 4.5. It was calculated from statistics of a 340 ns long MD simulation. The white region of the map was not visited by the trajectory. There are two local free energy minima corresponding to an α helix conformation ($\phi, \psi = -100, -40$) and to a β sheet conformation ($\phi, \psi = -100, 140$).

The height of the free energy barrier between the two metastable regions at 303K is less than $2k_BT$ and the transitions between the metastable states are rapid on the trajectory time scale so the MFPTs can be estimated from straightforward MD simulations directly. We have performed five independent MD simulations of 68 ns. In each of the simulations more than 1000 transitions between the metastable regions occurred. The MFPT of $\alpha \to \beta$ transition is 66.4 ps ($\pm$ 2.7 ps) and that of the opposite transition is 53.8 ps ($\pm$ 4.6 ps). We set up the Milestoning calculation by placing six images in the conformational space in the positions ϕ_i, ψ_i=-100°,-240°+60°i, ($i = 1, \ldots, 6$). The positions of the images were not optimized. They were placed equidistantly in the region of conformation space that is accessible to the molecule. Table 4.1 shows the results of the Milestoning calculations for this system; it also includes the results of Markovian Milestoning with Voronoi Tessellation method (Vanden-Eijnden and Venturoli 2009). The MMVT calculation was performed with the same settings as for DiM, with the exception of the image placement; images for

MMVT calculation were placed at $\phi_i', \psi_i' = -100^\circ, -210^\circ + 60^\circ i$ (for $i = 1, 2, \ldots, 6$) so that the Milestones are placed in the same positions as in Directional Milestoning.

Table 4.1: Results of the MFPT calculations on alanine dipeptide solvated in water with 6 cells placed as shown on Fig. 4.5. Exact MFPTs were calculated by running five 68 ns long MD trajectories. The standard deviation of predicted MFPT of DiM and MD calculations are given in the brackets. For DiM, standard deviation was calculated from a single execution by using Bayesian inference (details in Appendix D). The total cost for DiM is given as a sum of the simulation time of all trajectories and the simulation time used for preparation of initial distributions.

Method	MFPT [ps], (sd [ps]) $\alpha \to \beta$ / $\beta \to \alpha$	total cost [ns]
straightforward MD	**66.4** (2.7) / **53.8** (4.6)	68
DiM, 100 trajectories/Milestone	66.5 (11.1) / 39.0 (4.6)	5.0 + 0.6 = 5.6
DiM, 250 trajectories/Milestone	57.7 (5.4) / 46.5 (3.6)	12.5 + 1.0=13.5
DiM, 500 trajectories/Milestone	61.2 (4.2) / 46.8 (2.6)	22.8 + 2.0=24.8
DiM, 1000 trajectories/Milestone	57.0 (2.7) / 45.2 (1.8)	46.1 + 3.9=50.0
DiM, 5000 trajectories/Milestone	59.5 (1.3) / 44.2(0.8)	230 + 10.1=240.1
MMVT, 0.4 ns /cell	60.2 / 43.9	2.4
MMVT, 0.8 ns /cell	57.2 / 43.7	4.8
MMVT, 1.6 ns /cell	63.2 / 41.2	9.6
MMVT, 3.4 ns /cell	63.4 / 53.2	20.4
MMVT, 12 ns /cell	62.4 / 48.3	72.0

Note that the employed dynamics is almost deterministic and thus a trajectory reflected from an interface (procedure required in MMVT) would approximately track itself back in time. Therefore we have slightly modified the MMVT protocol in a way suggested by Vanden-Eijnden in a private communication: instead of reversing the velocities of all the degrees of freedom at a cell interface, only the velocities of peptide atoms are reversed. This modification should not influence the statistics of observed fluxes through the interfaces since only the peptide degrees of freedom are used in the definition of cell boundaries.

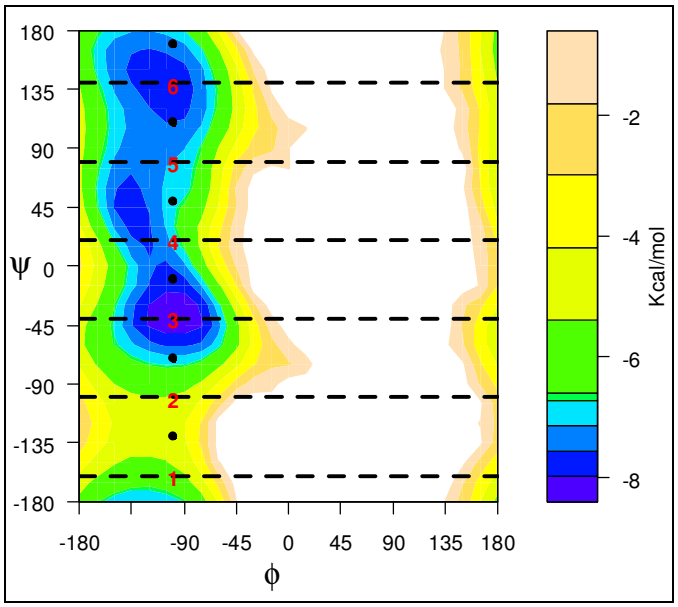


Figure 4.5: Free energy profile of alanine dipeptide as a function of the two dihedral angles ϕ and ψ. It was calculated from statistics of a 340 ns long MD simulation. Images for DiM calculations are placed at the positions of the red numbers and for MMVT calculation at the location of the black points. Both algorithms with these placement of images infer the Milestones in the positions of the dashed lines, in DiM, however there are two directional Milestones for each line.

Both methods, DiM and MMVT, perform well in this scenario, though MMVT is more efficient for this simple system. If enough sampling is done, both techniques provide reasonable estimates of MFPTs between the metastable regions, the systematic error is lower for MMVT (6 % and 10 %) as compared to our method (10 % and 18 %). Analysis of MMVT on the same system was performed recently (Maragliano, Vanden-Eijnden et al. 2009). A different force field was employed in (Maragliano, Vanden-Eijnden et al. 2009) and the MFPT reported differs by a factor of two from our calculations; however the relative error of MMVT for the reported $\alpha \to \beta$ transition is about 6%, which is comparable to our result. Results of $\beta \to \alpha$ transition were not reported in (Maragliano, Vanden-Eijnden et al. 2009). Table 4.1 shows that MMVT needs about 2-3 times less CPU time compared to DiM to converge. DiM requires more computations in these setting since each interface of MMVT is effectively doubled for the two different directions. Furthermore, additional

computation is needed in DiM to sample initial phase space points on each interface. In this one-dimensional set-up of Milestones with relatively large separation between Milestones and low free energy barrier MMVT is more efficient and as accurate as DiM. However, we will show below that with smaller separation between the interfaces, multi-dimensional arrangement of milestones, and rougher energy landscapes, DiM is better.

Even though previous Milestoning studies calculated accurately MFPTs on alanine dipeptide, memory effects in the system are not negligible. First hitting point distributions (in terms of ϕ angles) for the Milestones $M_{4\to5}$ and $M_{6\to5}$ are shown on Fig. 4.6. There is a noticeable difference between distributions of first hitting points on the Milestone $M_{4\to5}$ and on the Milestone $M_{6\to5}$. As shown on the figure, the approximate sampling described in Section 4.2.4 distinguishes the first hitting point distributions arriving from different directions to the region of image X_5 reasonably well.

In Table 4.2, we examine the use of directional Milestones on this system. The table shows that transitions between the six Milestones (if direction is not part of the description) are not Markovian. If no memory effects were present in the system then the probability of transiting to Milestone i+1 from Milestone i would not depend on the Milestone visited before i, i.e. the second and the forth columns of Table 4.2 would be the same within the error bars. We however see differences of up to 21% (for i=5) or by a factor of up to 2.2 (for i=1). One can see that the values of these relative probabilities estimated by Directional Milestoning (columns 3 and 5 in Table 2) are in good agreement with the true values.

Table 4.2: This table shows that dynamics of the alanine dipeptide system is not fully reducible to a Markov jump process between six hypersurfaces shown on Fig 4.5. The probability of jumping to the Milestone $i+1$ from the Milestone i depends on the Milestone visited before i. Probabilities (from a long MD trajectory) of jumping from i to $i+1$ if the Milestone $i-1$ ($i+1$) was visited before the hypersurface i are listed in the second (fourth) column. The third and fifth columns list these probabilities as measured by DiM calculation by starting 1000 trajectories from each Milestone. Note that in contrast to DiM, the original Milestoning assumes that $P(i \to i+1 / i-1 \to i) = P(i \to i+1 / i+1 \to i)$.

i	$P(i \to i+1 / i-1 \to i)$	$N_{M_{i-1\to i}M_{i\to i+1}} / N_{M_{i-1\to i}}$	$P(i \to i+1 / i+1 \to i)$	$N_{M_{i+1\to i}M_{i\to i+1}} / N_{M_{i+1\to i}}$
1	3.9	3.6	8.6	8.3
2	82.4	84.8	89.4	92.0
3	84.9	88.1	91.0	88.0
4	39.0	37.5	49.0	50.0
5	39.2	41.4	60.6	50.5
6	26.3	32.0	35.0	34.1

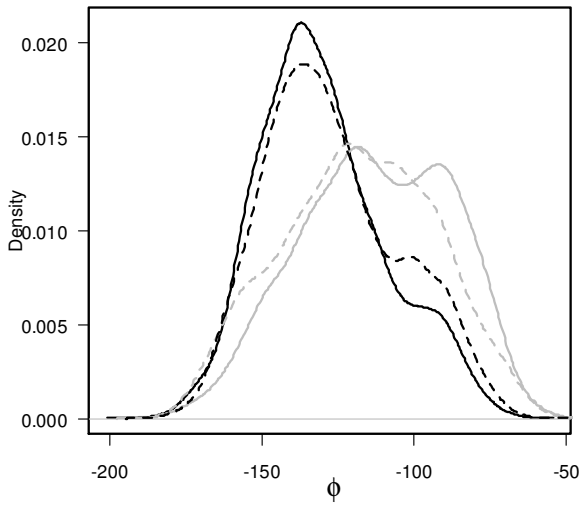


Figure 4.6: Distributions of ϕ angle of the first hitting point conformations of the region of image X_5 (located at $\psi = 80^\circ$): distributions observed in a long MD simulation for conformations arriving to the hypersurface at X_5 from the hypersurface of X_4 (black solid), or from that of X_6 (gray solid). Distributions sampled on the Milestone $M_{4\to5}$ (black dashed) and the Milestone $M_{6\to5}$ (gray dashed).

In the second experiment we examine both methods (DiM and MMVT) on the same system with Milestones in more than one dimension. This experiment is performed to empirically illustrate that placing Milestones in a non-linear arrangement does not compromise accuracy of DiM calculations. Images are placed in a two-dimensional grid covering the accessible space at the target temperature (conformations with torsional angle $\phi < 0$). For DiM, 18 images are placed in the positions marked $1, ..., 18$ on Fig. 4.7a). Each image has 8 incoming Milestones and 8 outgoing Milestones (displayed in solid and dashed on Fig. 4.7a) respectively). We calculated the MFPT from $M_{12\to11}$ (or $M_{10\to11}$) to the union of $M_{10\to9}$ and $M_{8\to9}$ for the $\beta \to \alpha$ transition. The MFPTs from these two Milestones differ from each other by about 0.3 ps and we report their average in Table 4.3. The opposite transition ($\alpha \to \beta$) was defined in the equivalent way.

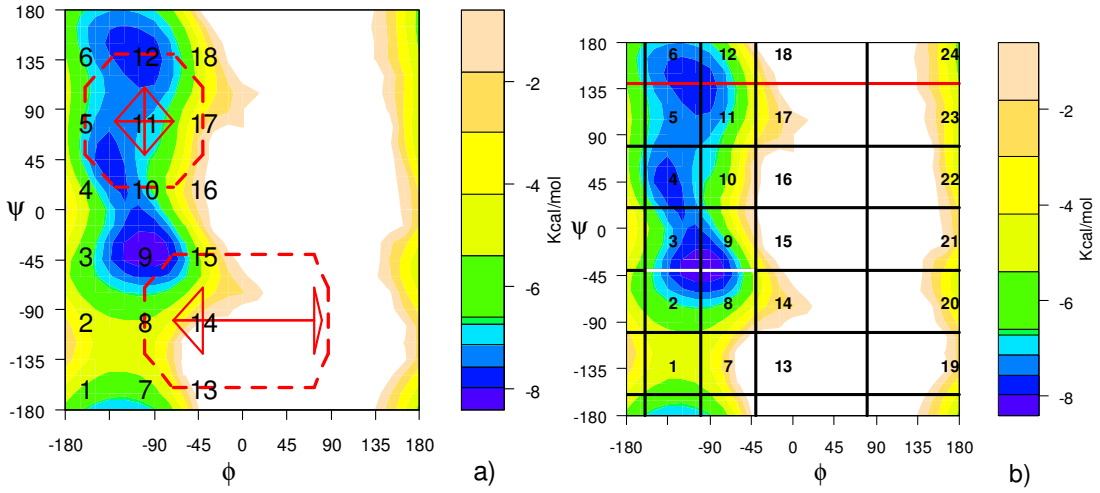


Figure 4.7: Placement of images on a two dimensional grid. (a) DiM settings: total of 18 images, located at position of numbers in the plot, are placed in a two dimensional grid. For two of the images, X_{11} and X_{14}, the outgoing (dashed) and incoming (solid) Milestones are shown. (b) Arrangement for MMVT. 24 images are placed in the conformational space so the resulting milestones are in the positions equivalent to DiM

For MMVT the images were placed in slightly different positions than for DiM (see Fig. 4.7b) such that the Milestones inferred by the Voronoi Tessellation are in equivalent positions to those used in Directional Milestoning. For the $\alpha \to \beta$ transitions, we calculated the MFPT of trajectories starting from the two white Milestones in Fig. 4.7b ($M_{2\leftrightarrow 3}$ and $M_{8\leftrightarrow 9}$) and terminating at the union of the red Milestones. MFPT of the transitions from these two starting points differ by less than 0.2 ps so only their average is reported in Table 4.3. The $\beta \to \alpha$ calculation was performed in the equivalent way (from the two central Milestones in the β sheet conformation ($\psi = 140^\circ$) to the union of all the Milestones with $\psi = -40^\circ$).

The results of both methods are listed in Table 4.3. The accuracy of Directional Milestoning is not compromised by multidimensionality; hence DiM works well for higher dimensions or higher connectivity of Milestones. The relative error of the MMVT method increased to 33 % (31 %). We think that this is mainly due to the corners between Milestones in the MMVT arrangement that cause rapid termination times between nearby Milestones and unwanted correlations between touching Milestones. Evidence of this can be seen in Fig. 4.8.

Table 4.3: Results of the MFPT calculations on alanine dipeptide solvated in water with 18 cells placed as on Fig. 4.7a). Standard deviations are in the brackets. Total cost for DiM is given as a sum of the simulation time of all trajectories and the simulation time used for preparation of the initial ensemble on each Milestone.

Method	MFPT [ps], (sd [ps]) $\alpha \to \beta$ / $\beta \to \alpha$	total cost [ns]
straightforward MD	**66.4** (2.7) / **53.8** (4.6)	68
DiM, 100 trajectories/Milestone	68.2 (10.0) / 56.9 (8.9)	10.0 + 2.6 = 12.6
DiM, 300 trajectories/Milestone	63.5 (4.9) / 56.6 (4.1)	31.1 + 8.7= 39.8
DiM, 1000 trajectories/Milestone	62.8 (2.5) / 53.2 (1.6)	103 + 26=129
DiM, 2000 trajectories/Milestone	65.7 (1.6) / 52.2 (1.1)	207 + 52=259
MMVT, 5 ns / cell	48.6 / 37.0	120
MMVT, 10 ns / cell	44.3 / 37.1	240

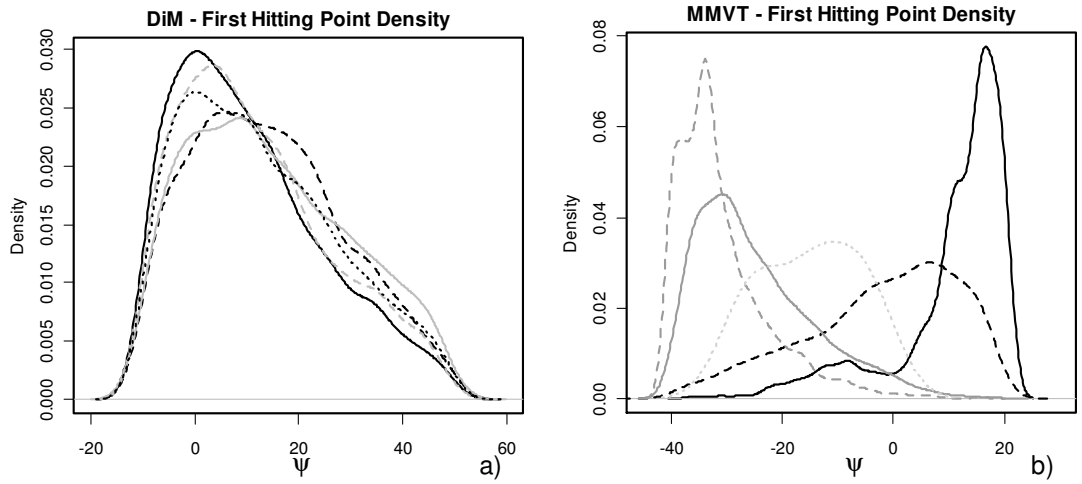


Figure 4.8: First hitting point distributions related to Fig. 4.7. (a) For DiM, distribution of ψ torsional angle of conformations arriving to the Milestone $M_{4\to10}$ from the Milestones $M_{9\to4}$ (black, solid), $M_{11\to4}$ (black, dashed), $M_{10\to4}$ (gray, solid), $M_{3\to4}$ (gray, dashed), and $M_{5\to4}$ (black, dotted). (b) For MMVT, distribution of ψ torsional angle of conformations arriving to the Milestone $M_{3\leftrightarrow9}$ from the Milestones $M_{10\leftrightarrow9}$ (black, solid), $M_{4\leftrightarrow3}$ (black, dashed), $M_{8\leftrightarrow9}$ (gray, solid), $M_{2\leftrightarrow3}$ (gray, dashed), and $M_{15\leftrightarrow9}$ (gray, dotted).

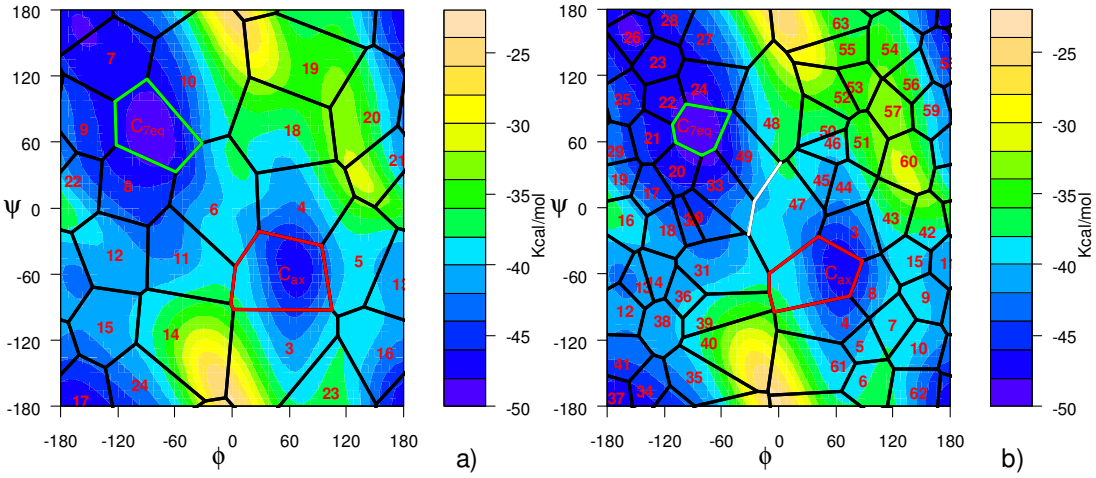


Figure 4.9: The shown landscape is an adiabatic ϕ, ψ energy map. The energy is minimized while constraining the ϕ and ψ dihedrals to specified values. Placement of (a) 24 images, (b) 63 images in the conformational space based on the algorithm described in Subsection 4.3.2.1 is shown. Also displayed is the Voronoi Tessellation based on the periodic Euclidean metric in the reduced space of ϕ and ψ torsions.

4.3.2 Alanine dipeptide in vacuum

In vacuum there are two stable conformers C_{7eq} and C_{ax} of alanine dipeptide (Fig. 4.9). The state C_{7eq} is further split into two sub-states denoted by C_{7eq} and C'_{7eq} (located at X_{26} in Fig. 4.9b) separated by a small barrier. We calculate the MFPT of transition from C_{7eq} to C_{ax} at two different temperatures, 400 K and 350 K, using Langevin dynamics. This is performed by calculating MFPT starting from each of the incoming Milestones to C_{7eq} region (green on Fig. 4.9) and considering union of the incoming Milestones to the region C_{ax} (red on Fig. 4.9) as the final state. The MFPT is not sensitive to exact identity of the starting Milestone (variation of less than 2%) therefore an average MFPT from all green Milestones is considered. The friction constant of Langevin dynamics was set to 30 ps^{-1}.

4.3.2.1 Image and cell generation

The images were generated by the following expansion. We start with the set of images $S = \{X_1, X_2\}$, where X_1 is a conformation located at C_{ax} and X_2 at C_{7eq}. Then we iteratively pick an image X from the set S and "expand" it: We launch trajectories starting from X with randomly initiated velocities and run each of these trajectories until it departs at least a pre-specified distance δ from X. Then we cluster the set of end points of these trajectories to existing images in S and potentially add new images to the set S if there are end points that are farther than δ from all images of S. We repeat this process until no new images are generated, i.e. we have tried launching trajectories from all images in S and all end coordinates are in S. There are three parameters in this algorithm: (i) the distance cutoff δ, (ii) the number of expanding trajectories N_e, and (iii) the clustering algorithm employed. For alanine dipeptide we have used expectation-maximization as a clustering algorithm (Hartley 1958), with N_e set to 400 and two different values of δ, $\delta_1 = 0.6$ Å and

$\delta_2 = 0.4$ Å. The root mean squared distance after optimal overlap (RMSD) (Kabsch 1976) is the distance metric (the RMSD between X_1 and X_2 is 1.25 Å) for the purposes of clustering as well as the distance function in the definition of Milestones (4.1).

4.3.2.2 Results for alanine dipeptide in vacuum

By using different values for δ we obtained sets of images of size 24 (for δ_1) and 63 (for δ_2); both are shown on Fig. 4.9. The tessellations shown in black in this figure are only approximate since they are based on the Euclidean distance in (ϕ,ψ) space, where the real interfaces (Milestones) are defined using the RMSD distance. The MFPT of the transitions between the metastable conformations are significantly longer than those in the solvated peptide due to higher free energy barriers. Tables 4.4 and 4.5 summarize the results of the Milestoning calculations in this system. At the high temperature (400 K) both methods, DiM and MMVT, predict accurate MFPT from C_{7eq} to C_{ax} (with systematic error of about 10%). MMVT needs to run about 1.5 μs MD simulations to obtain converged results, while DiM requires about 2.5 μs. Both of them provide significant speed up against straightforward MD simulation, even though a rough estimate of MFPT of the C_{7eq} to C_{ax} transition can be obtained by running about 11 independent MD simulations (equivalent to 4 μs of the total simulation time); however, both MMVT and DiM can be trivially parallelized to thousands of CPUs, shortening the actual time to perform the calculation.

Table 4.4: Results of the MFPT calculations on alanine dipeptide in vacuum with 24 cells placed as on Fig. 4.9a) at temperature 400 K. Standard deviations are in the brackets. Estimation of the exact MFPT was performed by launching five groups of 400 trajectories from C_{7eq} state and running them until C_{ax} state is reached (the MFPT reported in the table is calculated as the MFPT of all 2000 trajectories; the error is estimated by standard deviation of MFPTs calculated from each of the five groups). Total cost for DiM is given as a sum of the simulation time of all trajectories and the simulation time used for preparation of the initial ensemble on each Milestone.

method	MFPT [ns]	total cost [μs]
straightforward MD at T = 400 K	375 (16)	150
DiM, 500 trajectories/Milestone	630 (299)	0.13 + 0.09 = 0.22
DiM, 1K trajectories/Milestone	217 (103)	0.26 + 0.18 = 0.46
DiM, 3K trajectories/Milestone	306 (76)	0.78 + 0.47 = 1.25
DiM, 10K trajectories/Milestone	344 (37)	2.6 + 1.6 = 4.2
DiM, 20K trajectories/Milestone	387 (34)	5.2 + 3.1 = 8.3
DiM, 30K trajectories/Milestone	352 (31)	7.8 + 4.7 = 12.5
MMVT, 10 ns /cell	135	0.24
MMVT, 20 ns /cell	289	0.48
MMVT, 40 ns /cell	322	0.96
MMVT, 60 ns /cell	359	1.5
MMVT, 130 ns /cell	351	3.1
MMVT, 400 ns /cell	336	9.6

Table 4.5: Results of the MFPT calculations on alanine dipeptide in vacuum with cells placed as on Fig. 4.9a/b) at temperature 350 K. DiM was performed with 24 cells, MMVT in two different settings: 24 and 63 cells. Standard deviations are in the brackets. Estimation of the exact MFPT was performed by launching five groups of 200 trajectories from C_{7eq} state and running them until C_{ax} state is reached. Standard deviation and average of the MFPT calculated from each group are reported in the table. Total cost for DiM is given as a sum of the simulation time of all trajectories and the simulation time used for preparation of the initial ensemble on each Milestone.

method	MFPT [μs]	total cost [μs]
straightforward MD at T = 350 K	2.05 (0.3)	410
DiM, 5K trajectories/Milestone	2.78 (0.65)	2.3 + 1.4 = 3.7
DiM, 10K trajectories/Milestone	1.74 (0.40)	4.7 + 2.8 = 7.5
DiM, 20K trajectories/Milestone	1.75 (0.33)	9.4 + 5.6 = 15.0
DiM, 60K trajectories/Milestone	1.77 (0.20)	28 + 16.8 = 44.8
MMVT, 24 cells, 2.00 μs /cell	69.7	48
MMVT, 63 cells, 0.75 μs /cell	3798	47
MMVT, 63 cells, 2.25 μs /cell	855	142

When the temperature is lowered to 350 K (see Table 4.5) the C_{7eq} to C_{ax} transition is slower with MFPT of about 2.0 μs. As listed in Table 4.5, Directional Milestoning calculates the MFPT with systematic error of about 15% with as few as 7.5 μs of total simulation time. That is a significant speedup compared to straightforward MD since DiM can be easily parallelized on thousands of processors. MMVT fails to calculate the MFPT accurately. The main reason for this failure is poor statistics. An important difference between DiM and MMVT is that DiM allocates computational resources to each Milestone, where MMVT allocates the computational resources to a cell. If a transition between two specific interfaces in a cell is needed to describe the reaction and the transition is significantly less likely than transitions between other interfaces of the cell, then sampling this transition using MMVT is inefficient. A simple realization of this effect is the existence of a barrier in the middle of the cell. In that case MMVT trajectory is likely to be confined to a one minimum, to collide with the same interface many times (hits that do not count for the statistics) and to record only a few successful transitions to the other minimum. In contrast DiM launches a large number of short trajectories. These trajectories terminate quickly, and contribute to the statistics.

In DiM, sampling is done (extensively) at the interfaces, so the probability of observing a transition between interfaces of interest is greatly enhanced, since at least one end of the transitional event is sampled extensively. A potential problem in DiM is a large number of interfaces that may make sampling expensive. To avoid sampling irrelevant interfaces (at a given temperature) trajectories are initiated at few initial interfaces and only interfaces that are hit at least once during the DiM calculation are sampled and launched. We stop the DiM calculation when the process converges (i.e. no new interfaces besides those already sampled are reached).

For the MMVT calculation with 24 images, many cells cover a relatively large part of the conformational space with a rough energy landscape (see for example cell X_6 on Fig. 4.9a). This arrangement may cause poor statistics for those regions since the trajectories spend most of their time in low free energy regions, rarely visiting interfaces higher in free energy. To increase the probability of having a double hit at the two desired surfaces, we run the same calculation with 63 images as well. But even when 63 images are used, the allocation of computational resources is highly unbalanced. For example, we consider the frequency of hitting the interfaces $49 \to 47$ and $33 \to 47$ (displayed in white on Fig. 4.9b) that are both important for the overall MFPT. In both, 49 and 33 cells, confined simulations of total time of 2.25 μs hit a cell boundary more than $2 \cdot 10^7$ times. However, the interface $33 \to 47$ is hit only 17 times and the interface $49 \to 47$ only 7 times. In contrast DiM allocates equal number of starting trajectories to each of the Milestones and transitions from Milestones located near the transition states are sampled as well as other Milestones. We have not experimented with any selection criterions for allocation of computational resources to different cells (in MMVT) or to different Milestones (in DiM) but both methods may benefit from selective allocation of resources to "important regions" of conformational space.

4.3.3 Folding of a pentapeptide

We also performed DiM Calculations on a more realistic biophysical system. We studied folding thermodynamics and kinetics of a pentatpeptide Ac–WAAAH-NH_2 (wh5) at 300 K, which experimentally exhibits fast folding kinetics to an α-helical structure (Jas, Hegefeld 2010). The peptide molecule was solvated in a periodic box (30 Å)^3 of 801 TIP3P water molecules and one Cl- ion. The OPLS force field (Jorgensen and Tirado-Rives 2002) is used with the same settings as for alanine

dipeptide simulation described in Section 4.3.1. To cover conformational space accessible to the peptide by images, a 1 μs MD trajectory was executed with a structure being saved every ps. The obtained set of one million structures was clustered such that the set of clusters covers the conformational space accessed by the MD simulation. The distance metric, that is used to differentiate between structures, is Euclidean distance (with periodicity) in the ten-dimensional space given by ϕ and ψ backbone torsional angles of all five residues. A following greedy clustering algorithm was used: the fully helical structure is assigned as a center of the first cluster. Then, going sequentially along the structures from the MD trajectory, a new structure is assigned as a center of a new cluster if its distance from all other cluster centers is larger than 3 radians. 153 images were obtained in this way. A fast DiM calculation (by limiting each trajectory length to 3 ps) was performed to identify images that communicate rapidly. An image was removed from the initial set of images if there was a trajectory initiated in the image that terminated in less than 100 fs on a different image. This procedure reduced the number of images to ninety. A regular DiM calculation with the reduced set of images was performed. In total, there are 6186 directional Milestones between the images (according to Definition (4.1)) reachable at 300 K. From each interface, 50 trajectories were initiated, with overall mean termination time of a trajectory being 33.8 ps. The total accumulated simulation time is 11.8 μs (from which 1 μs was used for the initial MD sampling and 200 ns were used for preparation of the initial configurations on the Milestones). A markovian transition matrix Q between the Milestones is build from the collected data. From Q, we calculated the equilibrium probability of each Milestone and MFPT to the *native state*. The *native state* is considered as the union of all incoming Milestones to the α-helix image (by construction these Milestones are closer to the α-helix image than

any other image). The equilibrium probability of the native state defined in this way is about 2.5% (for comparison, fully unfolded states add up to 17.3%).

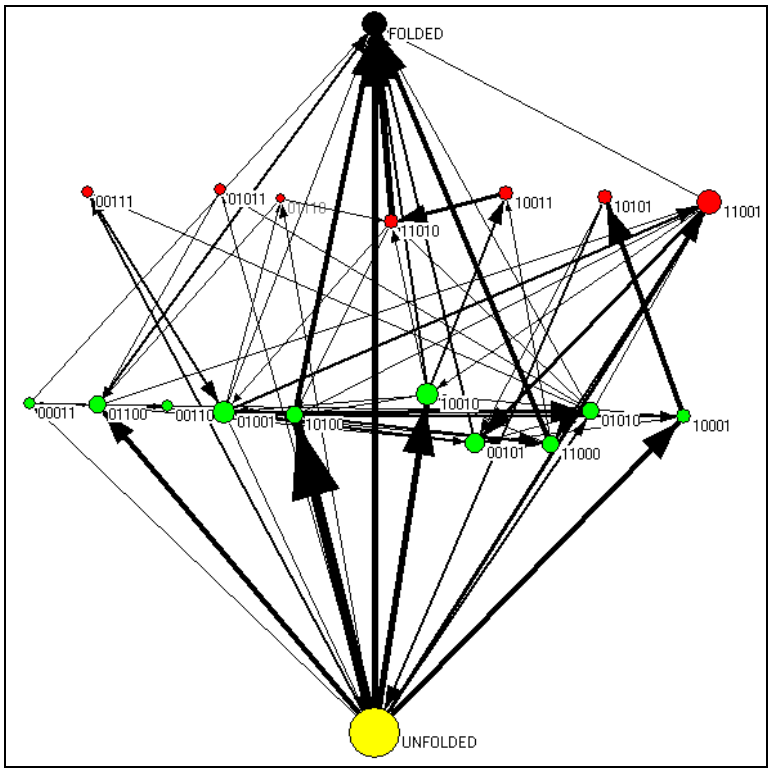


Figure 4.10: Schematic view of folding of wh5. Conformational space is divided to different groups by distinguishing state of each residue as helical (1) or non-helical (0) according to Ramachandran plot. By FOLDED group we denote images that have at most one residue in the non-helical state, by FOLDED those that have at most one residue in the helical state. The size of each state codes for its equilibrium weight and the width of each directed edge codes for amount of flux to the folding state along that edge.

The MFPT to the native state from all other Milestones weighted by the equilibrium probability is 4.0 ns what is in a good agreement with an estimation of MFPT from the 1 μs MD trajectory (9.3 ns). The value calculated by straightforward MD is of qualitative value since the equilibration of the MD trajectory might not be reached in 1 μs. Note that calculation of a long MD trajectory is not required for DiM in general. The initial set of images can be obtained by different less expensive techniques for example by those discussed in Chapter 2 or those in Section 4.3.2.1. A

schematic view of the folding conformational space for wh5 pentapeptide as calculated by DiM is shown in Figure 4.10. We calculated MFPT of folding also by MMVT with the same set of images and the same total simulation time as in the DiM calculation. However, the resulting average MFPT to the folding state as calculated by MMVT is of 113 ns, what is more than an order of magnitude larger than the value estimated by straightforward MD.

4.4 Discussions and conclusions

In this chapter, we proposed a method to compute dynamics in high dimensions called Directional Milestoning. We have shown that the mean first passage times between Milestones can be calculated accurately given that the distribution at which a Milestone is hit does not depend on the previously assigned Milestone (the assumption formulated in Equation (4.4)). Directional Milestoning arranges dividing hypersufaces in a special way, aiming to satisfy the above assumption: i) Milestones in DiM are made directional, so the local progress of the reaction (going from the region of X_i to X_j as opposed to being at the interface between X_i and X_j) is made part of the description, ii) the arrangement of Milestones guarantees a lower bound on spatial separation of any connected pair of Milestones so trajectories initiated on a Milestone have space and time to "lose memory" before terminating on a different Milestone.

The algorithm, while based on the trajectory fragments of Milestoning, is a step in the direction of Transition Interface Sampling (TIS) (Moroni, Bolhuis et al. 2004; Moroni, van Erp et al. 2004; van Erp and Bolhuis 2005) and Forward Flux Sampling (FFS) methods (Allen, Frenkel et al. 2006; Valeriani, Allen et al. 2007) compared to the original Milestoning. Here we use some trajectory tracking. The main difference between these methods and Directional Milestoning is that TIS and FFS are

tracking trajectories all the way back to the reactant state. This tracking has the advantage of not relying on any assumption about the initial ensemble on an interface like is done in Milestoning. On the other hand, sampling of trajectories in TIS and FFS is computationally more expensive than in Milestoning because every attempted trajectory in these methods is tracked back to the reactant state where in (Directional) Milestoning a trajectory is tracked only until it reaches a different Milestone. Computations of trajectory fragments can be done in Milestoning in a massively parallel way. The Partial Path Sampling method uses a conceptually similar approach of trajectory fragments (Moroni, Bolhuis et al. 2004).

An important distinction of Directional Milestoning compared to TIS, FFS, and the original Milestoning is that it allows for arbitrary arrangement of Milestones in conformational space, not necessarily following a linear arrangement along an order parameter or a reaction coordinate. A similar (arbitrary) arrangement of interfaces is used in the MMVT method (Vanden-Eijnden and Venturoli 2009), nonequilibrium umbrella sampling method (Warmflash, Bhimalapuram et al. 2007; Dickson, Warmflash et al. 2009), and Trajectory Parallelization and Tilting method (Vanden-Eijnden and Venturoli 2009). The last two techniques are using short trajectories in cells and balance the fluxes between cells. Recently the non-equilibrium umbrella sampling (Dickson, Warmflash et al. 2009) was illustrated to be more efficient than FFS (Allen, Frenkel et al. 2006). The Weighted Ensemble approach was also shown recently to work without a reaction coordinate (Zhang, Jasnow et al. 2010).

We have compared DiM with MMVT and showed that the performance of MMVT (in terms of effectiveness and correctness) is comparable to that of DiM in some of the examples, but that the correctness and/or effectiveness of MMVT can be compromised in systems with high free energy barriers, or in cells with two interfaces that are hard to reach. Another problem for straightforward implementation of MMVT

is the existence of corners between Milestones along more than one dimension that contribute to termination times that are too short. So while DiM is in general somewhat slower than MMVT it provides reliable results more consistently, including cases in which MMVT fails.

We also would like to comment on the similarities (and the differences) of our approach to the Markov State Model (MSM - for a recent study see (Noe, Schutte et al. 2009)). In the applications of MSM that we are aware of, long to very long Molecular Dynamics trajectories at normal conditions are used to estimate transition times and population of different cells. MMVT and DiM are designed to avoid such long trajectories (at the cost of approximate matching of probability densities at the interfaces). Once a sample of conformational space is available (which can be done in numerous ways, reaction path calculations, replica exchange simulations, or high temperature trajectories) only very short Molecular Dynamics trajectories are required to estimate the local kinetics. These short trajectories that can be trivially parallelized providing profound computational saving compared to straightforward Molecular Dynamics simulations. While significant progress has been made in parallelizing a single trajectory (Shaw, Deneroff et al. 2008), overhead still remains and special hardware that is frequently used is more expensive to buy and to maintain.

REFERENCES

Allen, R. J., D. Frenkel, et al. (2006). "Simulating rare events in equilibrium or nonequilibrium stochastic systems." The Journal of Chemical Physics **124**(2): 024102-16.

Dickson, A., A. Warmflash, et al. (2009). "Nonequilibrium umbrella sampling in spaces of many order parameters." The Journal of Chemical Physics **130**(7): 074104-12.

Elber, R. (2007). "A Milestoning Study of the Kinetics of an Allosteric Transition: Atomically Detailed Simulations of Deoxy Scapharca Hemoglobin." **92**(9): L85-L87.

Elber, R., A. Roitberg, et al. (1995). "MOIL: A program for simulations of macromolecules." Computer Physics Communications **91**(1-3): 159-189.

Ensing, B., M. De Vivo, et al. (2005). "Metadynamics as a Tool for Exploring Free Energy Landscapes of Chemical Reactions." Accounts of Chemical Research **39**(2): 73-81.

Faradjian, A. K. and R. Elber (2004). "Computing time scales from reaction coordinates by milestoning." The Journal of Chemical Physics **120**(23): 10880-10889.

Hartley, H. (1958). "Maximum likelihood estimation from incomplete data." Biometrics **14**: 174-194.

Jas, S. J., Hegefeld W., et al. (2010). "Fastest Simplest Helix." in preparation

Jorgensen, W. L. and J. Tirado-Rives (2002). "The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin." Journal of the American Chemical Society **110**(6): 1657-1666.

Juraszek, J. and P. G. Bolhuis (2008). "Rate Constant and Reaction Coordinate of Trp-Cage Folding in Explicit Water." **95**(9): 4246-4257.

Kabsch, W. (1976). "A solution for the best rotation to relate two sets of vectors." Acta Crystallographica Section A **32**(5): 922-923.

Kuczera, K., G. S. Jas, et al. (2009). "Kinetics of Helix Unfolding: Molecular Dynamics Simulations with Milestoning." The Journal of Physical Chemistry A **113**(26): 7461-7473.

Maragliano, L. and E. Vanden-Eijnden (2008). "Single-sweep methods for free energy calculations." The Journal of Chemical Physics **128**(18): 184110-10.

Maragliano, L., E. Vanden-Eijnden, et al. (2009). "Free Energy and Kinetics of Conformational Transitions from Voronoi Tessellated Milestoning with Restraining Potentials." Journal of Chemical Theory and Computation **5**(10): 2589-2594.

Moroni, D., P. G. Bolhuis, et al. (2004). "Rate constants for diffusive processes by partial path sampling." The Journal of Chemical Physics **120**(9): 4055-4065.

Moroni, D., T. S. van Erp, et al. (2004). "Investigating rare events by transition interface sampling." Physica A: Statistical Mechanics and its Applications **340**(1-3): 395-401.

Noe, F., C. Schutte, et al. (2009). "Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations." Proceedings of the National Academy of Sciences of the United States of America **106**(45): 19011-19016.

Ren, W., E. Vanden-Eijnden, et al. (2005). "Transition pathways in complex systems: Application of the finite-temperature string method to the alanine dipeptide." The Journal of Chemical Physics **123**(13): 134109-12.

Shalloway, D. and A. K. Faradjian (2006). "Efficient computation of the first passage time distribution of the generalized master equation by steady-state relaxation." The Journal of Chemical Physics **124**(5): 054112-8.

Shaw, D. E., M. M. Deneroff, et al. (2008). "Anton, a Special-Purpose Machine for Molecular Dynamics Simulation." Communications of the ACM **51**: 91-97

Torrie, G. M. and J. P. Valleau (1977). "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling." Journal of Computational Physics **23**(2): 187-199.

Valeriani, C., R. J. Allen, et al. (2007). "Computing stationary distributions in equilibrium and nonequilibrium systems with forward flux sampling." The Journal of Chemical Physics **127**(11): 114109-11.

van Erp, T. S. and P. G. Bolhuis (2005). "Elaborating transition interface sampling methods." Journal of Computational Physics **205**(1): 157-181.

Vanden-Eijnden, E. and M. Venturoli (2009). "Exact rate calculations by trajectory parallelization and tilting." The Journal of Chemical Physics **131**(4): 044120-7.

Vanden-Eijnden, E. and M. Venturoli (2009). "Markovian milestoning with Voronoi tessellations." The Journal of Chemical Physics **130**(19): 194101-13.

Vanden-Eijnden, E., M. Venturoli, et al. (2008). "On the assumptions underlying milestoning." The Journal of Chemical Physics **129**(17): 174102-13.

Warmflash, A., P. Bhimalapuram, et al. (2007). "Umbrella sampling for nonequilibrium processes." The Journal of Chemical Physics **127**(15): 154112-8.

West, A. M. A., R. Elber, et al. (2007). "Extending molecular dynamics time scales with milestoning: Example of complex kinetics in a solvated peptide." The Journal of Chemical Physics **126**(14): 145104-14.

Zhang, B. W., D. Jasnow, et al (2010). "The "weighted ensemble" path sampling method is statistically exact for a broad class of stochastic processes and binning procedures." The Journal of Chemical Physics **132**(5): 054107-7.

CHAPTER 5

CONCLUSIONS

This dissertation discusses three different strategies to reduce the computational costs of calculations addressing the large scale conformational transitions: action minimization algorithms combined with spatial coarse-graining, a systematic design of accurate coarse-grained potentials, and Milestoning algorithm generalized to complex processes. As introduced in the previous chapters, there are other methods available for calculating quantitative/qualitative descriptions of conformational transitions. Most of them, however, fail to scale to moderately sized (hundreds of residues) biological systems. To scale to the systems of this size we have concentrated on methods that reduce the complexity of the system, both in the spatial and the temporal terms. Moreover, the Milestoning method, even after the reductions, provides an accurate calculation of thermodynamics and kinetics along with the possibility of massive parallelization on a computer cluster.

Many of the ideas, as calculating the conformational transitions by an action minimization, usage of coarse-grained potentials, or Milestoning algorithm itself, have been around before the research described in this dissertation has been performed. We have, though, extended and combined these methods in ways that make them more applicable to practical problems faced in contemporary computational molecular biology. The set of methods presented in this dissertation, and implemented in the MOIL molecular modeling package, provides a set of tools capable of quantitative description of a moderately sized biophysical system.

There is, however, still room for improvements. The methods, as they stay now, are implemented to work on isolated and solvated protein systems without any atypical chemical modifications. Even on such ideal systems, interpretation of results

of these methods shall be performed carefully, with testing and verification. There are several internal parameters described in detail in the previous chapters that need to be set up properly to obtain accurate and statistically converged results. We need to, for example, keep in mind that a reduced quality of the employed atomistic force field would directly cause incorrectness of Milestoning or the action minimization algorithms. It is therefore recommended to verify applicability of the presented algorithms on a given system by first testing against available experimental evidence and test the stability of the algorithm with respect to its internal parameters, before any computational predictions are taken seriously.

In biology, moreover, many proteins consist of special residues, have bound ligands, or are in an interaction with different organic macromolecules (nucleic acids, sugars, or membranes). Significant amount of work is required to extend the methods presented in this dissertation to a set of robust tools applicable to such a set of biological systems.

The employed algorithms can be also improved to enable more accurate calculations or applications to larger biophysical systems. Here we list several suggestions for the future research: The efficiency of the action minimization algorithms can be improved significantly by employing more complex global function minimizers. The learning algorithm of FREADY potential can be modified to incorporate explicitly information about unfolded structures in the learning process and thus make the resulting potential applicable for example to protein folding. Milestoning algorithm without a reaction coordinate can accommodate several improvements: implementation of the exact sampling of the first hitting points instead of the approximate one described in Section 4.2.4, consideration of different Milestone geometries, or alternative image placement strategies, as in the current form, the

number of interfaces can become quite large (as demonstrated on the pentapeptide example in Section 4.3.3).

APPENDIX

APPENDIX A: PARALLEL CALCULATIONS OF BOUNDARY VALUE PATHWAYS

The calculation of trajectories with the SDEL formulation requires the determination of the paths that minimize the action S^l_{Gauss}, or more precisely minimize the target function T as described in (2.7). The simulated annealing procedure requires the values of $\partial T/\partial \mathbf{x}_j \quad j = 2,...,N-1$ for a gradient-based move. Let us examine the communication and computation required to calculate the $\partial T/\partial \mathbf{x}_j$'s. The most complex part of the function T is S^l_{Gauss} itself. S^l_{Gauss} is a sum of squared norms of $\partial \bar{S}/\partial \mathbf{x}_j$'s which are functions of $\mathbf{x}_j$, $\mathbf{x}_{j\pm1}$, $\partial U/\partial \mathbf{x}_j$, $U_{j\pm1}$, and U_j. See the exact formula in Appendix B. Here we use an abbreviation $\partial \bar{S}/\partial \mathbf{x}_j = \mathbf{F}(\mathbf{x}_{j-1}, \mathbf{x}_j, \mathbf{x}_{j+1}, \partial U/\partial \mathbf{x}_j, U_{j-1}, U_j, U_{j+1})$, and $\partial \bar{S}/\partial \mathbf{x}_{jk} = \mathrm{F_k}(\ldots)$. After substituting $\partial \bar{S}/\partial \mathbf{x}_j$ into $\partial S^l_{Gauss}/\partial x_{jm}$ we get

$$\begin{aligned}
\frac{\partial S^l_{Gauss}}{\partial x_{jm}} &= 2\sum_k \left(\frac{\partial \bar{S}}{\partial x_{j-1k}} \frac{\partial^2 \bar{S}}{\partial x_{j-1k} \partial x_{jm}} + \frac{\partial \bar{S}}{\partial x_{jk}} \frac{\partial^2 \bar{S}}{\partial x_{jk} \partial x_{jm}} + \frac{\partial \bar{S}}{\partial x_{j+1k}} \frac{\partial^2 \bar{S}}{\partial x_{j+1k} \partial x_{jm}} \right) = \\
&= 2\sum_k \begin{pmatrix}
\mathrm{F_k}\left(\mathbf{x}_{j-2}, \mathbf{x}_{j-1}, \mathbf{x}_j, \partial U/\partial \mathbf{x}_{j-1}, U_{j-2}, U_{j-1}, U_j\right) \dfrac{\partial}{\partial x_{j-1k}} \mathrm{F_m}\left(\mathbf{x}_{j-1}, \mathbf{x}_j, \mathbf{x}_{j+1}, \partial U/\partial \mathbf{x}_j, U_{j-1}, U_j, U_{j+1}\right) + \\
\mathrm{F_k}\left(\mathbf{x}_{j-1}, \mathbf{x}_j, \mathbf{x}_{j+1}, \partial U/\partial \mathbf{x}_j, U_{j-1}, U_j, U_{j+1}\right) \dfrac{\partial}{\partial x_{jk}} \mathrm{F_m}\left(\mathbf{x}_{j-1}, \mathbf{x}_j, \mathbf{x}_{j+1}, \partial U/\partial \mathbf{x}_j, U_{j-1}, U_j, U_{j+1}\right) + \\
\mathrm{F_k}\left(\mathbf{x}_j, \mathbf{x}_{j+1}, \mathbf{x}_{j+2}, \partial U/\partial \mathbf{x}_{j+1}, U_j, U_{j+1}, U_{j+2}\right) \dfrac{\partial}{\partial x_{j+1k}} \mathrm{F_m}\left(\mathbf{x}_{j-1}, \mathbf{x}_j, \mathbf{x}_{j+1}, \partial U/\partial \mathbf{x}_j, U_{j-1}, U_j, U_{j+1}\right)
\end{pmatrix} \\
&= \mathrm{G}_m\left(\mathbf{x}_{j\pm2}, \mathbf{x}_{j\pm1}, \mathbf{x}_j, \partial U/\partial \mathbf{x}_{j\pm1}, \partial U/\partial \mathbf{x}_j, \partial^2 U/\partial \mathbf{x}_j^2, U_{j\pm2}, U_{j\pm1}, U_j\right),
\end{aligned} \tag{A.1}$$

where $\mathbf{G}$ is a function that takes its listed inputs and returns a vector $\partial S^l_{Gauss}/\partial \mathbf{x}_j$. In order to calculate the derivative of T with respect to $\mathbf{x}_j$ we need the position and

potential information of five different structures, need to compute forces for three different structures, and compute a Hessian matrix for one structure. In previous studies the following protocol was used: A trajectory represented by *N* conformations is distributed among *P* processors, each processor being responsible for updating *N/P* successive conformations. Suppose that a processor *p* is responsible for conformations $\mathbf{x}_m,\ldots,\mathbf{x}_{m+N/P-1}$. According to Equation A.1, the processor *p* requires positions $\mathbf{x}_{m-2},\ldots,\mathbf{x}_{m+N/P+1}$ to update its conformations. However, the positions of $\mathbf{x}_{m-2},\mathbf{x}_{m-1},\mathbf{x}_{m+N/P},\mathbf{x}_{m+N/P+1}$ are modified on different processors, and these conformations have to be communicated from the processors responsible for their updates. The send and receive communications summed up to $4\cdot 3n_{pt}$ floating-point numbers per each step of simulated annealing (where n_{pt} is the number of particles in the system). This amount of communication may contribute significantly to the computation clock-time. Therefore, the forces $\partial U/\partial\mathbf{x}_{m-1}$ and $\partial U/\partial\mathbf{x}_{m+N/P}$ (together with $U_{m-1},U_{m+N/P}$) that are required as an input for the function $\mathbf{G}$ in Equation A.1, are recomputed on processor *p*, after the positions of $\mathbf{x}_{m-1},\mathbf{x}_{m+N/P}$ are received. It is recommended to communicate the values of U_{m-2} and $U_{m+N/P+1}$ since on most platforms their computation is more expensive than their communication. The proposed scheme requires *(N/P+2)* force computations, *N/P* Hessian matrix computations, and $\approx 4\cdot 3n_{pt}$ floating-point numbers received and sent in each step for each processor. The scheme provides reasonable scaling, unless the number of processors *P* approaches the number of conformations *N*. The last limit is approached for large system (like mGluR1) for which we wish to exploit the benefit of parallelization to the maximum. In these cases we assign a single structure to each processor (*P=N*), then we require *(N/P+2)=3* force calculations per algorithm step. The number of the force calculations could be reduced to one per step, if we allow for additional communication of $\partial U/\partial\mathbf{x}_{j\pm1}$ from neighboring processors. This reduces the

number of forces calculations per step per processor from $(N/P+2)$ to N/P and increases the amount of communication to $6\cdot 3n_{pt}$. The actual algorithm as implemented in MOIL uses slightly different reduction of Equation (A.1), which can be rewritten as

$$\frac{\partial S^{l}_{Gauss}}{\partial x_{jm}} = \mathbf{G}'_{m}\left(\mathbf{x}_{j\pm1}, \mathbf{x}_{j}, \partial\overline{S}/\partial\mathbf{x}_{j\pm1}, \partial U/\partial\mathbf{x}_{j\pm1}, \partial U/\partial\mathbf{x}_{j}, \partial^{2}U/\partial\mathbf{x}_{j}^{2}, U_{j\pm1}, U_{j}\right), \qquad (A.2)$$

where instead of $\mathbf{x}_{j\pm2}, U_{j\pm2}$, derivatives $\partial\overline{S}/\partial\mathbf{x}_{j\pm1}$ are used. This solution is in terms of computation, memory, and communication equivalent (up to negligible constants) to the former one; the advantage is that it can be implemented without requiring extra special cases in the code for the first and last processors and the code is more easily generalizable for the case $\mathrm{mod}(P,N)\neq 0$.

Additional reduction in the computation time can be obtained by transforming the problem of Hessian matrix computation to an additional force computation. This can be done because the Hessian $\partial^{2}U/\partial\mathbf{x}_{j}^{2}$ is used in calculation of $\partial T/\partial\mathbf{x}_{j}$ only for a multiplication with some other vector $\mathbf{v}$. The following first-order approximate reduction can be used to compute the product $\partial^{2}U/\partial\mathbf{x}_{j}^{2}\cdot\mathbf{v}$ (Eric Vanden Eijnden, private communication):

$$\left.\frac{\partial^{2}U}{\partial\mathbf{x}^{2}}\right|_{\mathbf{x}_{j}}\cdot\mathbf{v} \approx 1/\alpha\left(\left.\frac{\partial U}{\partial\mathbf{x}}\right|_{\mathbf{x}_{j}+\alpha\mathbf{v}} - \left.\frac{\partial U}{\partial\mathbf{x}}\right|_{\mathbf{x}_{j}}\right), \qquad (A.3)$$

The expression in Equation (A.3) becomes accurate for a sufficiently small scalar α. For the MOIL potential energy function with implicit solvent modeling, the calculation of Hessian matrix is approximately 50% more expensive than the calculation of the forces. However, the benefit of introducing this approximate reduction is not only in those 50% of run time, it also makes the code simpler and more understandable, since the formulas for the Hessian calculation is significantly

more complex than that for the forces calculation. Derivatives of other terms in function T (Equation (2.7)) can be computed from local information kept on each processor.

An Additional piece of global information required on each processor is $\langle \Delta l \rangle$, which is a slowly varying function of the number of optimization steps. Therefore, it suffices to re-compute $\langle \Delta l \rangle$ every 10 to 20 steps of simulated annealing. The parallel computation of $\langle \Delta l \rangle$ can be done classically in $2\log K$ communication rounds with a total number of $2P$ (single floating-point number) messages passed.

APPENDIX B: EXPLICIT EXPRESIONS FOR THE SDEL ACTION

The exact formulas for SDEL derivatives

$$\frac{\partial \overline{S}}{\partial x_{im}} = \frac{1}{2}\left(\frac{-\partial U/\partial x_{im}}{p_i}\left(\Delta l_{i,i+1} + \Delta l_{i-1,i}\right) + \left(p_i + p_{i+1}\right)\frac{x_{im} - x_{i+1m}}{\Delta l_{i,i+1}} + \left(p_{i-1} + p_i\right)\frac{x_{im} - x_{i-1m}}{\Delta l_{i-1,i}} \right)$$

$$p_i = \sqrt{2\left(E - U\left(\mathbf{x}_i\right)\right)}$$

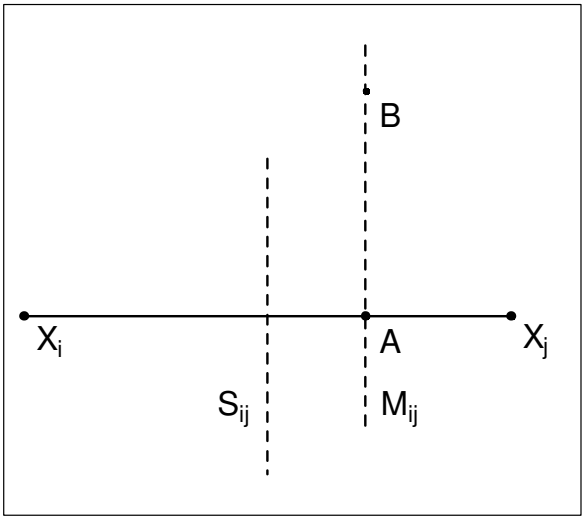
$$\frac{\partial^2 \overline{S}}{\partial x_{im}\partial x_{ik}} = \frac{1}{2}\left(\begin{array}{l} -\left(\frac{\partial^2 U/\partial x_{im}\partial x_{ik}}{p_i} + \frac{\partial U/\partial x_{im}\,\partial U/\partial x_{ik}}{p_i^3} \right)\left(\Delta l_{i,i+1} + \Delta l_{i-1,i}\right) \\ -\frac{\partial U/\partial x_{im}}{p_i}\left(\frac{x_{ik} - x_{i+1k}}{\Delta l_{i,i+1}} + \frac{x_{ik} - x_{i-1k}}{\Delta l_{i-1,i}} \right) - \frac{\partial U/\partial x_{ik}}{p_i}\left(\frac{x_{im} - x_{i+1m}}{\Delta l_{i,i+1}} + \frac{x_{im} - x_{i-1m}}{\Delta l_{i-1,i}} \right) \\ +\delta_{km}\left(\frac{\left(p_i + p_{i+1}\right)}{\Delta l_{i,i+1}} + \frac{\left(p_{i-1} + p_i\right)}{\Delta l_{i-1,i}} \right) \\ -\left(p_i + p_{i+1}\right)\frac{\left(x_{im} - x_{i+1m}\right)\left(x_{ik} - x_{i+1k}\right)}{\left(\Delta l_{i,i+1}\right)^3} - \left(p_{i-1} + p_i\right)\frac{\left(x_{im} - x_{i-1m}\right)\left(x_{ik} - x_{i-1k}\right)}{\left(\Delta l_{i-1,i}\right)^3} \end{array} \right)$$

$$\frac{\partial^2 \overline{S}}{\partial x_{i-1k}\partial x_{im}} = \frac{1}{2}\left(\begin{array}{l} -\frac{\partial U/\partial x_{im}}{p_i}\frac{x_{i-1k} - x_{ik}}{\Delta l_{i,i+1}\Delta l_{i-1,i}} - \frac{\partial U/\partial x_{i-1k}}{p_{i-1}}\frac{x_{im} - x_{i-1m}}{\Delta l_{i,i+1}\Delta l_{i-1,i}} \\ -\frac{p_{i-1} + p_i}{\Delta l_{i,i+1}\Delta l_{i-1,i}}\delta_{km} + \left(p_{i-1} + p_i\right)\frac{\left(x_{im} - x_{i-1m}\right)\left(x_{ik} - x_{i-1k}\right)}{\left(\Delta l_{i-1,i}\right)^3} \end{array} \right)$$

$$\frac{\partial^2 \overline{S}}{\partial x_{i+1k}\partial x_{im}} = \frac{1}{2}\left(\begin{array}{l} -\frac{\partial U/\partial x_{im}}{p_i}\frac{x_{i+1k} - x_{ik}}{\Delta l_{i,i+1}\Delta l_{i-1,i}} - \frac{\partial U/\partial x_{i+1k}}{p_{i+1}}\frac{x_{im} - x_{i-1m}}{\Delta l_{i,i+1}\Delta l_{i-1,i}} \\ -\frac{p_{i+1} + p_i}{\Delta l_{i,i+1}\Delta l_{i-1,i}}\delta_{km} + \left(p_{i-1} + p_i\right)\frac{\left(x_{im} - x_{i-1m}\right)\left(x_{ik} - x_{i+1k}\right)}{\left(\Delta l_{i,i+1}\right)^3} \end{array} \right)$$

$$\partial S^l_{Gauss} \Big/ \partial x_{jk} \approx \sum_{i,m} \frac{\partial\left(\partial \overline{S} \Big/ \partial x_{im}\right)^2}{\partial x_{jk}} = 2\sum_{i,m}\left(\partial \overline{S} \Big/ \partial x_{im}\right)\left(\partial^2 \overline{S} \Big/ \partial x_{im}\partial x_{jk}\right)$$

APPENDIX C: LEMMAS REGARDING THE MILESTONES GEOMETRY



Lemma C.1: Let X_i and X_j be two images in conformation space such that $M_{i\to j}$ exists. Let *A* be an intersection of the line segment X_iX_j with $M_{i\to j}$. Then a point B on the hyperplane perpendicular to X_iX_j and passing through A belongs to $M_{i\to j}$ iff $\forall k \quad d(X_k,B) \geq d(X_j,B)$.

Proof of Lemma C.1: From definition (4.1) of $M_{i\to j}$:

$$d(X_i,A)^2 - d(X_j,A)^2 = \Delta_i^2$$

By using the Pythagoras theorem for triangles X_iAB and X_jAB:

$$\begin{aligned} d(X_i,B)^2 - d(X_j,B)^2 &= \left(d(X_i,A)^2 + d(A,B)^2\right) - \left(d(X_j,A)^2 + d(A,B)^2\right) \\ &= d(X_i,A)^2 - d(X_j,A)^2 = \Delta_i^2 \end{aligned}$$

□

Consequence of Lemma C.1: $M_{i\to j}$ is a hyperplane segment perpendicular to X_iX_j.

Lemma C.2: Let S_{ij} be the hyperplane perpendicular to the line segment $X_i X_j$ and passing though its midpoint. Then $d(S_{ij}, M_{i\to j}) = \dfrac{\Delta_i^2}{2d(X_i, X_j)}$.

Proof of Lemma C.2: Since both S_{ij} and $M_{i\to j}$ are perpendicular to $X_i X_j$ the distance $d(S_{ij}, M_{i\to j})$ is equal to the distance of the $X_i X_j$ midpoint, P_{ij}, and the intersect of $M_{i\to j}$ with X_{ij}, A. Thus:

$$\left(d(X_i, P_{ij}) + d(S_{ij}, M_{i\to j})\right)^2 - \left(d(X_j, P_{ij}) - d(S_{ij}, M_{i\to j})\right)^2 = \Delta_i^2$$

$$d(S_{ij}, M_{i\to j}) = \frac{\Delta_i^2}{4d(X_i, P_{ij})} = \frac{\Delta_i^2}{2d(X_i, X_j)}$$

□

APPENDIX D: STATISTICAL REASONING

We describe an estimate of the statistical error of a milestoning calculation from a single set of collected data using Bayesian reasoning. As shown in Section 4.2, equation (4.5), repeated here as (D.1),

$$\left\langle \tau_{\alpha\beta} \right\rangle = \left\langle t_\alpha \right\rangle + \sum_\gamma P(\gamma \mid \alpha) \left\langle \tau_{\gamma\beta} \right\rangle \tag{D.1}$$

relates MFPTs ($\left\langle \tau_{\alpha\beta} \right\rangle$) and local kinetics entities ($\left\langle t_\alpha \right\rangle$ and $P(\gamma \mid \alpha)$). Milestoning aims to estimate $\left\langle t_\alpha \right\rangle$ and $P(\gamma \mid \alpha)$ by launching N_α trajectories from each Milestone α. $N_{\alpha\gamma}$ of them terminate on the Milestone γ and the mean incubation time (time to termination) of all N_α trajectories is T_α. In Bayesian inference a statistical model of the transitions among Milestones is needed. We closely follow and extend notation used in the analysis of Markovian Milestoning with Voronoi Tesselations (Vanden-Eijnden and Venturoli 2009). The same kinetic formulas (with different notation) are also available from (West, Elber et al. 2007). We assume continuous Markov jump process between the Milestones controlled by a transition matrix Q defined in the following way: Let the probability distribution of the system over all the Milestones be $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_N)$, where ρ_α is the probability that the system is assigned to a Milestone α. Under continuous Markov jump process $\boldsymbol{\rho}$ behaves as:

$$\dot{\boldsymbol{\rho}} = \boldsymbol{\rho} Q \,. \tag{D.2}$$

For transition matrix Q, by definition $q_{\alpha\alpha} = -\sum_{\beta \neq \alpha} q_{\alpha\beta}$ and it can be shown by simple algebra that $P(\beta \mid \alpha) = q_{\beta\alpha} / \sum_{\gamma \neq \alpha} q_{\gamma\alpha}$ and $\left\langle t_\alpha \right\rangle = 1 / \sum_{\gamma \neq \alpha} q_{\gamma\alpha}$ (for derivation see for example (Shalloway and Faradjian 2006; West, Elber et al. 2007; Vanden-Eijnden and Venturoli 2009)). By plugging the last three identities to the linear system (D.1) it reduces to

$$Q' \left\langle \boldsymbol{\tau} \right\rangle = -\mathbf{1}, \tag{D.3}$$

where $\langle \boldsymbol{\tau} \rangle$ is the row vector $\left(\langle \tau_{1\beta} \rangle, \ldots, \langle \tau_{\beta-1\beta} \rangle, \langle \tau_{\beta+1\beta} \rangle, \ldots, \langle \tau_{N\beta} \rangle\right)^T$ and Q' is a $(N-1)\times(N-1)$ matrix created from Q by skipping the row and the column related to the Milestone β. In order to infer $\langle \boldsymbol{\tau} \rangle$ from the collected data, $\{N_{\alpha\gamma}, T_\alpha\}$, using Eq. (D.3), a relation between $\{N_{\alpha\gamma}, T_\alpha\}$ and Q' is needed. Following the derivations from ref. (Vanden-Eijnden and Venturoli 2009): for a system ruled by (D.2) the probability of staying in a state α for time *t* and then jumping to a state β in the time interval $< t, t+dt >$ is $e^{-\sum_{\gamma\neq\alpha} q_{\alpha\gamma} t} q_{\alpha\beta}\, dt$. Using this equality, the likelihood of observing the collected data, $L(\{N_{\alpha\gamma}, T_\alpha\} \mid Q)$, is

$$L(\{N_{\alpha\gamma}, T_\alpha\} \mid Q) = \prod_\alpha \prod_{\gamma\neq\alpha} q_{\alpha\gamma}^{N_{\alpha\gamma}} e^{-q_{\alpha\gamma} N_\alpha T_\alpha} . \tag{D.4}$$

By using the Bayes' rule the likelihood that the true transition matrix is *Q* given the collected data, $L(Q \mid \{N_{\alpha\gamma}, T_\alpha\})$, is:

$$L(Q \mid \{N_{\alpha\gamma}, T_\alpha\}) \propto \prod_\alpha \prod_{\gamma\neq\alpha} q_{\alpha\gamma}^{N_{\alpha\gamma}} e^{-q_{\alpha\gamma} N_\alpha T_\alpha} P(Q) , \tag{D.5}$$

where *P(Q)* is the prior probability distribution of *Q* without seeing any data (typically this is set to uniform if we do not have any prior knowledge about the system). Equality (D.5) is typically used in maximum likelihood estimators, e.g. one estimates unknown entity *Q* with Q^*, the matrix that maximizes likelihood $L(Q \mid \{N_{\alpha\gamma}, T_\alpha\})$. In this particular case, Q^* has form $q^*_{\alpha\gamma} = N_{\alpha\gamma} / [N_\alpha T_\alpha]$, what is in agreement with estimators given in equation (4.6) in the main text. Instead of using purely Q^* for calculations of MFPTs we can examine whole distribution of transition matrices according to equation (D.5) and understand what is the distribution of MFPTs consistent with the data collected. Therefore we typically sample number of (typically 300) transition matrices from distribution (D.5) and look at the variance of MFPTs predicted by them. If standard deviation of MFPTs is large it suggests that more data

about the system shall be collected. We report standard deviation obtained by this algorithm in the results of Section 4.3.

APPENDIX E: SAMPLING EQUILIBRIUM DISTRIBUTION ON A MILESTONE

As described in Section 4.2.4 the equilibrium ensemble from a Milestone $M_{i\to j}$ is used to sample the first hitting point distribution on the Milestone $M_{i\to j}$. The Milestone $M_{i\to j}$ is defined in equation (4.1) as

$$M_{i\to j} \equiv \left\{ X \mid d(X,X_i)^2 = d(X,X_j)^2 + \Delta_i^2 \text{ and } \forall k\ d(X,X_j) \le d(X,X_k) \right\},$$

where $\{X_1,\ldots,X_K\}$ is a set of images in the conformational space. In practice we work with the following approximation of $M_{i\to j}$:

$$\begin{aligned} d_{ij}(X) &\equiv d(X,X_j)^2 - d(X,X_i)^2 + \Delta_i^2 \\ M'_{i\to j} &\equiv \left\{ X \mid \forall k, d(X,X_k) \ge d(X,X_j) \wedge -\lambda \le d_{ij}(X) \le 0 \right\} \end{aligned} \tag{E.1}$$

Clearly as $\lambda \to 0$, $M'_{i\to j}$ converges to $M_{i\to j}$. We have used $\lambda = 0.5^\circ$ or $\lambda = 0.01$ Å for the calculations on alanine dipeptide.

To sample conformations in $M'_{i\to j}$ from equilibrium distribution the following Umbrella Sampling protocol is employed. We run NVT trajectory of the system (using Andersen thermostat) with a modified potential function U and examine a conformation every few steps (every 100 – 400 fs for examples described in this chapter). If an examined conformation belongs to $M'_{i\to j}$ it is saved; otherwise it is discarded. If conformation is saved, corresponding velocities are sampled from Boltzmann distribution. The potential function U is modified to bias the system towards the region $M'_{i\to j}$ in the following way:

$$U'(X) = U(X) + U_{ij}^{1}(X) + U_{ij}^{2}(X)$$

$$U_{ij}^{1}(X) = \begin{cases} K_1 d_{ij}(X)^2 & \text{if } d_{ij}(X) > 0 \\ K_1 \left(d_{ij}(X) - \lambda \right)^2 & \text{if } d_{ij}(X) < -\lambda \\ 0 & \text{otherwise} \end{cases}$$

$$U_{ij}^{1}(X) = \begin{cases} K_2 \left(d(X, X_k) - d(X, X_j) \right)^2 & \text{if } d(X, X_k) < d(X, X_j) \\ 0 & \text{otherwise} \end{cases}$$

By definition for $X \in M'_{i \to j}$, $U'(X) = U(X)$ and therefore saved points from $M'_{i \to j}$ are sampled with the true equilibrium probabilities. If on the other hand NVT trajectory of the system is outside of the region $M'_{i \to j}$, the terms U_{ij}^{1} and/or U_{ij}^{2} force the system to return back to $M'_{i \to j}$, the strength of this bias is controlled by force constants K_1 and K_2 (both are set to 10^3 Kcal mol^{-1} rad^{-2} or 10^4 Kcal mol^{-1} Å^{-2} for alanine dipeptide system).