Optimizing the performance of Bias Exchange Metadynamics for Protein Folding

by

Pilar Cossio

Submitted in total fulfilment of
the requirements for the degree of

Master of Philosophy

Director: Ph. D. Alessandro Laio

Statistical Biophysics Sector
SISSA/ISAS

October, 2008

ABSTRACT

Computer simulation of conformational transitions in biomolecules, such as protein folding, is considered one of the main goals of computational chemistry. Due to the complexity of the systems, a quantitative description can only be provided at the price of a significant computational cost. A powerful methodology, called bias exchange metadynamics (BE-META) [16], has been recently developed. The approach combines replica exchange [15] with metadynamics [14], and allows exploring the free energy landscape of complex systems, like biomolecules. The primary objective of this thesis is to improve further this promising technique. This will be accomplished by searching for the optimal set of parameters (e.g. collective variables and exchange time) that enable the folding of a small protein 1E0G (48 amino acids) in the shortest possible time, using a coarse-grain force field (UNRES [18]). It will be shown that BE-META allows the accurate reconstruction of the folding free energies of 1E0G, with a small computational effort in comparison with other techniques like MREMD, and that a suffcient number of collective variables are necessary to increase the capability of each replica to diffuse through the conformational space.

ACKNOWLEDGMENTS

First of all, I would like to thank my thesis' director Alessandro Laio for teaching me the methods he knows and he has created, for his amiability and also for proposing me an interesting research problem for us to solve. I feel very proud to be his student.

I feel grateful with the Statistical Biophysics Sector for giving me the opportunity to fulfill my Ph.D studies at S.I.S.S.A, for the interesting lectures, and for the diversity of thoughts and discussions that we have had during this year. In particular, I would like to thank Fabio Pietrucci and Fabrizio Marinelli, for their patience and for the help they have given me with my research problem, also for the enlightening discussions and their good ideas. I know that without their help, I could not of learned and done as much as I present in this work.

For the finacial support and for giving me the possibility to work all day in a research project, for having very smart people helping me with the project, and for the technical facilities, I thank S.I.S.S.A. All of this has made the work environment very pleasent.

I thank my family for their unconditional love, for giving me all that I have, and for being the example I want to follow.

Last, I thank my friends for the very good times we have had here, and for bringing a lot of happiness into my life.

CONTENTS

LIST OF FIGURES

INTRODUCTION

A popular approach for studying biomolecules is to perform a molecular dynamics simulation with an all-atom force field describing explicitly both the protein and the solvent. Unfortunately, this description is computationally expensive. Nowadays direct simulation can access at most the microsecond time scale, which remains far from those that would be necessary to provide truly quantitative and statistically reliable predictions in protein folding. As a consequence, the progress of simulations still lags behind that of experiments. Decades of experimental work have led to a deep understanding of the nature of the folding process, and this knowledge is now summarized in several textbooks. While thousands of new protein structures are resolved every year using x-rays or NMR, the prediction of a protein fold by atomistic modeling is still considered to be extremely challenging [1], [2], [3].

An alternative to brute-force molecular dynamics is to rely on some methodology that is capable to accelerate rare events and that permits a faster exploration of the conformational space. Several methods have been developed with this scope in the last years, and important advances have been made in several fields, ranging from solid state physics to quantum chemistry [4], [5], [6], [7]. Most of these techniques are unfortunately only partially useful for biophysical applications. The efficiency of the methods based on a search on the potential energy surface [6], [8], [9] is hindered by the enormous amount of degrees of freedom involved. In the simulation of a normal size protein with explicit water one has to treat explicitly $20,000$ atoms, thus a 10^4-dimensional configuration space, which is far too large to be explored completely. What is commonly done is to reduce the dimensionality of the system into a few collective variables that are assumed to provide a coarse-grained description, and to explore the free energy surface as a function of these variables. The methods based on the exploration of a single or a few reaction coordinates, like WHAM [10], [11] and thermodynamic integration [4], [12] are of limited use in protein folding, since the phenomena involved in a biological process usually involve the concerted or sequential motion of several independent degrees of freedom. If an important degree of freedom is not considered explicitly, the simulation can show hysteresis effects and poor reproducibility of the results. Also history-dependent search methods, such as local elevation [7], Wang-Landau sampling [13], and metadynamics [14], allow a free energy reconstruction only as a function of a few variables. This is due to the fact that their performance deteriorates rapidly with dimensionality.

The only methodology that seems to offer a general route for studying complex rearrangements, such as protein folding, is the so-called replica exchange method (or parallel tempering, or multicanonical ensemble method [5], [15]). Applications to large biological systems can, however, be computationally very intensive. When a biomolecule is immersed in a solvent, the structure of the density of states imposes the use of a great amount of replicas even for small systems. This has so far limited the scope of this, otherwise extremely powerful, methodology.

In a recent paper [16] another technique for studying protein folding was introduced, called ”bias exchange metadynamics” (BE-META). The approach is based on the combined use of replica exchange [5], [15] and metadynamics [14], and is particularly efficient in accelerating rare events in biomolecules. In this method, a large set of collective variables is chosen and several metadynamics simulations are performed in parallel, biasing each replica with a time-dependent potential acting on just one or two of the collective variables. Exchanges between the bias potentials in the different variables are periodically allowed according to a replica exchange scheme. Due to the efficaciously multidimensional nature of the bias, the method allows exploring complex free energy landscapes with great efficiency. Applying this methodology, with a relatively moderate computational effort, it was possible to reversibly fold some small proteins (Triptophane cage, Villin and Advillin) [16], the structure of a mutant of Advillin that had previously never been investigated was predicted in [17], and the numerical prediction was successively validated by NMR experiments.

Before attempting to use this method to fold an average-size protein, it is necessary to understand in detail how the BE-META performance can be optimized. These are the most important questions that have to be addressed:

- What are the collective variables that are optimal for describing the folding of a generic protein?
- How many variables are necessary? Does this number depend on the size of the protein?
- How often should the exchanges be attempted?

The objective of this thesis is to address these issues in a systematic manner. This is done by using BE-META together with a coarse-grained force field with implicit solvent (UNRES) developed by Sheraga at al. in [18], [19]. UNRES is an united-residue force field, that has been carefully optimized, and is based on the physics of interactions, not on the native structure of the system. The greatest advantage of a simulation performed with a coarse-grain force field and BE-META is that since the computational cost is orders of magnitude smaller than using an all-atom force field, the protocol allows to perform the large number of simulations which are necessary to address the issues outlined above in a systematic manner.

The final goal of this research line will be to fold an average size protein using an all atom simulation with an explicit solvent description. It can be anticipated that this will require the use of very significant computational resources. Only with an optimal choice of collective variables and exchange rate this goal is expected to be fulfilled. Thus, the preliminary optimization of the BE-META parameters and a new set of better collective variables are essential.

In Chapter 1, the theoretical methods used in this work are discussed. A brief introduction to some of the methods used to calculate free energy profiles and accelerate rare events is also given. A full description of the metadynamics [14] methodology is given. Then, bias exchange metadynamics [16] is explained. Next, the coarse grain force field (UNRES) [18], used to describe the physics of the interactions between the amino acids in the protein is explained. Then, the implementation of bias exchange metadynamics in the molecular dynamics UNRES program is described. In the last section of this chapter the conventional collective variables that are used for protein folding are described, and a proposal of a new set of collective variables to describe proteins is made, starting from the contact matrix.

In the second chapter, the results of applying BE-META with the UNRES force field, for the folding of protein 1E0G (48 amino acids, [42]) with different sets of collective variables and bias exchange parameters, are shown. Next, the statistical results of a number of different short bias exchange simulations are given. Finally, a cluster search comparison between the long simulations is done. That allows finding the optimal and most efficient parameters.

The thesis ends with some conclusions and perspectives for future work.

1 THEORETICAL BACKGROUND

Computer simulations of biological systems offer the possibility of investigating the most basic mechanisms of life, but pose a formidable challenge to theoreticians, due to the level of complexity that arises from both the dimension of biomolecules and their heterogeneity. As a result, computer simulations are nowadays predictive only for phenomena that take place on a relatively short time scale or in a limited region of space. Major conformational changes that involve very complex free energy landscapes, like gating in ion channels, protein-protein interaction, and protein folding, are still out of reach to a direct atomistic simulation. This is due to the fact that the system gets trapped in local free energy minima and that there are not enough computational resources to allow the system to explore all the conformational space and find the global minima, which is for example the folded state of the protein. An alternative to brute-force molecular dynamics is to rely on some methodology that is capable to accelerate rare events and that allow a faster exploration of the conformational space. In the following an introducction to molecular dynamics simulations will be given, and the concepts of free energy and rare events will be introduced. In particular, Metadynamics and Bias Exchange Metadynamics, which is the method that we will use to do protein folding, will be fully described. At the end of this chapter the collective variables previously used to describe a protein's topology in bias exchange metadynamics will be explained.

1.1 MOLECULAR DYNAMICS

Molecular dynamics (MD) is a form of computer simulation in which atoms and molecules are evolved, for a period of time, under the action of a potential that provides approximations of known physics. Because molecular systems generally consist of a vast number of particles, it is impossible to find the properties of such complex systems analytically. MD simulation circumvents this problem by using numerical methods.

In a normal MD simulation, given a certain force field U and the positions and velocities of the particles $\vec{r}(t), \vec{v}(t)$ at time t, the accelerations over the particles are computed using $\vec{a} = -\vec{\nabla} U/m$, and then the equations of motion are integrated at a certain time step (Δt) as to find the final positions $\vec{r}(t+\Delta t)$ and final velocities $\vec{v}(t+\Delta t)$. There are many types of integrators [38], [41], the most commonly used is the Velocity Verlet [39] algorithm that calculates the final coordinates as follows:

$$\vec{r}(t+\Delta t) = \vec{r}(t) + \vec{v}(t)\Delta t + \frac{1}{2}\vec{a}(t)\Delta t^2, \tag{1.1}$$

$$\vec{v}(t+\Delta t) = \vec{v}(t) + \frac{\vec{a}(t) + \vec{a}(t+\Delta t)}{2}\Delta t. \tag{1.2}$$

This procedure is done iteratively and the system is evolved in time. What is important here is the value of the time step Δt. If it is too large the results will not be realistic and the simulation will crash. The appropriate Δt for atomistic simulations is usually of the order of $1fs$. Thus, if one wants to find the folded state of a small protein (folding time $10\mu s$) using a normal molecular dynamics simulation one should compute the forces 10^{10} times. This, with the computational resources that are nowadays available, is an almost unreachable amount of CPU time. This is why it is better to rely on others methodologies that can accelerate rare events.

1.2 RARE EVENTS

1.2.1 Metastability and Dimensional Reduction

Let us consider a system of particles of coordinates x, in space Ω, where x can include the normal coordinates ($\vec{r}$) but also generalized coordinates such as the box edge in Parrinello Rahman [20] or the volume. The system evolves under the influence of a potential $V(x)$ and is coupled to a thermostat bath of temperature T. According to the laws of thermodynamics, the system evolves following the canonical equilibrium distribution [21]

$$P(x) = \frac{1}{Z} e^{-\beta V(x)}, \tag{1.3}$$

where β is the inverse temperature and $Z = \int dx e^{-\beta V(x)}$ is the partition function of the system.

In normal biological systems, like proteins, there are of the order of 10^4 atoms. Thus $P(x)$ has an incredibly large dimensionality. What is done to overcome this problem is to consider the reduced probability distributions in terms of some collective variables or reaction coordinates $s(x)$. Namely, instead of monitoring the full trajectory $x(t)$ of the system, a reduced trajectory $s(t) = s(x(t))$ is analyzed. For an infinitely long trajectory the probability distribution $P(s)$ is given by the histogram of s:

$$P(s) = lim_{t \to \infty} \frac{1}{t} \int dt \delta(s - s(t)), \tag{1.4}$$

in real applications $P(s)$ is estimated as

$$P(s) \approx \frac{1}{n \Delta s} \sum_{t=1}^{n} \chi_s(s(t)), \tag{1.5}$$

where $\chi_s(x) = 1$ if $x \in [s, s + \Delta s]$ and zero otherwise. If the system is ergodic and the dynamics allows an equilibrium distribution at an inverse temperature β, the knowledge of $P(s)$ allows to define the free energy of the system in terms of s:

$$F(s) = -\frac{1}{\beta} ln(P(s)). \tag{1.6}$$

Qualitatively, a system will display metastability if the probability $P(s)$ is large in a set of disconnected regions A_i separated by regions in which the probability is low. A system is to be considered metastable if $F(s)$ has a characteristic shape with wells and barriers and if it presents more than one sharp minimum in its free energy profile (see Figure 1.1). A metastable system resides for the big majority of time in disconnected regions of the space and it will take the system a long time to go from one minimum to another.

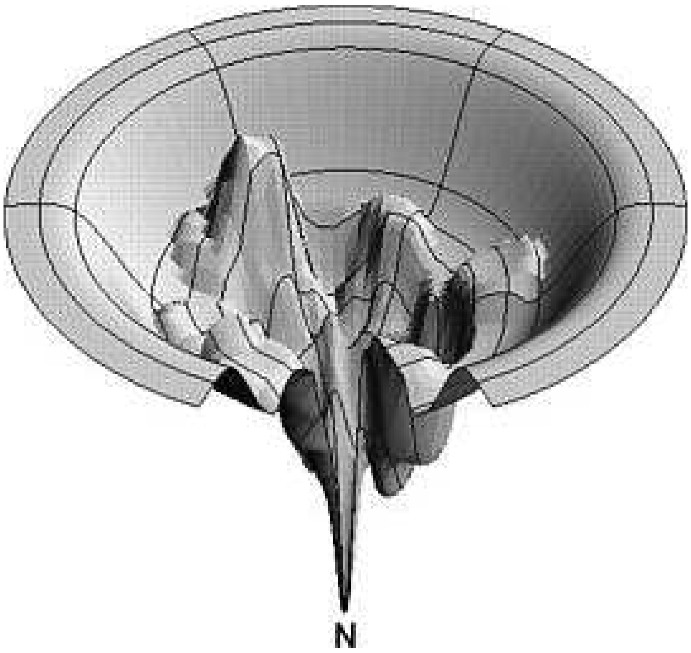


Figure 1.1: Free Energy profile of a metastable system.

1.2.2 Methods for Computing Free Energy

The free energy as a function of a relevant and smartly chosen set of variables provides a very important insight in the equilibrium and metastability properties of the system. For instance, the minima in a free energy surface correspond approximately to the metastable sets of a system: the system spends by definition a lot of time in the minima and only rarely it visits the barrier regions in between. The free energy profiles can be used to estimate the transition time between two metastable states and can give accurate estimations of interaction energies. For instance, in chemistry one can estimate the free energy needed to break a bond in a chemical reaction by using as a collective variable the distance between two atoms and studying its free energy profile. In the past decades, different methods for computing free energy profiles have been developed. In the following, some of the principal methods will be explained.

Umbrella sampling [22] is a commonly used method in which the normal dynamics of the system is biased by a suitably chosen bias potential $V^B(s(x))$ that depends on x only via $s(x)$. The new biased probability distribution is

$$P^B(x) = \frac{Z}{Z^B} e^{-\beta(V^B(s(x))+V(x))}, \tag{1.7}$$

where Z^B is the canonical partition function for the potential $V(x) + V^B(x)$. So, measuring a probability distribution in the presence of a bias $V^B(s(x))$ will provide a measure for the unbiased free energy and for the unbiased probability distribution. It can be shown [22] that the optimal choice for the biased potential is $V^B(s(x)) = -F(s)$ but in real systems $F(s)$ is not known, so the main problem that arises is how to construct $V^B(s(x))$ without a detailed knowledge of the system.

In order to solve this problem, an efficient strategy to apply is the weighted histogram method (WHAM) [10], [11], in which several histograms, constructed with different umbrellas $V^{B_i}(s(x))$, are combined in order to reconstruct a single estimate of $F(s)$. A typical bias potential would be of the form $V^{B_i}(s) = \frac{1}{2}k(s - s_i)^2$. The principal problem with this method is that the number of biasing potentials, that one has to use, scales exponentially with the dimensionality, so the computational price becomes unbearable in $d > 2$.

These methodologies are based on studying the properties of the system in equilibrium. Lately, some new algorithms have been generated to exploit non equilibrium dynamics in order to compute equilibrium observables. One of these methods makes use of Jarzynski's equality [23]

$$< e^{-\beta W_t} >= e^{-\beta \Delta F}, \tag{1.8}$$

where W_t is the work performed on the system in a trajectory of time duration t. This equation provides an explicit expression for the free energy of the system in terms of the average of the exponential of the work performed on it. The main problem of this method is that the average value of $e^{-\beta W}$ is dominated by the rare trajectories for which W is small and thus $e^{-\beta W}$ is large. This hinders accuracy especially if the time duration of the trajectory is short.

In the next section we will explain a powerful methodology called Metadynamics [14], in which the dynamics of the system is biased with a history-dependent potential that brings the system out of equilibrium but provides a full description of the system's free energy.

1.3 METADYNAMICS

In metadynamics, the dynamics in the space of the chosen CVs is driven by the free energy of the system and is biased by a history-dependent potential, $F_G(s,t)$ constructed as a sum of Gaussians centered along the trajectory followed by the collective variables up to time t. This history-dependent potential is expressed as

$$F_G(s(x),t) = \frac{w}{\tau_G} \int_0^t e^{-\frac{(s(x)-s(x(t')))^2}{2\delta s^2}} dt', \tag{1.9}$$

where τ_G is the rate at which the Gaussians are introduced and $w, \delta s$ represent the height and width of the Gaussian, respectively. In real simulations it can be calculated as

$$F_G(s(x),t) = w \sum_{t'<t} e^{-\frac{(s(x)-s(x(t')))^2}{2\delta s^2}}. \tag{1.10}$$

In the Monte Carlo or molecular dynamics simulation this bias is added to the normal potential of the system. The force generated by this biasing potential will discourage the system from revisiting the same spot and encourage an efficient exploration of the free energy surface (FES). Since the history-dependent potential iteratively compensates the underlying free energy, a system evolving with metadynamics will tend to escape from any free energy minimum through the lowest free energy saddle point. As the system diffuses on the FES, the Gaussian potentials accumulate and fill the FES wells, which permits the system to migrate, in a short time, from well to well through the lowest free energy saddle point. An example of a free energy profile filled by this biasing potential is shown in Figure 1.2.

After a while, the sum of the Gaussian terms will almost exactly compensate the underlying FES. So, for long t,

$$lim_{t\to\infty} F_G(s(x),t) \approx -F(s), \tag{1.11}$$

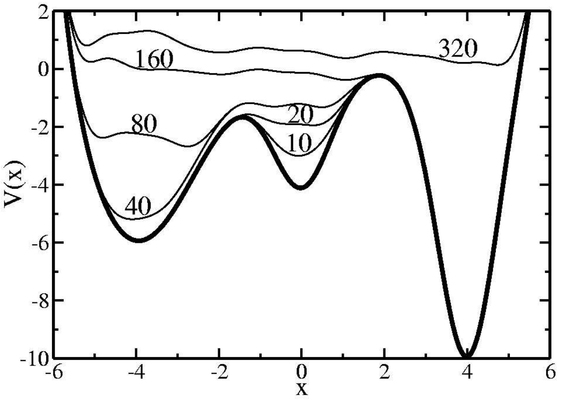


Figure 1.2: Free Energy profile filled by the biasing metadynamics potential.

this property does not derive from any ordinary thermodynamic identity, since the metadynamics is a non-equilibrium process. What makes it a flexible tool is that it can be used not only for efficiently computing the free energy but also for exploring new reaction pathways and accelerating rare events.

The problem of working with history-dependent dynamics is that the forces (or the transition probabilities) on the system depend explicitly on its history. Hence it is not a priori clear if, and in which sense, the system can reach a stationary state under the action of these dynamics. As follows, the validity of metadynamics will be demonstrated rigorously by introducing a mapping of the history-dependent evolution into a Markovian process in the original variable and in an auxiliary field that keeps track of the configurations visited. This will be done assuming that the system of study follows a Langevin type dynamics, but the same formalism can be extended to Monte Carlo like samplings such as Wang-Landau and, more generally, to all stochastic processes augmented by a history-dependent term which is an explicit function of the system's trajectory. In particular, it will be shown that the average over several independent simulations of the metadynamics biasing potential is exactly equal to the negative of the free energy (Eq. 1.11), and an explicit expression for the standard deviation will be found.

For Langevin dynamics it is possible to solve analytically the equilibrium distribution of the system. In this case the evolution of the CVs is given by

$$ds = -D\frac{\delta F(s)}{\delta s}\bigg|_{s=s(t)} dt + \sqrt{2D}dW, \tag{1.12}$$

where $ds = s(t+dt) - s(t)$, dW is a Wigner noise, D is a diffusion coefficient and the energies are measured in units of temperature. Let us also assume that the system is confined in a region Ω, and that the dynamics satisfies reflecting boundary conditions on $\partial\Omega$. In metadynamics, the normal equilibrium dynamics of the CVs is biased by an history-dependent term that discourages the system from revisiting positions in the CV space that has already been explored. The equation that shows the evolution of the CVs under the influence of the metadynamics biasing potential becomes

$$ds = -D\frac{\delta}{\delta s}\Big[F(s) + \int dt' g(s, s(t'))\Big]_{s=s(t)} dt + \sqrt{2D}dW, \tag{1.13}$$

where the time integral is an history-dependent potential, generated through the kernel $g(s, s')$, which so far it has been taken to be a Gaussian with the form

$$g(s, s(t')) = \frac{w}{\tau_G} e^{-\frac{(s(x)-s(x(t')))^2}{2\delta s^2}}, \tag{1.14}$$

but different kernels can be considered.

What will be further proven is that for large enough t, $\int dt' g(s, s(t'))$ is an unbiased estimator of $F(s)$, namely

$$< \int dt' g(s, s(t')) >= F(s), \tag{1.15}$$

where the average is done over several realizations of the dynamic process. To prove this, the kernel $g(s, s(t'))$ is required to be such that there exists a function $\phi_0(s)$ such that the equation

$$\int ds' g(s, s')\phi_0(s') + F(s) = 0 \tag{1.16}$$

has a solution, namely, it is required that the free energy can be representable as a convolution with the kernel g.

In order to study the average properties of an ensemble of independent metadynamics calculations, it is necessary to transform the stochastic description of Eq. 1.13 in a probabilistic description. When the stochastic evolution is Markovian, this is done using the Fokker-Planck equation. However, Eq. 1.13 contains a history-dependent term (the bias potential) and it is clearly non-Markovian. In order to circumvent this problem a time-dependent field $\phi(s;t)$ is defined as

$$\phi(s;t) = \int dt \delta(s - s(t)) + \phi_0(s), \tag{1.17}$$

which is made of two terms: the histogram of the positions already visited by the system and a time independent gauge term $\phi_0(s)$ defined by Eq. 1.16, where it is implicitly assumed that the initial conditions are $\phi_0(s;0) = \phi_0(s)$ so that $\int ds' g(s, s')\phi_0(s;0) = F(s)$. By substituting this term in Eq. 1.13 it is found that the equations that govern the evolution of the CVs are

$$\begin{aligned} ds &= -D \int ds' \frac{\partial g(s, s')}{\partial s} \phi(s';t)\Big|_{s=s(t)} dt + \sqrt{2D}dW, \\ d\phi(s, t) &= \delta(s - s(t))dt. \end{aligned} \tag{1.18}$$

This is the crucial step that allows the non-Markovian evolution of a single dynamic variable $s(t)$ in Eq. 1.13 to be replaced with a Markovian evolution for the extended set of variables which includes $s(t)$ and the field $\phi(s, t)$. The information related to the underlying free energy $F(s)$ has disappeared from the equation of motion but it is still present through the initial condition for $\phi(s, t)$. Using the Markovian property it is possible to analyze in a rigorous manner the behavior of Eq. 1.18. In particular, by using standard techniques, it is possible to write a generalized Fokker-Plank equation and to study its asymptotic behavior for large t. Let us consider an ensemble of independent metadynamics runs and define an ensemble density. Since our dynamic variables are the position of the CV, s, and the field $\phi(s)$, the probability density will be a function of s and a functional of ϕ. This probability will be denoted as $P(\phi, s, t)$, and its Fokker-Planck equation [24] is

$$\frac{\partial P}{\partial t} = -\frac{\delta P}{\delta \phi(s)} + DP \int ds' \frac{\partial^2 g(s,s')}{\partial s^2}\phi(s') + D\frac{\partial P}{\partial s}\int ds' \frac{\partial g(s,s')}{\partial s}\phi(s') + D\frac{\partial^2 P}{\partial s^2}, \tag{1.19}$$

if the dimensionality of the system is higher than 1, a trace is implied and the second derivative is in fact a Laplacian. The probabilistic description in Eq. 1.19 is completely equivalent to the coupled stochastic equations in Eq. 1.18. This equation describes the evolution of an ensemble of metadynamics runs and has far more general relevance than its application to the Langevin model in Eq. 1.13. In fact, this formalism would allow mapping the metadynamics equations into a Markovian form also before performing the dimensional reduction.

Let us now look for the limiting distribution of Eq. 1.19 when $t \to \infty$, namely, the probability density $\bar{P}$ which satisfies $\frac{\partial \bar{P}(\phi,s,t)}{\partial t} = 0$. If it is assumed that the equilibrium probability is independent on the walker's position, this is $\bar{P}(\phi, s) = \bar{P}(\phi)$ and also that $\frac{\partial^2 g(s,s')}{\partial s^2}$ is symmetric and negative definite, it is found that the solution of Eq. 1.20 is a Gaussian distribution in functional space and is given by

$$\bar{P}_\infty\left([\phi]\right) \propto \exp\left(\frac{\beta D}{2}\int ds ds' \left(\phi\left(s'\right) - \phi_0\left(s'\right)\right)\partial_s^2 g\left(s,s'\right)\left(\phi\left(s\right) - \phi_0\left(s\right)\right)\right), \tag{1.20}$$

where $\phi_0\left(s'\right)$ is defined in Eq.1.16.

Equation 1.20 expresses the probability of obtaining a given field ϕ at the end of a metadynamics simulation. Now expressing the reconstructed free energy at time t in terms of $\phi\left(s', t\right)$ as

$$F_G\left(s,t\right) = \int ds' \phi\left(s',t\right) g\left(s,s'\right), \tag{1.21}$$

and using Eq. 1.20 it is straightforward to prove that the average value of $F_G\left(s,t\right)$ over several independent metadynamics runs is exactly equal to $F\left(s\right)$. In fact, denoting by $\langle \cdot \rangle_M$ the average over several metadynamics realizations, Eq. 1.21 gives

$$\begin{aligned} \langle F_G\left(s\right)\rangle_M &= \int ds' g\left(s,s'\right)\langle \varphi\left(s'\right)\rangle_M = \\ &= \int ds' g\left(s,s'\right)\int d\varphi P_\infty\left(\varphi\right)\varphi \\ &= \int ds' g\left(s,s'\right)\varphi_0\left(s'\right) = F\left(s\right). \end{aligned} \tag{1.22}$$

Since the negative of the biasing potential is used to estimate the free energy, the error $\epsilon(s)$ is defined as the sum of the exact underlying free energy and the biasing potential. Using Eq. 1.22 we find that the deviation of $F_G\left(s\right)$ from $F\left(s\right)$ linearly related to the field ϕ through

$$\epsilon(s) = F\left(s\right) - F_G\left(s\right) = \int ds' g(s,s')\phi_0(s'). \tag{1.23}$$

This equation implies that for a specific realization of a metadynamics process the probability of finding large errors in the estimation of the free energy is small. Using Eq. 1.23 the expected average error of a series of runs can be explicitly calculated. Since the distribution is a Gaussian with respect to ϕ, the expectation value of this field is vanishing. The error $\epsilon(s)$ is linear in the field $\phi(s)$, and consequently also its expectation value will vanish

$$< \epsilon(s) >= 0. \tag{1.24}$$

Thus, the average of the biasing potential over a series of metadynamics runs provides an unbiased estimate for the underlying free energy.

The expected quadratic deviation in the free energy is given by

$$\begin{aligned} \epsilon^2(s) &= \left\langle (F_G(s) - F(s))^2 \right\rangle_M \\ &= \left\langle (F_G(s) - \langle F_G(s) \rangle_M)^2 \right\rangle_M . \end{aligned} \tag{1.25}$$

Using the explicit expression for the probability to observe a given ϕ, Eq. 1.20 allows computing explicitly $\epsilon^2(s)$ or $\overline{\epsilon}^2$ and these turns out to be independent on $F(s)$. The specific value depends only on the metadynamics parameters, on the shape of the domain on which the system is confined, on the diffusion coefficient, and on the temperature. For example, in a cubic domain of side S in d dimensions [30] the error is

$$\overline{\epsilon}^2 = \frac{S^2 w}{\beta D \tau_G} \left(\frac{\delta s}{S} \right)^d (2\pi)^{\frac{d}{2}} \sum_k \frac{1}{\pi^2 k^2} \exp\left(-\frac{k^2 \pi^2}{2} \left(\frac{\delta s}{S} \right)^2 \right), \tag{1.26}$$

where the sum is performed over all the d dimensional vectors of integers of norm different from zero.

The dependence of the error on the simulation parameters becomes more transparent if $\overline{\epsilon}$ is expressed as an explicit function of the total simulation time. Consider in fact a free energy profile $F(s)$ that has to be filled with Gaussians up to a given level $F_{\max}$, for example the free energy of the lowest saddle point in $F(s)$. The total computational time needed to fill this profile can be estimated as the ratio between the volume that has to be filled and the volume of one Gaussian times τ_G :

$$t_{sim} \approx \tau_G \frac{F_{\max}}{w} \left(\frac{S}{\delta s} \right)^d . \tag{1.27}$$

Substituting in Eq. 1.26 gives

$$\overline{\epsilon}^2 \approx \frac{\tau_S}{t_{sim}} \frac{F_{\max}}{\beta} f_d(\frac{\delta s}{S}), \tag{1.28}$$

where $\tau_S \doteq \frac{S^2}{D}$ is the average time required for the CVs to diffuse on a distance S and

$$f_d(\frac{\delta s}{S}) = (2\pi)^{d/2} \sum_k \frac{1}{k^2 \pi^2} \exp\left(-\frac{k^2 \pi^2}{2} \left(\frac{\delta s}{S} \right)^2 \right). \tag{1.29}$$

Equation 1.28 states that the error of a metadynamics reconstruction is inversely proportional to the square root of the total simulation time, measured in units of the diffusion time. The error will be large for slow diffusing systems, in which the walker takes a long time to explore the CVs space.

Even though the efficacy of Metadynamics has been proven in very different areas like condensed matter physics, chemistry, and biophysics [25], [26], [27], the method has some problems:

(i) Its efficiency scales badly with the dimensionality, since filling the free energy wells in high dimensions can be very expensive; (ii) If a relevant variable is forgotten the algorithm is inaccurate. If the system performs a transition in the hidden degrees of freedom, the thermodynamic forces become inconsistent with the Gaussian potential and hysteresis will be present.

To resolve the first issue, a new method that combines two different techniques, replica exchange and metadynamics, was proposed [16]. It is called Bias Exchange Metadynamics, and it will be explained in the next section. To address the second issue, in Section 1.3 the most commonly used collective variables to describe proteins will be shown, and a new set of collective variables explicitly designed for protein folding will be described.

1.4 BIAS EXCHANGE METADYNAMICS

As it has been shown, ordinary metadynamics is an algorithm that can be exploited for both efficiently computing the free energy and exploring new reaction pathways, i.e., for accelerating rare events. It is based on a dynamics performed in the space defined by a few collective variables $s(x)$, where dynamics is biased by a history-dependent potential constructed as a sum of Gaussians centered along the trajectory of the collective variables. Qualitatively, as long as the CVs are uncorrelated, the time required to reconstruct a free energy surface scales exponentially with the number of CVs [28]. Therefore, the performance of the algorithm rapidly deteriorates as the dimensionality of the CV space increases. This makes an accurate calculation of the free energy prohibitive when the dimensionality of the space is larger than three. This is often the case for protein folding, where it is very difficult to select a priori a limited number of variables that describe the process, if the structure of the native state is not known in advance.

To overcome these difficulties a new method called Bias Exchange Metadynamics (BE-META) was proposed by A. Laio and S. Piana [16]. BE-META is a combination of Replica Exchange [15] and Metadynamics [14], in which multiple metadynamics simulations of the system at the same temperature are performed. Each replica is biased with a time-dependent potential acting on a different collective variable. Exchanges between the bias potentials in the different variables are periodically allowed according to a replica exchange scheme. If the exchange move is accepted, the trajectory that was previously biased in the direction of the first variable continues its evolution biased by the second (and viceversa). In this manner, a large number of different variables can simultaneously be biased, and, ideally, a high dimensional space can be explored. This allows the system to cross the residual barriers, orthogonal to the reaction coordinates, in the available simulation time. The result of the simulation is not a free energy in several dimensions, but several low dimensional projections of the free energy surface along each of the collective variables.

In more detail, let us consider N_R non-interacting replicas of the system, all at the same temperature T, and each biased along a different collective variable $s^\alpha(x)$ with $\alpha = 1, ..., N_R$. Each different replica independently accumulates a history-dependent potential $V_G^\alpha(x,t) = V_G(s^\alpha(x),t)$ that, after a sufficiently long time, would provide an estimate of the free energy projected on s^α. Periodically, it is allowed to the replicas to exchange their configurations, like in the replica exchange method [15] and in the approach introduced in [30] where exchanges between replicas evolved by metadynamics at different temperatures are performed. So in this new method, two replicas a and b are selected at random among the N_R available. The exchange move consists of swapping the atomic coordinates x^a and x^b of the two replicas. Since the two replicas are evolved under the action of two different history-dependent potentials, the move is accepted with a probability P_{ab}:

$$P_{ab} = min(1, exp(\beta[V_G^a(x^a,t) + V_G^b(x^b,t) - V_G^a(x^b,t) - V_G^b(x^a,t)])). \tag{1.30}$$

The normal potential energy of the system cancels out exactly for this kind of move. If the move is accepted, the collective variables of replica a perform a jump from $s^a(x^a)$ to $s^a(x^b)$, and replica b from $s^b(x^b)$ to $s^b(x^a)$. Since the Gaussian potentials are time dependent, detailed balance is violated in BE-META, as in ordinary metadynamics. The exchange moves are not introduced for ensuring convergence to any distribution but to introduce a jump process on top of the ordinary molecular dynamics evolution. As in ordinary metadynamics the Gaussian potential converges to the negative of the free energy. However, the jumps greatly increase the capability of each replica to diffuse in the CV space, and hence the accuracy of the free energy reconstruction, which is primarily determined by the correlation time of the dynamics in CV space. Moreover, by using this algorithm each configuration evolves under the action of a history-dependent potential that changes every time an exchange move is accepted. Even if each bias is defined in just one CV at a time, after several accepted exchanges, the system will eventually explore the space spanned by all the collective variables. This improves greatly the capability of the system to explore the configuration space compared to metadynamics.

This method has been successfully used to fold some small proteins Adivillin, Villin and Tryptophan Cage [16], [17].

1.5 BE-META IN THE UNRES FORCE FIELD

1.5.1 UNRES

The united-residue (UNRES) force field has been developed by Scheraga, Liwo et al [18]. This force field is successful in predicting the native-like structures of some small globular proteins with simple topology [19].

In this model, a polypeptide chain is represented by a sequence of α-carbons, linked by virtual bonds, with attached side chains SC. United peptide groups P are located in the middle of the consecutive α-carbons. Only the peptide groups and the united side chains serve as interaction sites, while the α-carbons assist in the definition of the geometry (see Figure 1.3). All the virtual bond lengths (i.e., $C_\alpha - C_\alpha$ and $C_\alpha - C_{SC}$) are fixed, while the side-chain angles (α_{SC} and β_{SC}), as well as the virtual-bond angles (θ and γ) and the dihedral angles (χ) can vary.

UNRES is a force field that uses physics-based interactions to describe the forces over the coarse grain particles representing the amino acids of a protein. It has been derived as a restricted free energy function of an all-atom polypeptide chain plus the surrounding solvent, where the all atom energy function is averaged over the degrees of freedom that are lost when passing from the all-atom to the simplified system (they are: the degrees of freedom of the solvent, the dihedral angles for rotation about the bonds in the side chains, and the torsional angles for rotation of the peptide groups about the C_α virtual bonds). The restricted free energy is further decomposed into factors coming from interactions within and between a given number of united interaction sites. Expansion of these factors into generalized Kubo cumulants [35] enabled to derive approximate analytical expressions for the respective terms, including the multibody or correlation terms.

The energy of the virtual-bond chain is expressed by

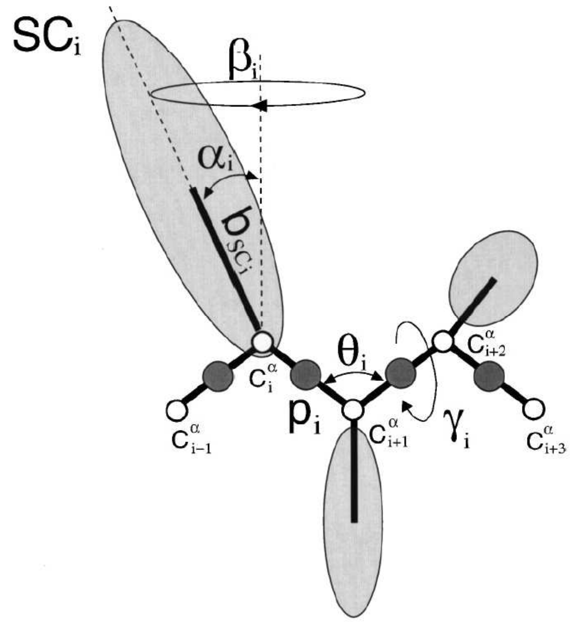


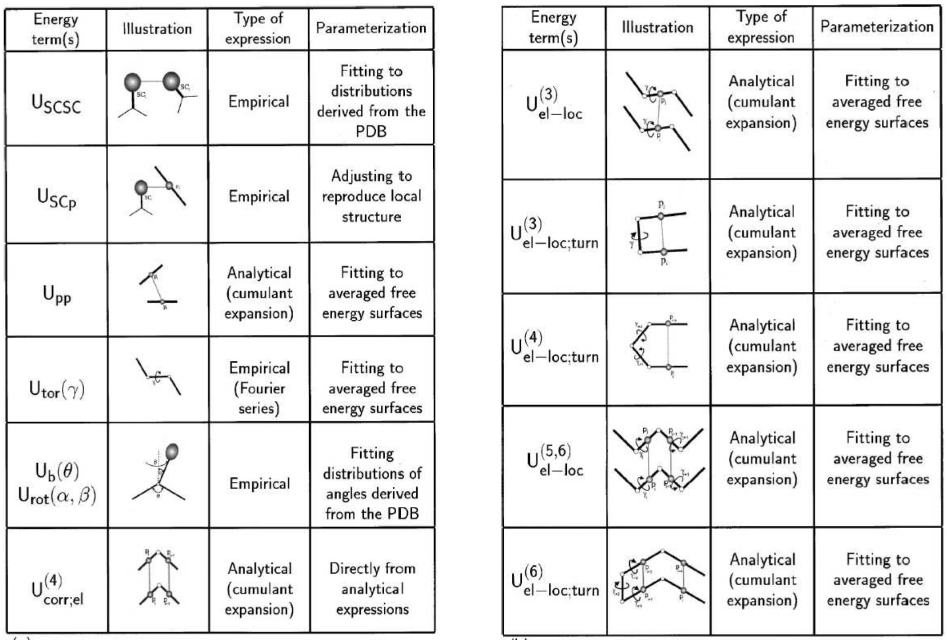
Figure 1.3: The UNRES model of a polypeptide chain.

$$\begin{aligned}U = {} & w_{SC}\sum_{i<j} U_{SC_iSC_j} + w_{SCP}\sum_{i\neq j} U_{SC_iP_j} + w_{PP}^{VDW}\sum_{i<j-1} U_{P_iP_j}^{VDW} + w_{PP}^{el}\sum_{i<j-1} U_{P_iP_j}^{el}\\ & + w_{tor}\sum_i U_{tor}(\chi_i) + w_{tord}\sum_i U_{tord}(\chi_i,\chi_{i+1}) + w_b\sum_i U_b(\theta_i) + w_{rot}\sum_i U_{rot}(\alpha_i,\beta_{i+1})\\ & + w_{corr}^3 U_{corr}^3 + w_{corr}^4 U_{corr}^4 + w_{corr}^5 U_{corr}^5 + w_{corr}^6 U_{corr}^6 + w_{turn}^3 U_{turn}^3 + w_{turn}^4 U_{turn}^4\\ & + w_{turn}^6 U_{turn}^6 + w_{bond}\sum_i^{nbond} U_{bond}(d_i). \end{aligned} \tag{1.31}$$

The U's denote energy components, they are functions of variables describing the geometry of the polypeptide chain, and they are summarized in a pictorial manner in Figure 1.4. The w's are weights of the energy terms, which are responsible for the balance between different kinds of interactions. The weights are determined by optimizing a Z-score function [36], [37], i.e., by maximizing the ratio of the difference between the energy of the native structure and the mean energy of non-native structures taken from a set of decoys from the PDB. The energy terms contain also internal coefficients, such as the Van der Waals radii and well depths in the case of $U_{SC_iSC_j}$, which are determined to reproduce a particular term as accurately as possible.

The energy components can be classified into the following two categories, according to the type of functional expression:

- *Empirical* ($U_{SC_iSC_j}$, $U_{SC_iP_j}$, U_{tor} , U_{tord} U_b , U_{rot}): The term $U_{SC_iSC_j}$ corresponds to the mean free energy of the hydrophobic-hydrophilic interactions between the side chains. It, therefore, implicitly contains contributions from the interactions with the solvent. This is the only fully sequence-dependent term in UNRES. The terms $U_{SC_iP_j}$ encode the excluded-volume potential of the side-chain peptide-group interactions. These terms have the functional form



Energy term(s)	Illustration	Type of expression	Parameterization
U_{SCSC}		Empirical	Fitting to distributions derived from the PDB
U_{SCp}		Empirical	Adjusting to reproduce local structure
U_{pp}		Analytical (cumulant expansion)	Fitting to averaged free energy surfaces
$U_{tor}(\gamma)$		Empirical (Fourier series)	Fitting to averaged free energy surfaces
$U_{b}(\theta)$ $U_{rot}(\alpha, \beta)$		Empirical	Fitting distributions of angles derived from the PDB
$U^{(4)}_{corr;el}$		Analytical (cumulant expansion)	Directly from analytical expressions

Energy term(s)	Illustration	Type of expression	Parameterization
$U^{(3)}_{el-loc}$		Analytical (cumulant expansion)	Fitting to averaged free energy surfaces
$U^{(3)}_{el-loc;turn}$		Analytical (cumulant expansion)	Fitting to averaged free energy surfaces
$U^{(4)}_{el-loc;turn}$		Analytical (cumulant expansion)	Fitting to averaged free energy surfaces
$U^{(5,6)}_{el-loc}$		Analytical (cumulant expansion)	Fitting to averaged free energy surfaces
$U^{(6)}_{el-loc;turn}$		Analytical (cumulant expansion)	Fitting to averaged free energy surfaces

Figure 1.4: UNRES force field terms.

of a repulsive Lennard Jones type potential, with parameters adjusted to obtain appropriate local geometries of model systems, such as α-helices. U_{tor} is the torsional potential for rotation about virtual bonds (variation of dihedral angles χ). U_{tord}, U_b, U_{rot} are the virtual-bond dihedral angle double-torsional terms, virtual-angle-bending and side chain-rotamer potentials, respectively.

- *Analytical* ($U_{P_iP_j}$, U^m_{corr}, U^m_{turn}): The term $U_{P_iP_j}$ is the energy of the average electrostatic interactions between backbone peptide groups. The correlation terms U^m_{corr} represent the multibody correlation contributions from the coupling between backbone local and backbone electrostatic interactions. The terms U^m_{turn} are correlation contributions involving m consecutive peptide groups, these were derived from a generalized cumulant expansion of the restricted free energy (RFE) of the system consisting of the polypeptide chain and the surrounding solvent. The multibody terms are necessary for reproducing regular α-helical and β-sheet structures.

1.5.2 Implementation of BE-META in UNRES

In our simulations, we used the UNRES software implementation provided by Adam Liwo in http://www.chem.univ.gda.pl/ adam/Downloads/index.html. The program is used to perform a molecular dynamics simulation of a protein. The equations of motions are integrated with the velocity Verlet algorithm [38] and the Berendsen thermostat [40].

We modified the original code introducing, in the normal molecular dynamics simulation, the metadynamics force, which is defined as:

$$F_{meta} = -\vec{\nabla} V_G(r, t), \tag{1.32}$$

where V_G is the biasing potential in Equation (1.10). So the force over the coarse grain particles is now

$$F_{tot} = F_{unres} + F_{meta}. \tag{1.33}$$

From now on this implementation will be called UNRES-META.

As shown in Section 1.3, in BE-META a number of metadynamics simulations, each biased on a different CV are run in parallel. After a certain time (τ_E) exchanges between the different replicas are attempted using the Metropolis scheme following Equation 1.30. This process is repeated iteratively until each free energy profile converges. The exchange procedure is implemented externally from the UNRES-META program through a bash script. In Table 1.1 a scheme of the algorithm is shown.

Steps	BE-META algorithm
1.	Choose a set of CVs of number N_R.
2.	Start N_R UNRES-META simulations, each biased on a different CV.
3.	After τ_E, stop the UNRES-META simulations.
4.	Randomly choose $N_R/2$ pairs of runs.
5.	Attempt an exchange between the biasing potentials of each pair, using the Metropolis scheme (see Equation 1.30).
6.	If the exchange is accepted, swap the biasing potentials .
7.	Restart the procedure from step 2.
8.	Stop when the system has converged.

Table 1.1: BE-META algorithm implemented in UNRES.

Convergence is found when the system has explored all the configuration space and then starts diffusing freely through it, i.e., the history-dependent potentials start growing parallel.

1.6 COLLECTIVE VARIABLES

As already discussed, the set of chosen collective variables $s(x)$, plays an essential role in determining the convergence and efficacy of the methods used for calculating free energy profiles and finding the metastable states of a system. If a set of CVs does not distingush different metastable states of the system the simulation will present hysteresis and not all of the important regions of the conformational space will be explored. To choose the proper set of CVs one needs to exploit some basic information on the topological, chemical and physical properties of the system. In the case of proteins, that are chains of amino acids with certain topological features (see APPENDIX), some intuitive CVs are:

- *Number of contacts*: This CV counts the number of atoms i in set A that have a distance smaller than r_o from an atom j in another set B. These could be, for example, the number of H-bonds, of C_γ contacts or the number of salt bridges. Mathematically it is expressed as

$$N = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} C_{ij}, \tag{1.34}$$

where N_A, N_B are the total number of atoms in A and in B respectively, and C_{ij} is a switch function like:

$$C_{ij} = \frac{1 - \frac{r_{ij}^m}{r_o^m}}{1 - \frac{r_{ij}^n}{r_o^n}}, \tag{1.35}$$

where r_{ij} is the distance between atoms i and j, and m, n are exponents that can vary depending on how sharp or wide the step function is wanted; common values are $m = 8$ and $n = 12$. If the sets A and B are the C_α carbons of the protein and if $r_o \approx 7.5\dot{A}$, the last equation defines an important mathematical object called the Contact Matrix, which in some recent studies has been proven to give an almost complete description of the protein's topology.

- *Contact Order*: It was introduced [34] to compare differences in topology (rather than in size) between proteins of different length. This parameter is small for proteins stabilized mainly by local interactions and is large when residues in a protein interact frequently with partners far away in the protein sequence. It is defined for the set of C_α as

  $$CO = \frac{1}{N} \sum_{i=1} \sum_{j=1} C_{ij} |i - j|, \tag{1.36}$$

  where N is the total number of contacts defined in Eq. 1.34, and C_{ij} is given by Eq. 1.35.

- *Dihedral value*: It measures the number of dihedrals ϕ_i (see APPENDIX), in a certain set A, that are similar to a reference dihedral ϕ_o. It is defined as

  $$\Phi_{\alpha\beta} = \sum_{i=1}^{N_A} \frac{1 + cos(\phi_i - \phi_o)}{2}. \tag{1.37}$$

  This CV is used to distinguish if residues belong to an α-helix or a β-sheets.

- *Dihedral correlation*: This CV measures the correlation between two successive dihedrals ϕ_i, ϕ_{i+1} of the protein backbone in a certain set A:

  $$\Phi_{SS} = \sum_{i=1}^{N_A} \frac{1 + cos(\phi_i - \phi_{i+1})}{2}. \tag{1.38}$$

  Since secondary structure elements α-helices or β-sheets have a correlation between its successive backbone dihedrals, this CV is commonly used to distinguish if a certain set of dihedrals has a secondary structure or not.

These collective variables were successfully used in [16] and [17] to fold some small proteins using BE-META. One could hope that these collective variables are enough to provide a satisfactory description also for longer proteins. Unfortunately this is not granted, and these CVs were sufficient because Adivillin and Tryptophan Cage's topologies are simple, but the problem becomes more complex when the protein gets larger. Some tests with these CVs were performed to fold SH3, that is a β-barrel type protein of 56 amino acids, but no complete folding from the random coil was observed in a long simulation time. In the next section we will propose a new set of collective variables that might allow for a more flexible description of protein topology.

1.6.1 New Topology-Based CVs

As our final goal, for future work, is to fold an average size protein, here we wish to improve some limitations of the previously used collective variables. At this scope, we propose a new set of CVs, based on a more topological approach. The new CVs are constructed starting from the contact matrix, and count approximately the number of 1^{st} and 2^{nd} nearest neighbor contacts or the number of closed loops.

A protein can be modeled by a set of points representing its amino acids. Topologically, this set of points is characterized by its contact matrix (contact map in graph theory), as defined in Equation 1.35. Recently, some properties of this matrix have been investigated: it has been shown that the contact map of the native structure of globular proteins can be reconstructed starting from the sole knowledge of the contact map principal eigenvector and that the reconstructed contact map allows in turn for the accurate reconstruction of the three-dimensional structure [31]. It has also been shown that the eigenvalues of these matrices are related with elementary excitations of the protein and that the eigenvectors indicate participation of each amino acids in these excitation modes [32]. Experimentally, it has also been shown that C_{ij} gives a convenient description of the transition state for protein folding [33].

Our idea is to propose a new set of collective variables that provide an approximate description of C_{ij}.

GENERALIZED CONTACT VARIABLES

The fist type of CVs that we introduce have the form

$$\phi_i^n = \sum_k (C_{ik}^n)^2, \tag{1.39}$$

where C_{ij} is the contact matrix (CM) defined in Equation 1.35 and n is a integer exponent. Denoting by λ_α and ν_i^α the eigenvalues and eigenvectors of the CM, and using the fact that ν_i^α is an orthonormal basis and that $\lambda_\alpha^{2n} = \sum_{lmk} \nu_l^\alpha C_{lk}^n C_{km}^n \nu_m^\alpha$, the vector defined in Eq. 1.39 can be written as

$$\phi_i^n = \sum_\alpha (\lambda_\alpha)^{2n} (\nu_i^\alpha)^2. \tag{1.40}$$

For $n = 1$, $\phi_i^0 = \sum_j (C_{ij})^2$, which counts the number of first nearest neighbor links, (i.e., the chemical bonds or coordination number N_i). The case $n = 2$ seems more useful: one sums the squares of the elements within a line of the matrix $(C^2)_{ij}$, this is the matrix counting the number of second nearest neighbor links (bridging atoms) between i and j. The vector defined in Eq. 1.40 has a cutoff, in the sense that if the topologies are the same until the $n-th$ nearest neighbor, the lower order ϕ_i^n are identical, but the higher orders are not.

The problem with this CV is that when it is computed on a molecular dynamics trajectory it fluctuates up to the order of N^2, where N is the total number of particles in the system. To avoid too large fluctuations, we also considered its block average:

$$\bar{\phi}_i^n = \frac{1}{7} \sum_{j=i-3}^{j=i+3} \phi_j^n, \tag{1.41}$$

where ϕ_j^n is defined in Eq. 1.39.

An example of ϕ_i^2 and $\bar{\phi}_i^2$ for the folded and unfolded states of protein 1E0G, is shown in Figure 1.5. It is found that for the unfolded state these CVs have a value near to 0, this is expected because they provide an estimate of the number of 2^{nd} nearest neighbors. It is shown that these CVs clearly distingush between the folded and unfolded states of 1E0G.

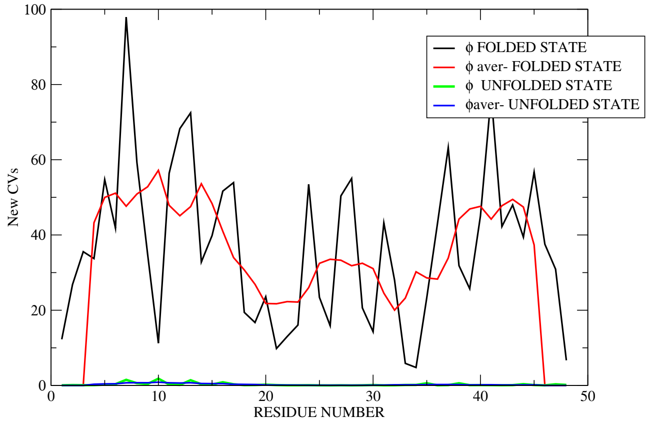


Figure 1.5: ϕ^2 and $\bar{\phi}^2$ for the folded and unfolded states of protein 1E0G.

LOOP VARIABLES

Another collective variable that we consider is the squared number of closed loops, of order n, containing the C_α carbon i. This is defined as

$$\psi_i^n = (C_{ii}^n)^2. \tag{1.42}$$

As we have done before to avoid too large fluctuations, we also considered its average as

$$\bar{\psi}_i^n = \frac{1}{7}\sum_{j=i-3}^{j=i+3} \psi_j^n. \tag{1.43}$$

It can be shown that also these variables clearly distinguish between the folded and unfolded states.

QUALITY MEASURE OF A SET OF CVS

In general a collective variable set is good if it is able to distinguish clearly the different metastable states of the system. Let us assume there exist two different metastable states A and B of the system, each one with a probability distribution over the CV space. We assume that the probability distribution of the collective variables in the metastable set is a Gaussian, of the form

$$P_A(\vec{s}) = \sqrt{\frac{Det(\hat{A})}{(2\pi)^n}} \int d^n s e^{(\vec{s}-\vec{\mu}^A)^T \hat{A}^{-1}(\vec{s}-\vec{\mu}^A)}, \tag{1.44}$$

where n is the dimension of the CV space (number of CVs), $\vec{\mu}_A$ is the center of the Gaussian and $\hat{A}$ is the correlation matrix defined as

$$A_{ij} =< (s_i - \mu_i^A)(s_j - \mu_j^A) > . \tag{1.45}$$

If the CV space is sufficiently good, then the two probability distributions P_A and P_B, in two different meta-stable sets A and B are almost non-overlapping. Consider in particular, the distance defined by the Jensen-Shannon divergence:

$$D = \frac{1}{2}\Big[\int d^n s P_A(\vec{s}) log\Big(\frac{P_A(\vec{s})}{P_B(\vec{s})}\Big) + \int d^n s P_B(\vec{s}) log\Big(\frac{P_B(\vec{s})}{P_A(\vec{s})}\Big)\Big]. \tag{1.46}$$

This is a measure of the separation of two probability distributions in the CV space. For distributions that occupy the same region of the space D is small, for distributions that are separated D is large. Thus a good set of CVs must provide a large D for all the pairs of relevant metastable states of the system.

If the two probaility distributions $P_A(\vec{s})$ and $P_B(\vec{s})$ are Gaussians (see Eq. 1.44), then the quantity D defined in Equation 1.46 can be expressed as

$$D = \frac{1}{2}(\vec{\mu}_A - \vec{\mu}_B)^T(\hat{B}^{-1} + \hat{A}^{-1})(\vec{\mu}_A - \vec{\mu}_B) + \frac{1}{2}(tr(\hat{A}\hat{B}^{-1} + \hat{B}\hat{A}^{-1})) - n, \tag{1.47}$$

where n is the dimension of the CV space, ($\vec{\mu}_A$, $\vec{\mu}_B$) and ($\hat{A}$, $\hat{B}$) are the centers and correlation matrices of P_A and P_B, respectively. A simple example of the measure of D, is to consider two probability distributions that have the same spread ($\hat{A} = \hat{B}$) but have different centers. In this case $D = (\vec{\mu}_A - \vec{\mu}_B)^T(\hat{A}^{-1})(\vec{\mu}_A - \vec{\mu}_B)$ is a parabolic function of the distance between the centers.

For example, consider two molecular dynamics simulations on two different metastable states A and B of protein 1E0G (see Section 2.1.1), which were performed using the UNRES force field. The trajectories obtained are two samples that represent two different probability distributions of the metastable states A and B, which can be mapped onto a collective variable space. Each of the previously defined collective variables in Equations 1.39, 1.41, 1.42, 1.43 with $n = 2$, were calculated over the two trajectories (A and B), for residues $i = 5, 11, 17, 23, 29, 35, 41, 47$ of protein 1E0G. Thus, for each CV we have an eight dimensional space. The probability distribution for the two meta-stable states in each CV space was estimated using Equation 1.44. The D between the two distributions (Equation 1.47) in each CV space was found. The results are shown in Table 1.2.

Set of CVs	Distance defined in Eq. 1.46
ϕ_j^2	$D = 47.9$
ϕ_j^2	$D = 98.2$
ψ_j^2	$D = 44.5$
ψ_j^2	$D = 103.8$

Table 1.2: Distance between two distribution probabilities, corresponding to two different metastable states, using diverse sets of CVs.

From this simple analysis we can conclude that $\bar{\phi}_j^2$ and $\bar{\psi}_j^2$, not only distinguish the folded and unfolded state, but also clearly two different meta-stable states of protein 1E0G. We are currently performing BE-META simulations using these variables in order to compare their performance with the normal CVs: preliminary results indicate that also these new Cvs enable to fold 1E0G in a short computer time compared to reference replica exchange simulations [18].

2 OPTIMIZATION OF THE BIAS EXCHANGE PARAMETERS

As discussed in Chapter 1, BE-META is a promising methodology that can be applied for studying complex conformational transitions like protein folding, and protein-protein interaction. Still, several issues remain to be addressed: what is the optimal exchange rate between replicas? What is the optimal set of collective variables?

In this chapter, we will use the coarse grain force field called UNRES [18], and BE-META [16] to fold protein 1E0G (48 aa). This will be done several times using different sets of collective variables and exchange times. We will compare how the different simulations explore the conformational space in order to optimize the performance of the method.

2.1 FOLDING OF 1E0G

2.1.1 Biological description of 1E0G (LysM Domain)

The LysM is a wide spread domain. It was originally identified in enzymes that degrade bacterial cell walls but is also present in many other bacterial proteins. Several proteins that contain the domain, such as Staphylococcal IgG binding proteins and Escherichia coli intimin, are involved in bacterial pathogenesis. LysM domains are also found in some eukaryotic proteins, apparently as a result of horizontal gene transfer from bacteria. Available evidence suggests that the LysM domain is a general peptidoglycan-binding module. Peptidoglycan is a component of the cell walls of both Gram positive and Gram negative bacteria, where it provides mechanical support and prevents bacteria from bursting under high internal osmotic pressure. The importance of peptidoglycan is highlighted by the fact that many antibiotics block its biosynthesis.

Since this protein was originally identified in bacterial lysins, the repeats have been termed LysM domains for lysin motif. To understand further the structure and function of this module the structure of a LysM domain from Escherichia coli membrane-bound lytic murein transglycosylase D (MltD) was determined by Alex Bateman and Mark Bycroft [42]. MltD consists of an N-terminal transglycosylase domain with two LysM repeated at the C terminus. The exact function of MltD is unknown. However, based on the phylogenetic distribution of the protein it has recently been suggested that it may be involved in fagellar biosynthesis [43]. Residues 389 to 452 of E. coli MltD encompassing the C-terminal copy of the two motifs in this enzyme, were overexpressed in E. coli and purified to homogeneity. Assignments for this peptide were obtained using a range of 2D and 3D heteronuclear NMR experiments using isotopically enriched samples. Non-random coil chemical shifts and long-range NOEs were only observed for the residues between 398 and 445, and structures were calculated for this region. Residues will now be referred to with Asp398 as residue 1.

As it is shown in Figure 2.1, 1E0G (LysM domain) protein has a $\beta\alpha\alpha\beta$ secondary structure with the two helices packing onto the same side of a two-stranded anti-parallel β sheet .

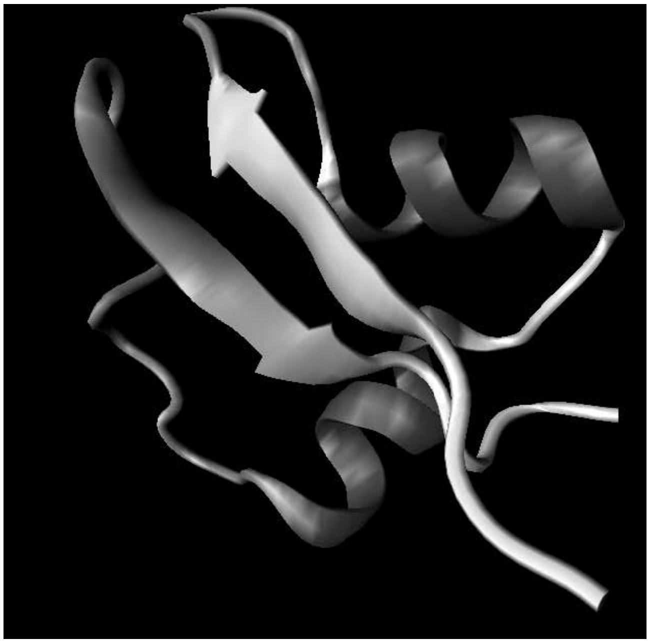


Figure 2.1: Tertiary structure of protein 1E0G.

This protein was used to parametrize the UNRES force field and it was folded using the multiple-replica exchange method in [18]. The authors used 50 million steps with a time step of $5fs$ in 80 processors, with 20 temperatures and 16 replicas per temperature. This is a good trial protein to optimize the BE-META parameters because it is known that its folded state is the global free energy minimum in the UNRES force field, and it contains both α and β secondary structure.

The choice of the protein to fold is limited by the fact that the UNRES force field might not always provide as a global minimum the folded state of the protein. Some molecular dynamics simulations were made with Advillin starting from the experimental PDB, and it was found that the folded state is not a global minimum using UNRES.

2.1.2 Setup of the Simulations

Since the objective of this thesis is to find out the best parameters for BE-META, several different simulations were ran to fold 1E0G.

The molecular dynamics parameters, which are fixed for all the runs, were $\Delta t = 5fs$, $T = 280K$, and an initial unfolded protein configuration was used as the initial state (same setup as in Liwo et al. [18]). The metadynamics parameters τ_G (time at which the Gaussians are introduced) and the height w of the Gaussians were also held fixed for all simulations with the values $\tau_G = 1ps$ and $w = 0.012kcal/mol$. The width of the Gaussian (δS) depends on the collective variable and it was chosen as the standard deviation of the value of the collective variable in an unbiased molecular dynamics simulation in the folded state.

POOL OF CVS

Using the definitions for the collective variables in Section 1.6 and due to the coarse grained nature of the force field, the following CVs, that do not explicitly contain hydrogen bonds but only C_α

atoms, were choosen to fold 1E0G:

i ***Hydrophobic contacts***: Equation 1.34 was used with the sets A and B as the hydrophobic atoms in the system. The width of the Gaussian is $\delta S = 1.0$.

ii ***Salt Bridges***: It is the contact number (Eq. 1.34) of the amino acids that can form salt bridges in the protein, those charged positively are in set A with those charged negatively in B ($\delta S = 0.2$).

iii ***Contacts of the 1^{st} half with the 2^{nd} half***: The C_α contacts (Equation 1.34) that can be made between the amino acids in the first half of the protein (set A) with those in the second half (set B) ($\delta S = 2$).

iv ***Uniformly distributed contacts***: The number of contacts that can be formed between residues $4, 12, 20, 28, 36, 44$ of the protein ($\delta S = 0.2$).

v ***Dihedral value for the 1^{st} quarter of the protein***: This CV is defined in Eq. 1.37. Since we are working in a coarse grain model, the dihedrals used are those formed between 4 sucesive C_α (See APPENDIX). It was found that for this dihedral the correct ϕ_o is $0.872 rad$ for the α-helices, and $1.754 rad$ for the β-structure. This provides a clear separation between these two secondary structures. The set A was the first quarter of the protein, the next three items are the same collective variable but defined for different sets A, these are $2^{nd}, 3^{rd}, 4^{th}$ quarters of the protein, respectively ($\delta S = 0.2$).

vi ***Dihedral value for the 2^{nd} quarter***

vii ***Dihedral value for the 3^{rd} quarter***

viii ***Dihedral value for the 4^{th} quarter***

ix ***Dihedral value for the 1^{st} half***: It is defined in Equation 1.37, and uses the first half of the protein as set A ($\delta S = 0.2$).

x ***Dihedral value for the 2^{nd} half*** As the last CV is uses Eq. 1.37, and set A is defined in the second half of the protein ($\delta S = 0.2$).

xi ***Dihedral correlation for the 1^{st} half***: Using Eq. 1.38, this collective variable is defined using as set A as the first half of the protein. ($\delta S = 0.2$)

xii ***Dihedral correlation for the 2^{nd} half***: As the previous CV, it is defined in the second half of the protein ($\delta S = 0.2$).

DIFFERENT SETUPS

What changed for the different BE-META simulations was the set of collective variables used (choosen from the previosly described pool) and the exchange time (τ_E). The setup for the 9 BE-META simulations is explained next.

1. **τ_E= 20ps and 10 CVs**: The CVs **i-viii** and **xi-xii**, described in the last section, were used in this simulation with an exchange time of $20ps$.

2. **τ_E= 20ps and 8 CVs**: This simulation uses the eight CVs **i-viii**, i.e. the four contacts and the four α, β dihedrals.

3. τ_E= **20ps and 6 CVs**: The CVs **i-iv** are used. Also two dihedral value CVs **ix** and **x**.

4. τ_E= **20ps and 4 CVs**: This simulation uses the contacts among hydrophobic residues (CV **i**) plus the 1^{st} half with the 2^{nd} half contacts (CV **iv**). The CVs **ix** and **x** are also used.

5. τ_E= **20ps and 4 Contacts**: It uses the first four CVs defined **i-iv** , i.e. only the contacts.

6. τ_E= **20ps and 4 Dihedrals**: Four dihedral value CVs are used, where each one is defined in a set that contains a quarter of the protein, CVs **v-viii**.

7. τ_E= **120ps and 8 CVs**: The same 8 collective variables as defined in simulation 2 are used, but the exchange time is 120ps.

   The following two simulations also use the 8 CVs **i-vii**, but have an exchange time of 60ps and 5ps, respectively.

8. τ_E= **60ps and 8 CVs.**

9. τ_E= **5ps and 8 CVs.**

The results obtained in these runs will be shown in the next section.

2.1.3 RESULTS

CONVERGENCE

As it was explained in Section 1.3, the result of a BE-META simulation is a number of low dimensional projections of the free energy surface along each of the chosen collective variables. In Figure 2.2, for the simulation that uses 4CVs and τ_E= 20ps, these projections are shown as the system is evolving in time. After a certain time the free energy profiles start to grow parallel. This means that the system has filled all his free energy profiles with the biasing potentials and can diffuse freely through the CV space. At this time one can say that the simulation has converged.

Remarkably, all the nine BE-META simulations previously described in Section 2.1.2 reach convergence, and find the folded state of the protein within an RSMD of $5.5A$, the same accuracy obtained with REMD by Liwo et al [18]. As expected, the value of the CVs in the folded state correspond to the position of the global free energy minimum in each projection. The length of the simulation for each replica was of the order of $80ns$ (due to the coarse grained force field, the simulation time is not directly comparable to all-atom force fields or experiments), and convergence was found approximately at $50ns$. The total amount of simulation time for the longest BE-META simulation was $800ns$. This compares with a total simulation time of $80000ns$ in the refernce MREMD by Liwo, and it shows that BE-META is much more efficient than MREMD. This is also shown by the fact that instead of needing 80 processors to fold $1E0G$ using MREMD, we used only 8 processors and 300 hours of CPU time with BE-META.

CLUSTER SEARCH COMPARISON

Since all the simulations are able to fold 1E0G, what we want to know is which one of these simulations is the most efficient.

As a primary step, we performed a statistical analysis on the time required in each BE-META simulation to find the folded state of the protein. The first time in which a replica in each BE-META simulation finds the folded state of the protein, is shown in Figure 2.3: the best performanced is obtained for the simulation that uses $\tau_E = 20ps$ and 4CV (see simulation **4** described

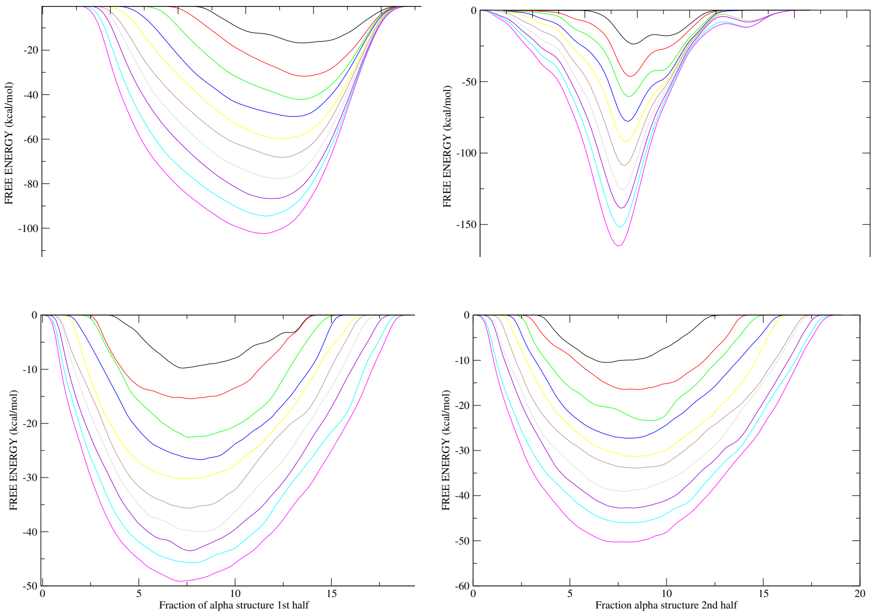


Figure 2.2: Free energy profiles evolving in time, for the BE-META simulation that uses 4CV and τ_E= 20ps.

in section 2.1.2). The BE-META simulations of 10CV - $\tau_E = 20ps$ and 4 dihedrals - $\tau_E = 20ps$ did not find the folded state of the protein in the time interval shown in this graph ($15ns$).

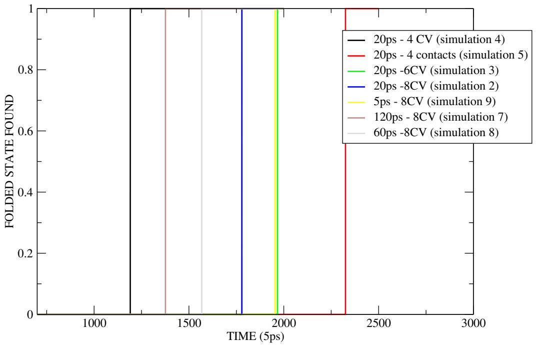


Figure 2.3: First time the folded state was found, for the different BE-META simulations.

Of course the moment in which the folded state if found is influenced by stochastic fluctuations. Thus, in order to assess more reliably the efficiency of the different simulations, 13 independent runs ($12ns$) where run for each of the 9 BE-META simulations, initialized with different seeds for the random number generator. The first time at which a replica in BE-META simulation found the folded state was extracted for all the different setups. The results are shown in Table 2.1.

The total simulation time (Time of the Walker $*$ Total Number of Walkers) taken by each BE-META simulation to find the folded state of the protein, are shown in Table 2.2, for all the different runs.

The average folding time ($\bar{\tau}_{fold}$), for each BE-META setup, is estimated by fitting the data

10CV	8CV	4CV	4cont	4dihe	6CV	60ps	120ps	5ps
NF*	914.30	1304.01	2317.96	NF	963.35	914.25	1242.47	1396.46
1399.79	NF	NF	2251.54	NF	1547.66	418.43	2331.53	125.14
NF	1778.86	1796.80	2325.12	NF	1417.55	1657.45	1889.79	NF
286.03	NF	1190.64	NF	NF	1966.64	1567.38	1376.28	1953.48
NF	1603.32	489.32	2260.62	NF	343.71	NF	2239.76	1771.95
NF	429.89	2335.08	NF	NF	1623.63	278.57	1376.79	351.88
1324.81	NF	1747.38	1214.60	NF	NF	1333.44	1425.27	1115.14
571.75	1309.58	1089.35	1857.91	1029.75	NF	1607.04	134.31	NF
NF	1416.98	2174.30	1046.69	NF	389.42	1305.34	2175.13	NF
2747.39	893.46	2646.13	NF	NF	NF	NF	1566.06	1284.83
NF	NF	1388.06	2005.66	NF	90.38	1984.23	2809.75	87.83
NF	388.78	NF	NF	NF	1506.08	2185.15	1050.41	565.56
NF	NF	NF	NF	NF	1641.61	1676.00	1280.81	1724.66

Table 2.1: FIRST TIME (units of 5ps) at which a walker in each BE-META simulation found the folded state. (*NF=NOT FOUND)

10CV	8CV	4CV	4cont	4dihe	6CV	60ps	120ps	5ps
NF*	7314.41	5216.07	9271.84	NF	5780.13	7314.07	7454.87	11171.7
13997.9	NF	NF	9006.19	NF	9286.01	3347.49	13989.2	1001.17
NF	14230.9	7187.22	9300.51	NF	8505.32	13259.7	11338.8	NF
2860.36	NF	4762.58	NF	NF	11799.9	12539	8257.7	15627.8
NF	12826.6	1957.28	9042.49	NF	2062.3	NF	13438.6	14175.6
NF	3439.14	9340.34	NF	NF	9741.82	2228.56	8260.79	2815.05
13248.2	NF	6989.56	4858.4	NF	NF	10667.6	8551.65	8921.14
5717.56	10476.7	4357.4	7431.67	4119	NF	12856.3	805.901	NF
NF	11335.9	8697.2	4186.79	NF	2336.55	10442.8	13050.8	NF
27474	7147.74	10584.5	NF	NF	NF	NF	9396.41	10278.7
NF	NF	5552.27	8022.67	NF	542.305	15873.9	16858.5	702.644
NF	3110.26	NF	NF	NF	9036.5	17481.2	6302.51	4524.5
NF	NF	NF	NF	NF	9849.68	13408	7684.88	13797.3

Table 2.2: TOTAL SIMULATION TIME (units of 5ps) at which each BE-META simulation found the folded state. (*NF=NOT FOUND)

presented in Table 2.2 to an exponential distribution of the form

$$P(t) = \frac{e^{-\frac{t}{\bar{\tau}_{fold}}}}{\bar{\tau}_{fold}}. \tag{2.1}$$

As in many cases the folding event is not observed the optimal $\bar{\tau}_{fold}$ is obtained by maximum likelihood [44]. These results are shown in Figure 2.4 for the simulations that have the same exchange time $\tau_E = 20ps$, and in Figure 2.5 for the simulations that have the same collective variable setup (8CVs like in simulation 2) but different exchange times..

From Figure 2.4 we can conclued that the optimal set of collective variables for folding 1E0G with $\tau_E = 20ps$, are the CVs **i**, **iv**, **xi** and **xi**: two tertiary contacts and two torsional dihedrals. From the fact that the longest folding time is found for the BE-META simulation that has only dihedral values (CVs **v-viii**), it is shown that CVs that give tertiary contacts are necesary to improve the method's efficiency.

From Figure 2.5 it is found that there are two relevant process in the exchange time that compete for the efficiency in the method. If exchanges are frequent (τ_E small) then the replicas are

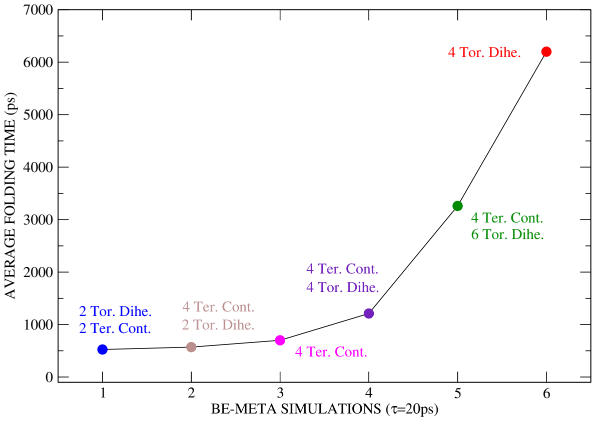


Figure 2.4: Average folding times for the BE-META simulations that use different sets of CVs and $\tau_E = 20ps$.

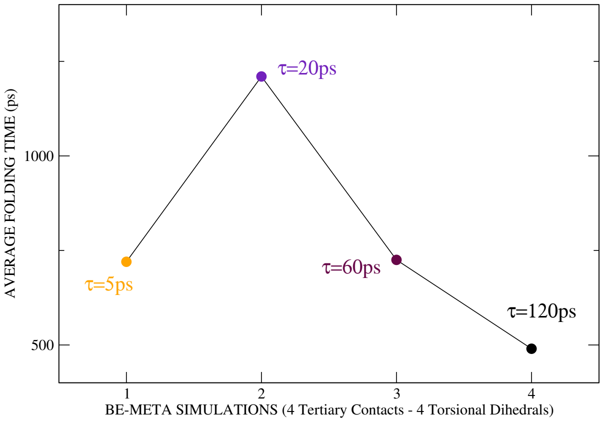


Figure 2.5: Average folding times for the BE-META simulations that use different τ_E and the set of CVs defined for simulation 2.

able to diffuse fast through the conformational space but exchanges are performed before local free energy wells are filled. Instead if τ_E is large then the replicas have enough time to fill the local wells, and the system will be encouraged to explore other parts of the conformational space.

The capability of BE-META to explore systematically the configuration space was further investigated by selecting the 80 most populated clusters found by a REMD run on 1E0G. For each of the BE-META simulations, we then computed the number of the reference clusters that are found as a function of time. A reference cluster is assumed to be explored whenever the RMSD of a frame in the trajectory is lower than $3.5\dot{A}$. The results are shown in Figure 2.6 for the simulations that have the same exchange time $\tau_E = 20ps$, and in Figure 2.7 for the simulations that have the same collective variable setup (8CV) but different exchange times.

From Figure 2.6 we can conclude that the BE-META simulations that explore more efficiently the conformational space, for $\tau_E = 20ps$, are those that have a big number of collective variables (6CV, 8CV and 10CV), thus the results shown in Figure 2.3 are due to stochastic fluctuations. A proof that collective variables that describe tertiary structures topology are very important, is given by the fact that the most inefficient simulation is the one that uses only 4 dihedrals, i.e. only collective variables that describe secondary structure elements.

From the comparison of different exchange times, we can conclude that, for the short times all the BE-META explore approximately the same number of clusters, but at long simulations times frequent exchanges become marginally more efficient.

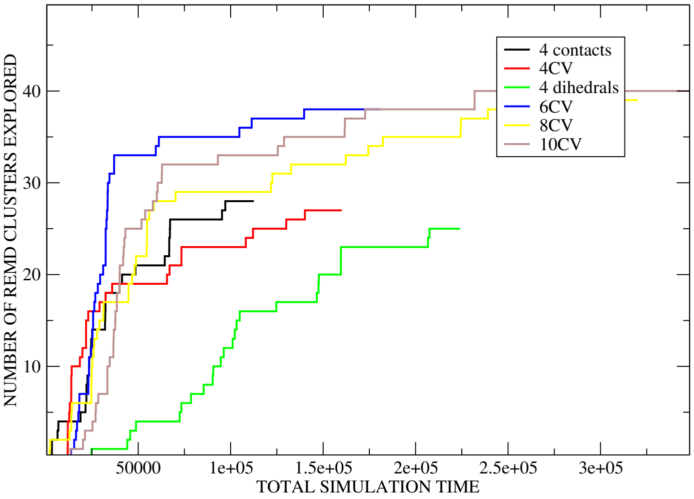


Figure 2.6: REMD reference clusters found as a function of time for the BE-META simulations with different collective variable setups.

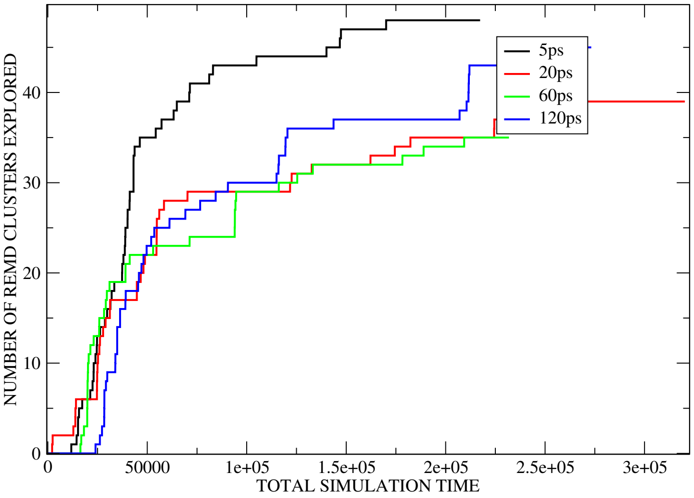


Figure 2.7: REMD reference clusters found as a function of time for the BE-META simulations that use different exchange times and the CVs **i-viii**.

CONCLUSIONS

In this thesis it is shown that bias exchange metadynamics allows the accurate reconstruction of the folding free energies of 1E0G using the coarse grained UNRES force field with a relatively small computational effort. It is also shown that this promising methodology is able to reproduce much more efficiently the results by Liwo et al. [18] with MREMD. This is very encouraging because BE-META can be easily implemented in an all-atom force field with explicit solvent.

It was shown that BE-META is a robust approach, and even though some of the sets of CVs used were small, it was found that all of the 9 BE-META simulations, employing different setups of CVs and exchange times, converged and found the folded state of the protein. This methodologhy is superior to conventional methods whenever it is not possible to describe the system with a few reaction coordinates.

Comparing the results from the different BE-META simulations, it is found that collective variables which can describe tertiary structure elements in proteins are necessary for an efficient exploration of the conformational space. A sufficient number of CVs provides an increased capability of each replica to diffuse through the conformational space.

BIBLIOGRAPHY

[1] G. Jayachandran, V. Vishal, and V.S. Pande, *J Chem Phys*, **124** 164902 (2006).

[2] C. D. Snow, E. J. Sorin, Y. M. Rhee and V. S. Pande,. *Ann. Rev. Biophys. Biomol. Struct.* **34**, 43 (2005).

[3] X. Daura, K. Gademann, B. Jaun, D. Seebach and W.F. van Gunsteren, *Angew Chem Int* Ed **38**, 236 (1999).

[4] E. A. Carter, G. Ciccotti, J. T. Hynes and R. Kapral, *Chem. Phys. Lett.*, **156**, 472 (1989).

[5] B. A. Berg and T. Neuhaus, *Phys. Rev. Lett.*, **68**, 9 (1992).

[6] G. Henkelman and H. Jonsson, *J. Chem. Phys.*, **111**, 7010 (1999).

[7] T. Huber, A. E. Torda and W. F. van Gunsteren, *J. Comput. Aided. Mol. Des.* **8**, 695-708 (1994).

[8] G. Henkelman, B. Uberuaga and H. Jonsson, *J. Chem. Phys.,* **113**: 9901 (2000).

[9] A. F. Voter, *Phys. Rev. Lett.*, **78**, 3908 (1997).

[10] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman, *J. Comput. Chem.*, **16**, 1339, (1995).

[11] B. Roux, *Comput. Phys. Comm.*, **91**: 275 (1995).

[12] M. Sprik and G. Ciccotti, *J. Chem. Phys.*, **109**, 7737 (1998).

[13] F. Wang and D.P. Landau, *Phys. Rev. Lett.* **86**, 2050 (2001).

[14] A. Laio and M. Parrinello, *Proc Natl. Acad. Sci. USA*, **99**, 12562, (2002).

[15] Y. Sugita and Y. Okamoto, *Chem Phys Lett.*, **314**, 141 (1999).

[16] S. Piana, and A. Laio, *J Phys Chem B*, **111**, 4553-4559(2007).

[17] S. Piana, A. Laio, F. Marinelli, M. Van Troys, D. Bourry, C. Ampe, and J. C. Martins, *J. Mol. Biol.* **375**, 460 (2008).

[18] A. Liwo, M. Khalili, C. Czaplewski, S. Kalinowski, S. Oldziej, K. Wachucik and H. A. Scheraga, *J. Phys. Chem. B* , **111**, 260-285 (2007).

[19] A. Liwo and C. Czaplewski, *J. Chem. Phys.*, **115**, 2323-2347 (2001).

[20] M. Parrinello and A. Rahman, *Phys. Rev. Letters*, **45**, 1196 (1980).

[21] F. Reif, *"Fundamentals of Statistical and Thermal Physics"*, McGraw-Hill Series (1997).

[22] S. Kumar and P. W. Payne and M. Vásquez, *J. Comp. Chem.*, **17**, 1269-1275 (1996).

[23] C. Jarzynski, *Phys. Rev. Lett.*, **78**, 2690 (1997).

[24] H. Risken,*"The Fokker-Planck Equation"*, Springer-Verlag, (1998).

[25] F. Zipoli and M. Bernasconi and R. Martoňák,*Eur. Phys. J. B*, **39**, 41 (2004).

[26] R. Martoňák and A. Laio and M. Parrinello, *Phys. Rev. Lett*, **90**, 75503 (2003).

[27] V. Churakov and M. Iannuzzi and M. Parrinello, J. Phys. Chem. B, **108**, 11567 (2004).

[28] A. Laio, A. Rodriguez-Fortea, F. Gervasio, M. Ceccarelli, M. Parrinello, *J. Phys. Chem. B*, **109**, 6714-6721 (2005).

[29] G. Bussi, F. Gervasio, A. Laio, M. Parrinello, *J. Am. Chem. Soc.*, **128**, 13435-13441 (2006).

[30] Bussi, G and Laio, A and Parrinello, M, Phys. Rev. Lett., **96**, 090601 (2006).

[31] M. Porto, U. Bastolla, H. E. Roman, and M. Vendruscolo, Phys. Rev. Lett., **92**, 218101 (2002).

[32] J.F. Sadoc1, *Eur. Phys. J. E*, **18**, 321333 (2005).

[33] A. R. Fersht, *Proc Natl. Acad. Sci. USA*, **97**, 1525 (2000).

[34] K. W. Plaxco, K. T. Simons, and D. Baker, *J. Mol. Biol.*, **277**, 985-994 (1998).

[35] R. Kubo, *J. Phys. Soc. Jpn.* **17**, 1100 (1962).

[36] R. A. Goldstein, Z. A. Luthey-Schulten, and P. G. Wolynes, *Proc. Natl. Acad. Sci. USA* **89**, 9029 (1992).

[37] D. K. Klimov and D. Thirumalai, *J. Chem. Phys.*, **109**, 4119 (1998).

[38] L. Verlet, *Phys. Rev.*, **159**, 98 (1967).

[39] L. Verlet, *Phys. Rev.*, **165**, 201 (1967).

[40] H. J. C. Berendsen, *Comp. Phys. Commun.*, **44**, 233 (1987).

[41] E. Forest, R.D. Ruth, *Physica D*, **43**: 105 (1990).

[42] A. Bateman, M. Bycroft, *J. Mol. Biol.*, **299**: 1113-1119 (2000).

[43] M. Pellegrini, E. M. Marcotte, M. J.Thompson, D. Eisenberg, T. O. and Yeates, *Proc. Natl Acad. Sci. USA*, **96**, 4285-4288 (1999).

[44] http://en.wikipedia.org/wiki/Maximumlikelihood

A APPENDIX

A.1 Amino acid

In Chemistry, an amino acid is a molecule containing both amine and carboxyl functional groups. In biochemistry, this term refers to α-amino acids with the general formula $H_2NCHRCOOH$, where R is an organic substituent (see Figure A.1).

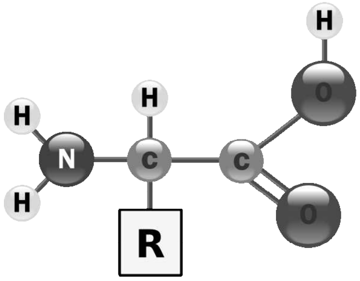


Figure A.1: The general structure of an α-amino acid, with the amino group on the left and the carboxyl group on the right.

In the α-amino acids, the amino and carboxylate groups are attached to the same carbon, which is called the αcarbon. The various alpha amino acids differ in which side chain (R group) is attached to their αcarbon. They can vary in size from just a hydrogen atom in glycine through a methyl group in alanine to a large heterocyclic group in tryptophan. Amino acids are usually classified by the properties of the side chain into four groups, they can make them behave like a weak acid, a weak base, a hydrophile if they are polar, and hydrophobe if they are nonpolar. There are only 20 standard amino acids used to form proteins.

A.2 Proteins

Proteins are large organic compounds made of amino acids arranged in a linear chain and joined together by peptide bonds between the carboxyl and amino groups of adjacent amino acid residues. The sequence of amino acids in a protein is defined by a gene and encoded in the genetic code, which specifies a primary sequence made of 20 L-α amino acids.

Like other biological macromolecules such as polysaccharides and nucleic acids, proteins are essential part of organisms and participate in every process within cells. Many proteins are enzymes that catalyze biochemical reactions and are vital to metabolism. Proteins also have structural or mechanical functions, such as actin and myosin in muscle and the proteins in the cytoskeleton, which form a system of scaffolding that maintains cell shape. Other proteins are important in cell signaling, immune responses, cell adhesion, and the cell cycle. Proteins are also necessary in animals' diet, since animals cannot synthesize all the amino acids they need and must obtain essential amino acids from food. Through the process of digestion, animals break down ingested protein into free amino acids that are then used in metabolism.

Most proteins fold into unique 3-dimensional structures. The shape into which a protein naturally folds is known as its native state. Although many proteins can fold unassisted, simply

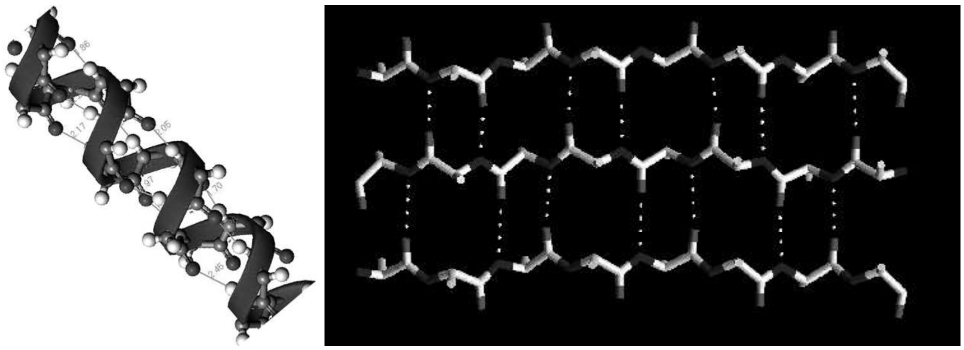


Figure A.2: Secondary structure elements of Proteins: a) α-helices, b) β-sheets

through the chemical properties of their amino acids, others require the aid of molecular chaperones to fold into their native states. Biochemists often refer to four distinct aspects of a protein's structure:

- Primary structure: the amino acid sequence.
- Secondary structure: regularly repeating local structures stabilized by hydrogen bonds. The most common examples are the α helix and β sheet (see Figure A.2). Because secondary structure are local, many regions of different secondary structure can be present in the same protein molecule.
- Tertiary structure: is the overall shape of a single protein molecule, i.e. the spatial relationship of the secondary structures to one another. Tertiary structures are generally stabilized not only by nonlocal interactions, most commonly the formation of a hydrophobic core, but also through salt bridges, hydrogen bonds, disulfide bonds, and even post-translational modifications. The term "tertiary structure" is often used as synonymous with the term fold.
- Quaternary structure: the shape or structure that result from the interaction of more than one protein molecule, usually called protein subunits in this context, which function as part of the larger assembly or protein complex.

A.3 Dihedral Angle

In geometry, the angle between two planes is called their dihedral or torsion angle. The dihedral angle of two planes can be seen by looking at the planes along their line of intersection, so the dihedral angle Φ_{AB} between two planes A and B is simply the angle between their two normal unit vectors $\vec{n}_A$ and $\vec{n}_B$:

$$\Phi_{AB} = cos^{-1}(\vec{n}_A.\vec{n}_B). \tag{A.1}$$

Since a plane can be defined in several ways (e.g., by vectors or points in them, or by their normal vectors), there are several equivalent definitions of a dihedral angle. Any plane can be defined by two non-collinear vectors lying in that plane; taking their cross product and normalizing yields the normal unit vector to the plane. Thus, a dihedral angle can be defined by three non-collinear vectors $\vec{b}_1, \vec{b}_2$ and $\vec{b}_3$. The vectors $\vec{b}_1$ and $\vec{b}_2$ define the first plane, whereas $\vec{b}_2$ and $\vec{b}_3$ define the second plane (see Figure A.3). The dihedral angle corresponds to an exterior spherical angle ϕ, which is a well-defined, signed quantity.

$$\phi = atan2(|\vec{b}_2|\vec{b}_1.[\vec{b}_2 \times \vec{b}_3], [\vec{b}_1 \times \vec{b}_2].[\vec{b}_2 \times \vec{b}_3]) \tag{A.2}$$

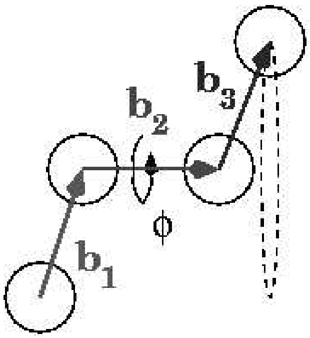


Figure A.3: Dihedral defined by three noncolineal vectors.

where the two-argument $atan2$ is defined as

$$atan2(y,x) = \begin{cases} \psi sgn(y) & x > 0, \\ \frac{\pi}{2} sgn(y) & x = 0, \\ (\pi - \psi) sgn(y) & x < 0. \end{cases}$$

with $tan(\psi) = |y/x|$.