# Insights on protein structure and dynamics from multiple biased molecular dynamics simulations

International School for Advanced Studies (SISSA-ISAS)

Trieste

Fabrizio Marinelli

Advisor: Alessandro Laio

# Contents

CHAPTER

**ONE**

# Introduction

The function of a protein is determined by its three dimensional structure and by its interactions with other biomolecules. From molecular biology and biochemical experiments the linear sequence of a high amount of proteins and several possible interaction partners are now available. However there is less information about their structure and mode of action. Although there are now well established bioinformatics tools[1–3] aimed at obtaining qualitative information on biochemical processes, many important questions cannot be easily addressed with a knowledge-based approach and are still open to investigation: what is the kinetic mechanism of folding and protein-ligand interaction? What is the driving force of the conformational changes involved in these processes (enthalpic or entropic)? What is the role of the solvent molecules?

Atomistic simulations in explicit solvent are a natural candidate to address quantitatively these important issues. These techniques are now very popular and have brought new insights in many fields of molecular physics ranging from solid state physics to biochemistry. The continuous increase of computing power[4] and of the reliability of empirical force fields [5–7] have led to simulations that resemble experiments more and more closely. The timescale of molecular dynamics (MD) simulations containing tens of thousands of atoms reach now routinely the $\mu$s time scale. Although this improvements increase significantly the range of applicability of MD many of the most interesting biological processes like enzyme reactions, protein folding and protein-protein/DNA interaction, are still unaffordable by plain MD as in most cases they occur on a time scale far beyond the $\mu$s. Furthermore, to extract reliable thermodynamics and kinetics information from the simulations data, the process under investigation must be reversibly observed several times[8]. Indeed more than 1 $\mu$s MD simulation is required to converge kinetic rates even for a simple system like the Ace-$Ala_3$-Nme peptide (see chapter 3).

To address this issue, simulations techniques were developed that allows accelerating rare events[9–30].

A technique that is particularly appropriate for studying complex conformational transitions and has been often used for studying biomolecules is metadynamics[30,

31]. This technique allows both accelerating rare events and computing free energies by adding to the normal forces a history-dependent bias which discourages the system to remain close to the same configuration. The bias potential acts on selected collective variables(CVs) that must be chosen *a priori*. For a complex system, if the relevant conformations are unknown, it is difficult to select the optimal CVs. In principle this problem could be alleviated by choosing simultaneously a large number of CVs, but this decreases significantly the acceleration factor. Typically for metadynamics it is not practical to use more than 3 CVs[31]. This is a common drawback of all the approaches that requires the choice of CVs like e.g umbrella sampling[32].

An alternative enhanced simulation technique that does not require chosing CVs is replica exchange molecular dynamics[28, 33] (REMD). This method consists in performing multiple simulations at different temperatures in which the sampling at a certain temperature is enhaced by exchanging the conformations between pairs of replicas. Unfortunately this approach is also affected by severe convergence problems when used with an explicit solvent potential energy function[34, 35]. To overcome this limitations hamiltonian replica exchange approaches were recently developed[36]. In these methods exchanges are performed between replicas having the same temperature but a different hamiltonian. For example the hamiltonian can be changed by adding a bias potential along a different CV in each replica[37]. These approaches require much fewer replicas for an enhanced sampling compared to standard REMD. Bias exchange metadynamics[38] (BE) is a recently developed approach that is based on a similar idea in a metadynamics framework. This method uses a time-dependent bias that acts on different CVs for each replica. Exchanges between the bias potentials are attempted at a certain frequency with a metropolis acceptance criterion in which the force field contributions cancels out. This scheme was shown[38] to allow an efficient exploration of the relevant configurations and to improve the convergence of free energies. Indeed exchanges among different replicas significantly reduce hysteresis problems[31] in the free energy reconstruction. Using BE it was possible to reversibly fold Trp-cage [38], villin headpiece, advillin headpiece together with two of their mutants [39] and Insulin chain B[40] using an explicit solvent force field, in less than 100 nanoseconds of simulation with only eight replicas. Recently this method was also used for exploring the mechanism of enzymatic reactions[41].

Another critical issue in molecular simulations is that long time scale MD and enhanced sampling techniques produce a large amount of data that must then be analyzed to obtain the relevant information. Several methods[42–47] have been developed to extract from MD trajectories the metastable conformations, to assign their occupation probability, and to compute the rates for transitions among them. These methods have the big advantage of reducing a complex dynamics in a high-dimensional configuration space to a Markov process describing transitions among a finite number of metastable states. This states are identified by a separation of time scales:

- the time required for the equilibration within each state,
- the time required for transitions among metastable states.

The latest process is typically the slowest. Their use is also justified by the empirical observation that for most of the biological processes, transitions among metastable conformations are stochastic in nature[42], due to thermal effects, the presence of small energetic barriers and collisions with solvent molecules. These models are very useful for extracting the long time scale behaviour of a complex system which can be directly compared with the relaxation times measured experimentally[42, 48]. Typically these methods are used for analyzing the outcome of a long ergodic MD trajectory or a large number of short MD trajectories. These trajectories can be obtained for example from worldwide distributed computing[49] or from REMD[42, 46, 47]. Using short trajectories to construct markovian models requires a proper choice of the procedure for dividing the conformational space (*binning*). If statistics is poor it can be difficult to obtain accurate kinetic rates between states as high free energy conformations are rarely explored. Reliable rates in those cases can be obtained by increasing the bin size, but this affects the reliability of the kinetic description. This problems can be in principle solved by increasing the sampling of non converged regions or by using biased trajectories for extracting the populations. However in enhanced sampling techniques the transitions probabilities are altered by the bias, and the trajectories cannot be used for computing the rates.

In this thesis we developed an approach that allows using trajectories generated by bias exchange metadynamics for constructing a reliable kinetic and thermodynamic model of complex biological processes. The approach aims at extracting the same information from a BE simulation as one can obtain from the analysis of a long ergodic MD run or of several shorter runs[42–47]. The rate model is constructed following three steps:

1. A cluster analysis is performed on the BE trajectories in a possibly extended CV space, assigning each configuration explored during the biased dynamics to a reference structure (bin) that is close by in a high-dimensional CV space.

2. Next, the equilibrium population of each bin is calculated from the BE simulations using a weighted histogram analysis method(WHAM)[13] exploiting the metadynamics bias potentials.

3. Finally, a kinetic model is constructed by assigning rates to transitions among bins. The transition rates are assumed to be of the form introduced in Ref. [50], namely to depend exponentially on the free energy difference between the bins with a prefactor that is determined by a diffusion matrix $D$ and by the bins relative position. The only free parameter in the model is $D$, as the free energies are already assigned. Following Ref. [46] $D$ is estimated maximizing the likelihood of an unbiased MD trajectory (not necessarily ergodic).

The model constructed in this manner is designed to optimally reproduce the long time scale dynamics of the system. It can be used, for example, for characterizing the metastable misfolded intermediates of the folding process[51]. The advantage of using biased trajectories, besides the acceleration of slow transitions, is a greatly enhanced accuracy of the estimated free energy in the transition regions.

This approach is first illustrated on the solvated Ace-$Ala_3$-Nme peptide (see chapter 3). This system is simple enough to allow benchmarking the results against a long standard MD simulation ($\sim 2\mu$s). Thermodynamics (e.g., bins and basins free energies) and kinetics (e.g., mean first passage time between attractors and first passage time distribution) properties calculated with the model are shown to be in excellent agreement with the extended MD simulations.

The same approach is then applied to much more complex systems. The first realistic application reported is the Trp-cage folding[48] (chapter 4). This is a designed 20-residue polypeptide that, in spite of its size, shares several features with larger globular proteins. Although the system has been intensively investigated experimentally and theoretically, its folding mechanism is not yet fully understood. Indeed, some experiments suggest a two-state behavior, while others point to the presence of intermediates. For the Trp-cage folding the kinetic model predicts a two state like relaxation time of ∼2300 ns in agreement with experiments (3100 ns). Despite of the single exponential kinetics the presence of several metastable intermediates was also detected, one of which is a molten globule structure that acts as a kinetic trap and is responsible of the observed relaxation time. Instead, non-compact structures relax to the folded state on the sub-microsecond timescale. Thus, surprisingly, the relaxation time measured by fluorescence may not be directly related to the "folding" transition, if one calls "folding" the transition from a random coil to the native state. The model also predicts the presence of a state at $C_\alpha$-RMSD of 4.4 Å from the NMR structure in which the Trp strongly interacts with Pro12. This state may explain the abnormal temperature dependence of the Pro12-$\delta$3 and Gly11-$\alpha$3 chemical shifts. The structures of the two most stable misfolded intermediates are also in agreement with NMR experiments on the unfolded protein.

The second application is on a larger biologically relevant protein: the insulin chain B[40]. Insulin is a highly investigated protein with the important function of regulating the glucose levels in the blood. It is composed by two chains (chain A and B) linked with two difulfide bridges. Here we investigated the folding mechanism of insulin chain B. This study is motivated by the following reasons: first, chain B of insulin is believed to retain much of its structure independently of chain A[52–54], second, structure-activity studies of insulin indicate the C-terminus of chain B as integral to receptor information[55–57] and the terminal regions of this chain are shown to be quite flexible. For this system the model allows identifying three main basins separated from each another by large free energy barriers. The characteristic native fold of chain B was observed in one basin, while the other two most populated basins contained molten-globule conformations stabilized by

electrostatic and hydrophobic interactions, respectively. Transitions between the three basins occur on the microsecond time scale. The implications and relevance of this nding to the folding mechanisms of insulin were investigated and are discussed in chapter 5

The last application discussed in this thesis is the study of the binding mechanism of a peptide substrate to the Human Immunodeciency Virus Type-1 Protease[58] (HIV-1 PR). This protein cuts polyproteins to smaller fragments and is essential for the virus life cycle. It is one of the main targets of anti-AIDS drug design as if inhibited the infection of the virus is reduced. Most FDA-approved HIV-1 PR mimic the structure of a fragment of the natural substrate. Investigating the mechanism by which substrates and drugs bind to this protein is crucial to understand the molecular rational of drug resistance. HIV-1 PR is a symmetric homodimer with a large binding pocket covered by two relatively flexible hairpins (flaps). These flaps during the binding of the natural substrate adopt an open conformation to allow the access to the binding site. Also for this system a kinetic model model was constructed analyzing the BE simulation data. The computed binding free energies and the kinetic constants measured were compatible with experimental results. Surprisingly, the binding mechanism extracted with the model shows that full opening of the flaps is not necessary for a short peptide, of size comparable to that of a drug to enter in the binding pocket. Thus, it can be inferred that natural substrate and drugs may bind throught different pathways and mutations of HIV-1 PR may affect in a different manner the binding pathway of the natural substrate and of the drugs.

CHAPTER

**TWO**

# Methods

## 2.1 Molecular Dynamics Simulations

MD is a useful tecnique to study the microscopic behavior of a molecular systems. In this section we will briefly summarize its most important features.

In atomistic simulation the main aspect are:

- The phase space sampling algorithm
- The choice of the interaction potential, $V(\mathbf{r})$, between the atoms of the system.

Several simulations approaches were developed in the last decades that differs in the method to sample the phase space. In MD simulations the atoms trajectories are extracted by integrating the Newton equations.

$$\mathbf{F}_i = m_i \mathbf{a}_i \text{ with } \mathbf{F}_i = -\frac{\partial V(\mathbf{r})}{\partial \mathbf{r}_i} \tag{2.1}$$

where $V(\mathbf{r})$, the potential, is a function of the atoms positions. In this equation one assumes that the motion of the atomic nuclei can described by classical dynamics. This can be considered a good approximation if the distance in the energetic levels of the involved degrees of freedom is $<< kT$, where $k$ is the Boltzmann constant and $T$ the temperature. Within the classical approximation statistical ensemble averages of selected observables can be performed using the ergodic hypothesis. Namely thermodynamics information on the system can be estimated as time averages.

### 2.1.1 Algorithms for integrating the Newton equations

MD simulations are based on the integration of the Newton equation. Due to the complexity of these equations an analytic solution is unaffordable and approximated methods must be used.

The features of a good integration algorithm can be summarized as follows:

- It must allow using a large time step $\Delta t$.
- It must be time reversible.

For the second point, as the Newton equations are time reversible also the algorithm is supposed to satisfy the same symmetry. The algorithms that are not time reversible do not normally preserve the phase space volume, i.e. they do not satisfy the Liouville theorem.

A good way to check the accuracy of the algorithm is to follow the temporal evolution of an observable $A$ that should be conserved (e.g. the total energy). In general a good algorithm must be such that:

$$\frac{\mid A(t_n) - A(t_0) \mid}{\langle A(t) \rangle} \ll 1, \qquad for\,(t_n - t_0) \gg \Delta t \tag{2.2}$$

The algorithm employed in the MD code used for this thesis is leap-frog[59], a variant of the Verlet algorithm:

$$\mathbf{v}_i(t + \Delta t/2) = \mathbf{v}_i\,(t - \Delta t/2) + \frac{\mathbf{F}_i(t)\Delta t}{m_i} \tag{2.3}$$

$$\mathbf{r}_i(t + \Delta t) = \mathbf{r}_i(t) + \mathbf{v}_i(t + \Delta t/2)\Delta t \tag{2.4}$$

where $\mathbf{F}_i$,$\mathbf{r}_i$,$\mathbf{v}_i$ are respectively the force acting on the atom $i$, the atom position and the atom velocity.

### 2.1.2 The interaction potential

The potential function $V(\mathbf{r})$ from which the forces used in MD are derived depends on the atomic coordinates $\mathbf{r}_i$.

$V(\mathbf{r})$ used in this thesis has the following expression:

$$\begin{aligned} V(r_1, r_2, \ldots, r_N) &= \sum_{bonds} \frac{1}{2} K_d (d - d_0)^2 + \qquad (2.5) \\ &+ \sum_{angles} \frac{1}{2} K_\theta (\theta - \theta_0)^2 + \\ &+ \sum_{improper\,dihedrals} \frac{1}{2} K_\xi (\xi - \xi_0)^2 + \\ &+ \sum_{dihedrals} K_\phi \left[1 + \cos(n\phi - \delta)\right] \\ &+ \sum_{ij\,LJ} \left[ \left( \frac{C_{ij}^{12}}{r_{ij}^{12}} - \frac{C_{ij}^{6}}{r_{ij}^{6}} \right) \right] \qquad (2.6) \\ &+ \sum_{ij\,coulomb} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \end{aligned}$$

The first two terms (two and three body interactions respectively) represent the bonds and angles potentials, that are approximated by harmonic functions. The third and fourth term describe 4 body interactions. Improper dihedral terms are typically described by an harmonic function. Instead proper dihedrals are described by periodic functions (i.e. cosine functions) of a given periodicity $n$. The last two terms are a Lennard-Jones ($LJ$) potential and a coulomb potential between pair ($ij$) of atoms. The parameters used in this kind of potentials are typically obtained from quantum chemical calculations and experimental data (e.g. crystallographic data, spectroscopic data, etc). Among the popular sets of parameters (force fields) for MD simulations of proteins we can cite for example AMBER[60], GROMOS[61], CHARM[62] and OPLS[63]. They all use the potential function expression given above for all the atoms of the simulated system except for the GROMOS(and CHARMM19 force field) force field in which a united atom description is used for non-polar hydrogens.

In MD simulations the description of the solvent (water for most of the biologically interesting systems) can be explicit or implicit. In the first case solvent molecules with a full atomistic force field description are added in the simulation box at the experimental density. In the implicit solvent description the solvent is treated as a dielectric medium in which the system is embedded. This is clearly a more approximated description but it is also computationally much more efficient since in many practical cases the solvent constitutes the majority of the atoms. In this thesis we used for all the MD simulations the AMBER03[64] force field with an explicit description of the solvent given by the TIP3P[65] water model.

### 2.1.3 Constraints

Constraints are used in MD to fix bonds to their equilibrium value. This allows increasing the simulation time step $\Delta t$. Constraining the bond lenght does not alter significantly the statistics as these are quantum degrees of freedom being mostly in their ground state at the normal simulation temperature. Using the bonds constraints it is possible to use $\Delta t \sim 2fs$ [66] (2-4 times larger than the one that can be used without constraints). A common method to introduce constraints is the algorithm SHAKE [66], in which after each time step the atoms positions iteratively are modified in order to satisfy the constraint.

SHAKE may have convergence problems when applied to large planar groups and its implementation could hinder the efficiency of computing. To improve these aspects the LINCS algorithm was recently introduced[67]. For water molecules it is also possible to use an analytic solution of SHAKE called SETTLE[68]. Both LINCS and SETTLE (for the water molecules) have been used in all the simulations performed in this thesis.

### 2.1.4 Boundary conditions

To simulate a finite size system, boundary conditions are needed to avoid artifacts

near the border of the simulation box. Typically periodic boundary conditions (PBC) are used. In this scheme short range non bonded interactions are calculated using the minimal image convention (only the nearest replica is considered). Typically a cut-off radius ($R_c$) is used for $LJ$ interactions of the order of 10 Å. To avoid interactions between a particle and its periodic image each box side must be larger than $2R_c$.

The coulomb energy is instead treated considering the full periodicity of the system. For a periodic lattice made by $N$ particles it is given by:

$$E = \frac{1}{8\pi\epsilon_0} \sum_{|n|=0}^{\infty}{}^{\star} \left[ \sum_{i=0}^{N} \sum_{j=0}^{N} \frac{q_i qj}{|r_{ij} + n|} \right] \tag{2.7}$$

where $n$ indicates the periodic images, $i$,$j$ the particles and the symbol $\star$ indicates that the summation does not contain the term with $i = j$ if $n = 0$.

The method used in this thesis to evaluate this energy is Particle Mesh Ewald[69].

### 2.1.5 Pressure and Temperature coupling

To compare simulation results with experiments it is necessary to control pressure and temperature using thermostats like Nose Hoover[70], or Berendsen[71]. These approaches introduce extra auxiliary variables that are evolved with suitable dynamics, designed with scope of endorsing the exploration of the correct temperature or pressure.

### 2.1.6 The molecular partition function

If a classical MD simulation is performed in the NVT ensemble the atoms positions distribution is canonical (here we assume to move only classical degree of freedom, while the bond length and angles are constrained to their equilibrium value):

$$\rho_{can}(\mathbf{r}/\mathbf{b^0}) = \frac{\exp\left(-V(\mathbf{r}/\mathbf{b^0})/T\right)}{\int \exp\left(-V(\mathbf{r'})/T\right)\delta(\mathbf{b} - \mathbf{b^0})d\mathbf{r}} \ , \tag{2.8}$$

where $V(\mathbf{r})$ is the potential energy function (force field), $\mathbf{r}$ are the atom positions, $\mathbf{b}$ is the vector of bond length and $\mathbf{b^0}$ is their equilibrium value. The notation $\mathbf{r}/\mathbf{b^0}$ is used to specify that the variations of the atoms coordinates are performed in the surface of constrained bond lengths (the constraint is that they are fixed to $\mathbf{b^0}$). Boltzmann constant unit are used in eq. 2.8, i.e. $k_B = 1$, where $k_B$ is the Boltzmann constant. This units of measure will be assumed for all the other equations of this Thesis in which $k_BT$ is involved. Here it is assumed that the bonds are mostly in their quantum ground state. Otherwise the summation over all the accessible quantum levels should be included.

## 2.2 Enhanced sampling techniques

Although MD is a powerful tool for exploring the microscopic behaviour of molecular systems it may suffer from limitations especially if applied to the study of complex conformational transitions. Indeed the sampled MD distribution is given by eq. 2.8 and this means that only configurations having energy within few $T$ from the global minimum will be explored. This is a clear advantage considering that only relevant configurations will be explored, but if there is an energy (in general a free energy) barrier between two relevant configurations one of the two may not be explored in a finite simulation time. To overcome this problem avoiding the use of simplified models (e.g. "coarse-grained" force fields or an implicit solvent description) advanced simulation methods are required.

Broadly speaking these methods can be classified in four categories, according to their scope and range of applicability:

1. Methods aimed at reconstructing the probability distribution or enhancing the sampling as a function of one or a few predefined CVs. For instance, in a chemical reaction one would choose the distance between two atoms that have to form a bond or, in the study of nucleation, the size of the nucleus and enhance the sampling as a function of these coordinates. Examples of these methods include thermodynamic integration [9, 72], free energy perturbation[10], umbrella sampling [11], conformational flooding [12], weighted histogram techniques [13, 73, 74], Jarzynski's identity-based methods[14, 75], adaptive force bias [15, 76], steered MD [16] and adiabatic molecular dynamics [17]. These approaches are very powerful but require a careful choice of the CVs that must provide a satisfactory description of the reaction coordinate. If an important variable is forgotten they suffer from hysteresis and lack of convergence. Moreover, when more than a few CVs are used, the computational performance rapidly degrades as a function of the number of variables.

2. Methods aimed at exploring the transition mechanism and constructing reactive trajectories[18], such as nudged elastic band [77], finite-temperature string method[19, 78], transition path sampling [20, 79, 80], transition interface sampling[81], milestoning [21] and forward flux method [82]. These methods do not require in most of the cases the explicit definition of a reaction coordinate, but require the *a priori* knowledge of the initial and final states of the process that has to be simulated. For instance, if applied to the study of folding, these methods require the knowledge of the folded and "unfolded" state[83].

3. Methods for exploring the potential energy surface and localizing the saddle points that correspond to the transition states like, for example, eigenvalue following [22], the dimer method[23], hyperdynamics[24], multiple-time scale accelerated molecular dynamics[25] event-based relaxation[26]. These

approaches are extremely powerful for exploring potential energy surfaces of low dimensionality, but their reliability degrades with the complexity of the system. Indeed, for very large or complex systems the number of possible transition states surrounding a minimum becomes rapidly too large for a deterministic search. Even if strategies have been designed to alleviate this problem that are effective in some special cases[26], in solvated systems the concept of saddle point on the potential energy surface becomes fuzzy, and these approaches cannot easily be applied.

4. Methods in which the phase space is explored simultaneously at different values of the temperature, such as parallel tempering [27] and replica exchange [28], or as a function of the potential energy, such as multicanonical MD [84] and Wang-Landau[29]. These approaches are very general and powerful, however they are not immune from some of the limitations listed in point 1. Indeed, these methods exploit more or less explicitly the potential energy as a generalized CV. In several cases, ordered and disordered states may correspond to the same value of potential energy, or be present in the thermal ensemble at the same temperature. This may lead to hysteresis and convergence problems[35].

### 2.2.1 Flattening the free energy profile

In order to introduce in more detail the enhanced sampling approach used in this thesis we here consider a simple example: the transition from $\alpha$ to $\beta$ of the Ace-$Ala_3$-Nme peptide (hereafter $Ala_3$, see Fig. 2.1). This process can be described using the $\psi$ backbone dihedral angle.

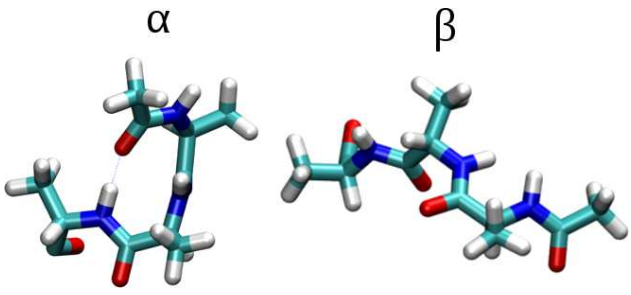


Figure 2.1: $\alpha$ **and** $\beta$ **structures of** $\mathbf{Ala_3}$. Structures of $Ala_3$ corresponding to the central $\psi$ dihedral angle in $\alpha$ and $\beta$ positions. The hydrogen bond formed in the $\alpha$ conformation is displayed in the figure

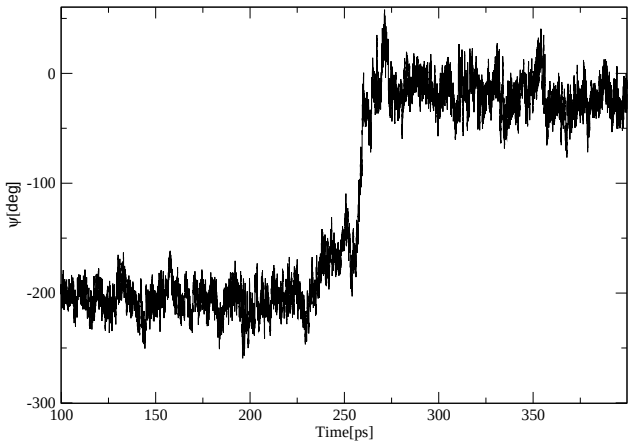


Figure 2.2: **$\alpha$-$\beta$ transition in Ala$_3$** $\psi$ dihedral angle of Ala$_3$ as a function of time taken from a MD simulation of Ala$_3$ in explicit TIP3P[65] water

In Fig 2.2 it is shown the temporal behavior of $\psi$ taken from a MD simulation of Ala$_3$. The peptide remains for a long time in a $\beta$ conformation (time=100-250 ps). Then a thermal fluctuation occurs stimulating the transition to a $\alpha$ conformation ($\psi\sim0$). The transition itself, if it happens, is relatively fast (it occurs within 5 ps). The behaviour observed in Fig. 2.2 is prototypical of a metastability between two states $\alpha$ and $\beta$. This can be quantified in terms of a time scale separation between the equilibration within each state ($\alpha$ or $\beta$; fast) and the transitions among the states ($\alpha$-$\beta$ transitions; slow).

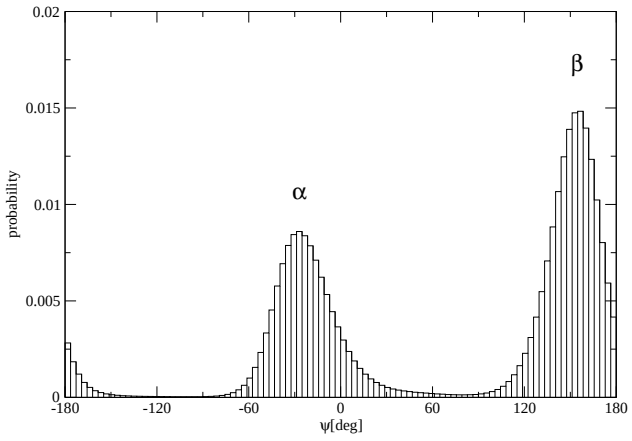


Figure 2.3: **Histogram along the $\psi$ dihedral angle of Ala$_3$**.The histogram is calculated from a $\sim$300 ns MD simulation of Ala$_3$ in explicit solvent

In Fig. 2.3 it is shown the histogram of $\psi$ calculated from a much longer MD trajectory of 300 ns. As it can be noted the probability $p(\psi)$ of being in $\alpha$ or $\beta$ is

higher than the one to be in an intermediate conformation. From a thermodynamic point of view this can be quantified by looking at the free energy as a function of $\psi$, evaluated from the histogram as $F(\psi) = -T \log p(\psi)$. This function is displayed in Fig. 2.4

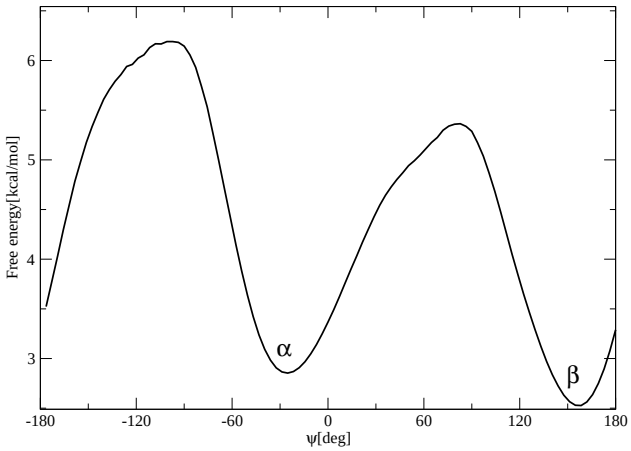


Figure 2.4: **Free energy as a function of the the $\psi$ dihedral angle of Ala$_3$.** The free energy is calculated from the histogram outlined in Fig. 2.3 using the equation $F(\psi) = -RT \log p(\psi)$.

The height of the maximum of the free energy function is related to the rate of $\alpha$-$\beta$ interconversion. The lower is the height of the barrier the faster will be the interconversion rate, i.e. the frequency of interconversion will increase if the barrier is decreased. If one could artificially decrease the height of the barrier the process would be no longer rare and could be observed most frequently during an MD simulation. The highest interconversion rate is in fact obtained for a flat free energy surface.

The free energy in the canonical ensemble, is formally defined as follows. Here we suppose to have a system simulated using a MD simulation in the NVT ensemble. The atoms position distribution is described by eq. 2.8. To simplify the notation hereafter we omit the constraint specification in eq. 2.8. If we call $\mathbf{s} = (s_1, s_2, ..., s_N)$ a vector of N collective variables, the free energy F($\mathbf{s}$) can be obtained by integrating eq. 2.8 on all the other degrees of freedoms.

$$F(\mathbf{s}') = -T \ln \int \rho_{can}(\mathbf{r}) \delta(\mathbf{s}' - \mathbf{s}(\mathbf{r})) d\mathbf{r} = -\ln \frac{\int \exp\left(-V(\mathbf{r})/T\right) \delta(\mathbf{s}(\mathbf{r}) - \mathbf{s}') d\mathbf{r}}{\int \exp\left(-V(\mathbf{r}')/T\right) d\mathbf{r}'} \,, \tag{2.9}$$

Under this framework a flat free energy surface can be obtained by adding a potential $V_G(\mathbf{s}(\mathbf{r})) = -F(\mathbf{s}(\mathbf{r}))$ to $V(\mathbf{r})$. In fact substituting $V(\mathbf{r})$ with $V(\mathbf{r}) - F(\mathbf{s}(\mathbf{r}))$ in eq. 2.9 one readily obtains:

$$\frac{\int \exp\left(-V(\mathbf{r})/T + F(\mathbf{s}(\mathbf{r}))/T\right)\delta(\mathbf{s}(\mathbf{r}) - \mathbf{s}')d\mathbf{r}}{\int \exp\left(-V(\mathbf{r}')/T + F(\mathbf{s}(\mathbf{r}'))/T\right)d\mathbf{r}'} \propto \frac{\exp\left(-F(\mathbf{s}')/T + F(\mathbf{s}')/T\right)}{\int \exp\left(-V(\mathbf{r}')/T + F(\mathbf{s}(\mathbf{r}'))/T\right)d\mathbf{r}'}$$

This corresponds to an uniform distribution in $\mathbf{s}$. This imply that if one knows the free energy profile as a function of the CVs, the rare event could be accelerated by adding to the MD force field an artificial potential equal to $-F(\mathbf{s}(\mathbf{r}))$. This idea of adding an additional potential to let the system explore different regions of the configurations space is common in many free energy calculation methods. Indeed in the standard umbrella sampling formulation[13] a bias potential $V_G(\mathbf{s}(\mathbf{r}))$ is added to the potential energy function and the canonical sampling is then recovered by means a reweighting procedure:

$$\rho_{can}(\mathbf{r}) \propto \rho_{umbrella}(\mathbf{r}) \exp(V_G(\mathbf{s}(\mathbf{r}))) \tag{2.10}$$

where $\rho_{umbrella}(\mathbf{r})$ is the distribution function sampled during the umbrella sampling simulation. Using this reweighting procedure also $F(\mathbf{s})$ can be reconstructed with an accuracy that depend on the shape of the bias potential. The highest accuracy is obtained using $V_G(\mathbf{s}(\mathbf{r})) = -F(\mathbf{s}(\mathbf{r}))$ [13].

## 2.3 Metadynamics

The metadynamics method[30] can be considered as an evolution of umbrella sampling in which is not necessary to specify $V_G$. Like umbrella sampling it requires a preliminary identification of a few CVs which are assumed to be able to describe the process of interest. In this method the normal MD forces are combined with forces derived from a history-dependent potential $V_G(\mathbf{s}, t)$ defined as a sum of Gaussians of height $w$ centered along the trajectory in CVs space. This manner of biasing the evolution was first used by the taboo search method [85] and, in the context of MD, by the local elevation method [86]. A similar approach is also used in the Wang and Landau algorithm [29] and adaptive force bias[15]. As we will discuss in the following in metadynamics the sum of Gaussians is exploited to reconstruct iteratively an estimator of the free energy.

The metadynamics bias potential can be written as:

$$V_G(\mathbf{s}, t) = w \sum_{t'=\tau_g, 2\tau_g, \ldots}^{t'<\tau} \exp\left(-\sum_{i=1}^{N} \frac{\left(s_i - s_i(t')\right)^2}{2\delta s_i^2}\right) \tag{2.11}$$

where $\tau$ is the total simulation time $\tau_g$ is the frequency at which the Gaussians are added and $\delta s_i$ is the width of the Gaussian for each CV.

This potential has the remarkable property of gradually filling $F(\mathbf{s})$. In the example of fig. 2.5 starting from one free energy minimum the lowest transition state from that minimum is the first to be explored. If the metadynamics simulation is continued, at the end the bias potential will fill all the available CV space.

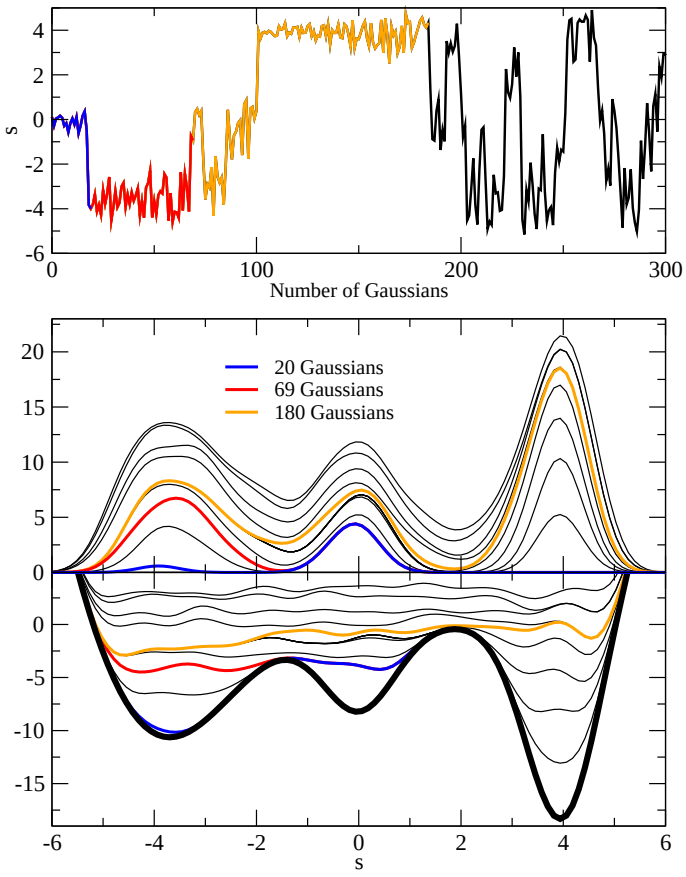


Figure 2.5: Upper panel: trajectory of a one-dimensional system evolved by a Langevin equation on the 3-minima potential represented in the lower panel. The dynamics is biased with a metadynamics potential $V_G$ as defined by Eq. 2.11. The parameters are $\delta s = 0.4$, $w = 0.3$ and $\tau_G = 300$. Middle panel: time evolution of the metadynamics bias potential $V_G$. Blue line: $V_G$ as when the first minimum is filled and the system "escapes" to the second minimum; Red line: $V_G$ as when also the second minimum is filled; Orange line: $V_G$ when the entire profile is filled and the dynamics becomes diffusive. Lower panel: time evolution of the sum of the metadynamics potential $V_G$ and of the external potential, represented as a thick black line.

The novel idea of metadynamics is that if the walker is able to keep memory of all the positions in which he has deposited the Gaussians he will be able to reconstruct a negative image of the underlying free energy. More precisely, one assumes the time dependent potential defined by the sum of Gaussians deposited up to time $t$ provides an unbiased estimate of the free energy in the region explored during the dynamics. This property, that does not follow from any ordinary thermodynamic identity, such as umbrella sampling [11], was postulated on a heuristic basis in Ref. [30], and afterward verified empirically in several systems of increasing complexity. Successively[87](see section below), it was shown that this property derives from rather general principles, and can be demonstrated rigorously for a system evolving under the action of a Langevin dynamics.

### 2.3.1 Estimating the error

In order to best allocate the available computational resources to study with metadynamics a given system, it is useful to estimate *a priori* the performance of the method, and choose the parameters to obtain the best possible accuracy in a given simulation time. The accuracy and efficiency of the free energy reconstruction is determined by the Gaussian width $\delta s$ , the Gaussian height $w$ and the Gaussians deposition time $\tau_G$. The parameter $w$ and $\tau_G$ determine the height and the rate at which the Gaussians are placed. In Ref. [88] it has been shown that the error on the reconstructed profile is approximately determined by the ratio $w/\tau_G$ and not by $w$ and $\tau_G$ separately. Indeed, adding a Gaussian of height, say, 0.2 kcal/mol every ps is approximately equivalent to adding a Gaussian of height 0.1 kcal/mol every 0.5 ps, as long as $\tau_G$ remains much shorter than the time required to fill the free energy basin. In order to understand how $\delta s$ and $w/\tau_G$ influence the accuracy and construct an explicit expression for the error, consider first the idealized case in which the CV evolves following an over-damped Langevin dynamics:

$$ds = -\frac{1}{T} D \frac{dF(s)}{ds} dt + \sqrt{2D} dW(t) \tag{2.12}$$

where $dW(t)$ is a Wiener process and $D$ is the diffusion coefficient. The motion of the walker described by Eq. 2.13 is assumed to satisfy reflecting boundary conditions at the boundary of a region $\Omega$. The evolution of this system under the action of metadynamics is modeled adding a history-dependent term to the free energy:

$$ds = \frac{1}{T} D \frac{d}{ds} \left( F(s) + \int_0^t dt' g(s, s(t')) \right) dt + \sqrt{2D} dW(t) \tag{2.13}$$

where $g(s, s')$ is a kernel that specifies how fast the metadynamics potential changes. In normal implementation $g$ is a Gaussian of width $\delta s$ and height $w/\tau_G$:

$$g(s, s') = \frac{w}{\tau_G} \exp\left( -\frac{(s - s')^2}{2\delta s^2} \right)$$

It should be remarked that in real systems the evolution of the CVs is described by a much more complex stochastic differential equation[89, 90] with memory and inertial terms. Still, as it will be discussed in the following, the quantitative behavior of metadynamics is reproduced rather precisely by this simple model. This is because, if the CV set is properly chosen, all the relaxation times are smaller than the time required to fill the free energy wells.

Equation 2.13 describes a non-Markovian process in CV space. In fact, the forces acting on the CVs depend explicitly on their history. Due to this non-Markovian nature it is not clear if, and in which sense, the system can reach a stationary state under the action of this dynamics. In Ref.[87] a formalism was introduced which allows mapping this history-dependent evolution into a Markovian process in the original variable and in an auxiliary field that keeps track of

the visited configurations. Defining

$$\varphi(s,t) = \int_0^t dt' \delta(s - s(t')) \tag{2.14}$$

equation 2.13 can in fact be written as

$$\begin{aligned} d\varphi &= \delta(s - s(t))\, dt & (2.15)\\ ds &= \frac{1}{T} D \left( \frac{d}{ds} \left( F(s) + \int ds' \varphi(s',t)\, g(s', s(t)) \right) \right) dt + \sqrt{2D} dW(t) & (2.16) \end{aligned}$$

These equations are fully Markovian, i.e. the state of the system at time $t + dt$, $(s(t+dt), \varphi(s, t+dt))$, depends only on the state of the system at time $t$, $(s(t), \varphi(s,t))$. Using this property Eq. 2.13 can be rigorously analyzed to obtain, for instance, its long time behavior.

The history-dependent potential at time $t$ is related to $\varphi(s',t)$ by

$$V_G(s,t) = \int ds' \varphi(s',t)\, g(s,s') \tag{2.17}$$

In order to characterize the average properties of a system described by Eq. 2.16 it is convenient to consider the probability $P(s,[\varphi],t)$, to observe $s$ and the field realization $\varphi$. $P(s,[\varphi],t)$ satisfies a Fokker-Planck equation that can be directly derived from Eq. 2.16 using standard techniques[91]. Ref.[87] shows that, for large $t$, $P(s,[\varphi],t)$ converges to a distribution $P_\infty([\varphi])$ that does not depend on $s$. This distribution is Gaussian in functional space and is given by

$$P_\infty([\varphi]) \propto \exp\left( \frac{D}{2T} \int ds ds' \left(\varphi(s') - \varphi_0(s')\right) \partial_s^2 g(s,s') \left(\varphi(s) - \varphi_0(s)\right) \right) \tag{2.18}$$

where $\varphi_0(s')$ is defined in such a way that its convolution with the kernel $g$ gives minus the free energy of the system $F(s)$ :

$$\varphi_0(s') \;:\; \int ds' \varphi_0(s')\, g(s,s') = -F(s) \tag{2.19}$$

Using Eq. 2.18 it is straightforward to prove that the average value of $V_G(s,t)$ over several independent metadynamics runs is exactly equal to $-F(s)$. In fact, denoting by $\langle\cdot\rangle_M$ the average over several metadynamics realizations, Eq. 2.17 gives

$$\begin{aligned} \langle V_G(s)\rangle_M &= \int ds' g(s,s') \langle\varphi(s')\rangle_M = \\ &= \int ds' g(s,s') \int d\varphi P_\infty(\varphi)\, \varphi = \\ &= \int ds' g(s,s')\, \varphi_0(s') = -F(s) \end{aligned} \tag{2.20}$$

The metadynamics error in $s$ is given by the expected deviation of $V_G(s,t)$ from

$-F(s)$:

$$\varepsilon^2(s) = \left\langle \left(V_G(s) + F(s)\right)^2 \right\rangle_M = \tag{2.21}$$

$$= \left\langle \left(V_G(s) - \langle V_G(s)\rangle_M\right)^2 \right\rangle_M \tag{2.22}$$

Using the explicit expression for the probability to observe a given $\varphi$, Eq. 2.18, allows computing the error that 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 temperature. For example, in a cubic domain of side $S$ in $d$ dimensions [87] the error is

$$\overline{\varepsilon}^2 = \frac{S^2 w T}{D\tau_G}\left(\frac{\delta s}{S}\right)^d (2\pi)^{\frac{d}{2}} \sum_k \frac{1}{\pi^2 k^2} \exp\left(-\frac{k^2\pi^2}{2}\left(\frac{\delta s}{S}\right)^2\right) \tag{2.23}$$

where the sum is performed over all the $d$ dimensional vectors of integers of non-zero norm.

Eq. 2.23 is an expression of the error of a single metadynamics simulation as a function of the simulation parameters $w/\tau_G$, $\delta s$, and the system-dependent parameters, $T$, D and S. The error increase linearly with the filling speed $w/\tau_G$ and is proportional to the inverse of D. The error increase also with $\delta s$ but the functional form is influenced also by the dimensionality $d$.

### 2.3.2 Disadvantages of standard metadynamics

Although metadynamics is a powerful technique used to accelerate rare events and to reconstruct the free energy it suffers of several limitations that prevent its applicability for complex processes like protein folding or protein-protein interaction[31]. A major problem is the filling speed that exponentially decreases with the dimensionality. This limits the use of metadynamics to not more than 3 CVs. For slowly diffusing systems the error in the free energy reconstruction may increase significantly as indicated by eq. 2.23. Finally similarly to other methods that reconstruct the free energy in a set of generalized coordinates, the reliability of metadynamics is strongly influenced by the choice of the CVs. What happens if a relevant CV is neglected? In this respect, a simple metadynamics run on an idealized model can be enlightening. Consider the Z-shaped two-dimensional free energy depicted in Fig. 2.6. If a metadynamics simulation is performed biasing only CV1 and neglecting CV2 the simulation, that is started in basin B, is not able to perform in due time a transition towards A, and metadynamics goes on overfilling this minimum. A transition is finally observed only when the height of the accumulated Gaussians will largely exceed the true barrier height. This behavior will continue indefinitely without ever reaching a situation in which the free energy grows evenly like in the example of Fig. 2.5.

A similar behavior is often observed in real cases and is a strong indication that an important CV is missing.

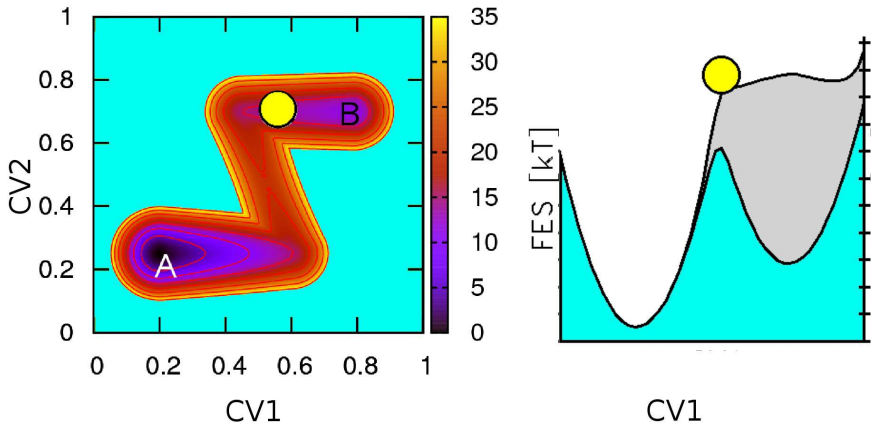


Figure 2.6: The effect of neglecting a relevant degree of freedom. Left side: 2D Z shaped potential. Right side: the trajectory of a metadynamics simulation generated using only s1 as CV. Transitions from A to B are not properly described by CV1, causing strong hysteresis in the reconstructed free energy.

## 2.4 Bias exchange metadynamics

BE[38] was developed in an attempt to alleviate the limitations of metadynamics cited above. A *large set* of CVs that are expected to be relevant for the process under investigation is chosen. A number NR (number of replica) of MD simulations (*walkers*) are run in parallel, biasing each walker with a metadynamics bias acting on just one or two collective variables. In BE the sampling is enhanced by attempting, at fixed time intervals of a few ps, swaps of the bias potentials between pairs of walkers. The swap is accepted with a probability

$$\min\left\{1, \exp\left[\left(V_G^a\left(r^a,t\right)/T + V_G^b\left(r^b,t\right)/T - V_G^a\left(r^b,t\right)/T - V_G^b\left(r^a,t\right)\right)/T\right]\right\}\ , \tag{2.24}$$

where $r^a$ and $r^b$ are the coordinates of walker a and b and $V_G^{a(b)}\left(r,t\right)$ is the metadynamics potential acting on the walker a(b). In this manner, each trajectory evolves through the high dimensional free energy landscape in the space of the CVs sequentially biased by different low dimensional potentials acting on one or two CVs at each time. A clear advantage of this method is that a large number of different variables can simultaneously be biased, and, ideally, the dimensionality of the space explored by metadynamics can be made so large that all the residual barriers orthogonal to the reaction coordinates can be crossed in the available simulation time. However the result of the simulation is not a free energy in several dimensions, but several low dimensional projections of the free energy surface along all the collective variables. Due to the efficaciously multidimensional nature of the bias the system is capable of exploring a complex free energy landscape with high efficiency. Moreover, since all the replicas are simulated at the same temperature, it is not necessary to use a large number of replicas for systems described with explicit solvent, as it is instead compulsory in replica exchange and parallel tem-

pering [27, 28]. In BE the convergence of the bias potential to the corresponding free energy projection is monitored like in standard metadynamics: if the CVs are properly chosen and describe all the "slow" degrees of freedom, after a transient time, $V_G$ reaches a stationary state in which it grows evenly fluctuating around an average that estimates the free energy[31].

## 2.5 Multidimensional free energy from bias exchange metadynamics

Here we introduce a method for calculating free energies in several dimensions from a bias exchange metadynamics(BE)[38] simulation. The approach aims at extracting the same information from a BE simulation as one can obtain from the analysis of a long ergodic MD run or of several shorter runs[42–47]. The method relies on the projection of the BE trajectory on the space defined by a set of variables, which are assumed to describe the relevant physics of the system. These variables are not necessarily the ones that are used for the BE simulation, can be chosen *a posteriori* and can be up to $\sim 10$, greatly extending the scope of normal metadynamics.

The idea is to exploit the low-dimensional free energies obtained from BE to estimate, by a weighted-histogram procedure, the free energy of a finite number of structures that are representative of all the configurations explored by the system. These structures are determined by performing a cluster analysis, namely grouping all the frames of the BE trajectories in sets (*bins*) in which all the elements are close to each other in CV space. Since the scope of the overall procedure is constructing a model that describes also the kinetic properties of the system, it is important that the bins are defined in such a way that they satisfy three properties:

1. The bins must cover densely all the configuration space explored in BE, including the barrier regions.

2. The distance in CV space between nearest neighbor bin centers must not be too large. This, as it will be shown in the following, is necessary for constructing the rate model.

3. The population of each bin in the BE trajectory has to be significant, otherwise its free energy estimate will be unreliable.

A set of bins that satisfy these properties is here defined dividing the CV space in small hypercubes forming a regular grid. The size of the hypercube is defined by its side in each direction: $ds = (ds_1, ds_2, ..., ds_n)$ where $n$ is the number of collective variables. This determines directly how far the bin centers are. Each frame of the BE trajectory is assigned to the hypercube to which it belongs and the set of frames contained in a hypercube defines a bin. This very simple approach is used here only in order to keep directly under control the distance between the bins, but the results

presented in this section also apply if the cluster analysis is performed with one of the other approaches that have been developed for this scope [42, 43, 92]. The canonical weight of each bin is estimated by a weighted histogram procedure based on the metadynamics bias potentials. Denote by $V_G^i(s,t)$ the history-dependent potential generated by the walker $i$ up to time $t$ expressed in Boltzmann constant units. After a certain time $t_F$ metadynamics has explored all the available CV space. At the end of the simulation, an estimate of the free energy is the average of $V_G^i(s,t)$ after $t_F$ [31, 93]:

$$V^i(s) = \frac{1}{t_{tot} - t_F} \int_{t_F}^{t_{tot}} dt\ V_G^i(s,t) \tag{2.25}$$

where $t_{tot}$ is the total simulation time. During the last part of the BE run $V_G^i(s,t)$ fluctuates around $V^i(s)$ (except for an irrelevant additive constant that grows linearly with time), but these fluctuations are small if the deposition rate of the Gaussians is not excessive. In order to keep the error induced by these fluctuations under control it is convenient to consider two different bias potentials of the form of Eq. 2.25, one obtained extending the integral from $t_F$ up to $(t_{tot}+t_F)/2$, the other from $(t_{tot}+t_F)/2$ up to $t_{tot}$. Only the configurations collected after $t_F$ in which the two bias potentials are consistent within a few $T$ ($T$ for $Ala_3$ and $2T$ for the Trp-cage) are retained for further analysis. The unbiased probability to observe bin $\alpha$ is estimated on walker $i$ using the standard umbrella sampling reweighting formula:

$$p_\alpha^i = \sum_{k \in \Omega_\alpha^i} e^{\frac{1}{T}\left(V^i\left(s_k^i\right) - f^i\right)} \tag{2.26}$$

where $f^i$ is a parameter that fixes the normalization and $\Omega_\alpha^i$ is the set of frames in the walker $i$ that are assigned to bin $\alpha$. The $p_\alpha^i$ -s are used to construct the best possible estimate of the probability $p_\alpha$ of observing bin $\alpha$. This requires estimating the error on $p_\alpha^i$. Here it is assumed that the error on a bin free energy estimate is:

$$\sigma^2\left(p_\alpha^i\right) = g \sum_{k \in \Omega_\alpha^i} e^{\frac{2}{T}\left(V^i\left(s_k^i\right) - f^i\right)} = g p_\alpha^i e^{\frac{1}{T}\left(\overline{V}_\alpha^i - f^i\right)} \cong g p_\alpha e^{\frac{1}{T}\left(\overline{V}_\alpha^i - f^i\right)} \tag{2.27}$$

where $g$ is a constant that takes into account the correlation time and

$$\overline{V}_\alpha^i = T \log\left(\frac{\sum_{k \in \Omega_\alpha^i} e^{\frac{2}{T} V^i\left(s_k^i\right)}}{\sum_{k \in \Omega_\alpha^i} e^{\frac{1}{T} V^i\left(s_k^i\right)}}\right) . \tag{2.28}$$

In order to simplify the notation we have neglected the position-dependence of $g$. In the last passage in Eq. (2.27) the fact that $p_\alpha^i$ is an unbiased estimator of $p_\alpha$ is assumed. The combined probability $p_\alpha$ is now written as a linear combination of the $p_\alpha^i$-s, namely $p_\alpha = C \sum_i \pi_\alpha^i p_\alpha^i$, where the weights $\pi_\alpha^i$ are parameters that have to be determined and $C$ is normalization constant. The expected error on $p_\alpha$ is $\sigma^2(p_\alpha) = C^2 \sum_i \left(\pi_\alpha^i\right)^2 \sigma^2\left(p_\alpha^i\right)$. The optimal weights for each bin $\alpha$ are determined

separately minimizing this error with the constraint $\sum_i \pi_\alpha^i = 1$. This gives $\pi_\alpha^i = e^{\frac{1}{T}\left(f^i - \overline{V}_\alpha^i\right)}(\sum_j e^{\frac{1}{T}\left(f^j - \overline{V}_\alpha^j\right)})^{-1}$ and, finally,

$$F_\alpha = -T \log p_\alpha = -T \log \sum_i \pi_\alpha^i p_\alpha^i = -T \log \frac{\sum_i \overline{n}_\alpha^i}{\sum_j e^{\frac{1}{T}\left(f^j - \overline{V}_\alpha^j\right)}} \tag{2.29}$$

with $\overline{n}_\alpha^i = \sum_{k \in \Omega_\alpha^i} e^{\frac{1}{T}\left(V^i\left(s_k^i\right) - \overline{V}_\alpha^i\right)}$. The constants $f^i$ are obtained iteratively from the condition

$$e^{-\frac{1}{T}f^i} = \frac{1}{\sum_\alpha \overline{n}_\alpha^i} \sum_\alpha e^{-\frac{1}{T}\overline{V}_\alpha^i} p_\alpha = C \frac{1}{\sum_\alpha \overline{n}_\alpha^i} \sum_\alpha e^{-\frac{1}{T}\overline{V}_\alpha^i} \frac{\sum_k \overline{n}_\alpha^k}{\sum_j e^{\frac{1}{T}\left(f^j - \overline{V}_\alpha^j\right)}} \,. \tag{2.30}$$

The free energy estimate given by Eq. 2.29 is affected by an error

$$\sigma^2\left(F_\alpha\right) = T^2 \frac{\sigma^2\left(p_\alpha\right)}{p_\alpha^2} = \frac{gT^2}{\sum_i \overline{n}_\alpha^i} \tag{2.31}$$

consistently with what is found in the normal weighted histogram analysis method[13].

Within this framework, the average value of an observable $O$ can be calculated, using the estimated free energies, as

$$\langle O \rangle = \frac{\sum_\alpha O_\alpha \exp\left(-F_\alpha/T\right)}{\sum_\alpha \exp\left(-F_\alpha/T\right)} \tag{2.32}$$

where the sums run over all the bins, $T$ is the temperature and $O_\alpha$ is the average value of $O$ in the bin $\alpha$. If the bin size is small enough, the bias potentials are approximately constant for the configurations belonging to the same bin [39]. Thus $O_\alpha$ can be reliably estimated as the arithmetic average of $O$ in all the configurations explored by the BE trajectory belonging to the bin $\alpha$. Corrections deriving from the variation of the bias potentials inside a bin have also been considered but they lead to negligible effects for small $ds$.

The enthalpy $H_\alpha$ of bin $\alpha$ is obtained averaging the enthalpy over the structures belonging to the bin. The entropy $S_\alpha$ is estimated as $S_\alpha = \left(H_\alpha - F_\alpha\right)/T$. Neglecting the dependence of the entropy on the temperature, the free energy at a temperature $T'$ different from $T$ is estimated as

$$F_\alpha\left(T'\right) = H_\alpha - T'S_\alpha = H_\alpha - \frac{T'}{T}\left(H_\alpha - F_\alpha\left(T\right)\right) \tag{2.33}$$

with an error of $\sigma^2(F_\alpha(T')) = (\frac{T'}{T})^2\sigma^2(F_\alpha(T)) + (1 - \frac{T'}{T})^2\sigma^2(H_\alpha)$.

Using Eq. 2.32 together with Eq. 2.33 allows extrapolating the average value of the observables for a few tens of K around the temperature at which the simulation is performed. The uncertainty on $O$ can be derived at each temperature from the error on $F_\alpha$, $H_\alpha$, and $O_\alpha$ using error propagation on Eqs. 2.32 and 2.33:

$$\sigma^2(\langle O\rangle) = \frac{\sum_\alpha e^{-2F_\alpha(T')/T'}\left[\frac{(\langle O\rangle - O_\alpha)^2}{T'^2}\sigma^2(F_\alpha(T')) + \sigma^2(O_\alpha)\right]}{\left(\sum_\beta e^{-F_\beta(T')/T'}\right)^2} \tag{2.34}$$

where $\sigma^2(O_\alpha)$ is the standard deviation of $O$ inside bin $\alpha$.

## 2.6 Kinetic model from a multi-dimensional free energy

The free energy provides direct information on the probability distribution along the CVs (thermodynamics), but can also be used to extract kinetic information, e.g. the transition rate. For example the height of the barrier can be used in transition state theory to estimate the rate between two states [94]. Another way to extract kinetic information from the free energy is to assume that the dynamics along the CVs can be approximated by a diffusion process:

$$\dot{s_i}(t) = -\sum_j \frac{D_{ij}}{T}\frac{\partial F(\mathbf{s})}{\partial s_j} + W_i \tag{2.35}$$

where $\mathbf{D}$ is the diffusion matrix, $F(\mathbf{s})$ the free energy and $\mathbf{W}$ is a white noise that obeys to the following relations:

$$\langle \mathbf{W}(t)\rangle = 0 \tag{2.36}$$

$$\langle W_i(t')W_j(t)\rangle = 2D_{ij}\delta(t'-t) \tag{2.37}$$

The corresponding Fokker-Planck equation is:

$$\frac{\partial \rho(\mathbf{s},t)}{\partial t} = \sum_i -\frac{\partial J_i}{\partial s_i} = \sum_i\sum_j \frac{\partial}{\partial s_i} D_{ij}\left(\frac{\rho}{T}\frac{\partial F(\mathbf{s})}{\partial s_j} + \frac{\partial \rho(\mathbf{s},t)}{\partial s_j}\right) \tag{2.38}$$

where $\rho(\mathbf{s},t)$ is the probability density. The vector $\mathbf{J}$ describe a flux and is introduced in eq. 2.38 to show that the Fokker-Planck eq. can be written as a continuity equation. For a one dimensional process in which the free energy is flat eq. 2.38 reduces to the Fick's second law: $\frac{\partial \rho}{\partial t} = D\frac{\partial^2 \rho}{\partial x^2}$. Eq. 2.35 and 2.38 require estimating the diffusion matrix $\mathbf{D}$. Moreover in realistic systems, the dynamics of a process can be described by eq. 2.38 only on an appropriate time scale. This will be discussed in details in the following.

Different approaches are possible for integrating eq. 2.38. In order to construct a bin based kinetic model, it is useful to write it as a master equation:

$$\frac{\partial p(\alpha)}{\partial t} = \sum_\beta k_{\beta\alpha}p(\beta) - k_{\alpha\beta}p(\alpha) \tag{2.39}$$

This requires discretizing the CVs space in bins. If these form a regular grid an explicit expression for the transition rates can be given[46]:

$$k_{\alpha\beta} = k^0_{\alpha\beta} e^{-\frac{1}{2T}\left(F_\beta - F_\alpha\right)} \tag{2.40}$$

where $\alpha$ and $\beta$ are two neighbouring bins and $k^0_{\alpha\beta} = k^0_{\beta\alpha}$ are the rates associated to simple diffusion on a flat free energy surface. This form ensures that the limiting probability distribution of the dynamics is correct, namely that the probability to observe bin $\alpha$ at long times scales is proportional to $\exp\left(-F_\alpha/T\right)$.

If the bins form a hypercubic grid in CV space the rates $k^0_{\alpha\beta}$ can be exactly expressed as a function of the (possibly position-dependent) diffusion matrix $\mathbf{D}^\alpha$ and of the hypercube side $ds$ [46]. In the following to simplify the notation we denote by $\mathbf{D}$ the diffusion matrix appearing in the transition rate between two bins $\alpha$ and $\beta$ assuming that $\mathbf{D}$ is the average of $\mathbf{D}^\alpha$ and $\mathbf{D}^\beta$[46]. In one dimension the bins are labeled by a single integer $(i)$ and, following Refs [46, 50], $k^0_{(i)(i\pm1)} = \frac{D}{ds^2}$ and zero otherwise. In $d$ dimensions the bins are labeled by $d$ integers $(i_1, i_2, \cdots, i_d)$. If $\mathbf{D}$ is diagonal, the one-dimensional expression for the rates can be generalized straightforwardly. If $\mathbf{D}$ is non-diagonal the only rates different from zero are those in which one or two of the components of $(i_1, i_2, \cdots, i_d)$ vary by one:

$$\begin{aligned} k^0_{(\cdots,i_k,\cdots)(\cdots,i_k\pm1,\cdots)} &= \frac{D_{kk}}{ds_k^2} - \sum_{j\neq k}\left|\frac{D_{jk}}{ds_k ds_j}\right| \\ k^0_{(\cdots,i_k,\cdots,i_j,\cdots)(\cdots,i_k\pm1,\cdots i_j\pm1,\cdots)} &= \max\left(\frac{D_{jk}}{ds_k ds_j}, 0\right) \end{aligned} \tag{2.41}$$

This form of the rates can be derived discretizing the Fokker-Planck equation for diffusion on the regular grid defined by the hypercube centers. To do this we use a general formulation to derive finite difference schemes for the numerical solution of partial differential equation, using Taylor expansion. As we want to derive the prefactor $k^0_{\alpha\beta}$ of eq. 2.40 we consider eq. 2.38 for a flat free energy surface, i.e. $\frac{\partial F(\mathbf{s})}{\partial s_j} = 0$:

$$\frac{\partial \rho(\mathbf{s},t)}{\partial t} = \sum_i \sum_j \left(\frac{D_{ij}\partial^2\rho(\mathbf{s},t)}{\partial s_i \partial s_j}\right) \tag{2.42}$$

If the space is discretized eq. 2.42 can be approximated as:

$$\frac{\partial \rho(\mathbf{s}^\alpha,t)}{\partial t} \sim a^0_\alpha \rho(\mathbf{s}^\alpha,t) + \sum_\beta k^0_{\beta\alpha}\rho(\mathbf{s}^\beta,t) \tag{2.43}$$

where $\beta$ indicate the $\alpha$ neighbouring bins. The choice and the number of neighbouring bins affects the properties of the discretization scheme. $k^0_{\beta\alpha}$ and $a^0_\alpha$ are treated here as parameters to be determined. We Taylor expand $\rho$ in the neighbours of $s^\alpha_i$:

$$\begin{aligned}\rho(\mathbf{s}^\beta,t) &= \rho(\mathbf{s}^\alpha,t)+\sum_i(s_i^\beta-s_i^\alpha)\frac{\partial\rho(s_\alpha,t)}{\partial s_i}\\ &+ \frac{1}{2!}\sum_i\sum_j(s_i^\beta-s_i^\alpha)(s_j^\beta-s_j^\alpha)\frac{\partial^2\rho(s_\alpha,t)}{\partial s_i\partial s_j}\\ &+ \frac{1}{3!}\sum_i\sum_j\sum_k(s_i^\beta-s_i^\alpha)(s_j^\beta-s_j^\alpha)(s_k^\beta-s_k^\alpha)\frac{\partial^3\rho(s_\alpha,t)}{\partial s_i\partial s_j\partial s_k}+o(|\Delta\mathbf{s}|^4)\end{aligned} \tag{2.44}$$

We now substitute this expression in eq. 2.43;

$$\begin{aligned}\frac{\partial\rho(\mathbf{s}^\alpha,t)}{\partial t} &= a_\alpha\rho(\mathbf{s}^\alpha,t)+\left(\sum_\beta k^0_{\beta\alpha}\right)\rho(\mathbf{s}^\alpha,t)\\ &+ \sum_i\left(\sum_\beta k^0_{\beta\alpha}(s_i^\beta-s_i^\alpha)\right)\frac{\partial\rho(s_\alpha,t)}{\partial s_i}\\ &+ \frac{1}{2!}\sum_i\sum_j\left(\sum_\beta k^0_{\beta\alpha}(s_i^\beta-s_i^\alpha)(s_j^\beta-s_j^\alpha)\right)\frac{\partial^2\rho(s_\alpha,t)}{\partial s_i\partial s_j}\\ &+ \frac{1}{3!}\sum_i\sum_j\sum_k\left(\sum_\beta k^0_{\beta\alpha}(s_i^\beta-s_i^\alpha)(s_j^\beta-s_j^\alpha)(s_k^\beta-s_k^\alpha)\right)\frac{\partial^3\rho(s_\alpha,t)}{\partial s_i\partial s_j\partial s_k}+...\end{aligned} \tag{2.45}$$

This equation coincides with eq. 2.42 with a second order accuracy if:

$$a^0_\alpha+\sum_\beta k^0_{\beta\alpha} = 0 \tag{2.46}$$

$$\sum_\beta k^0_{\beta\alpha}(s_i^\beta-s_i^\alpha) = 0 \tag{2.47}$$

$$\sum_\beta k^0_{\beta\alpha}(s_i^\beta-s_i^\alpha)(s_j^\beta-s_j^\alpha) = 2D_{ij} \tag{2.48}$$

$$\sum_\beta k^0_{\beta\alpha}(s_i^\beta-s_i^\alpha)(s_j^\beta-s_j^\alpha)(s_k^\beta-s_k^\alpha) = 0 \tag{2.49}$$

It is straightforward verifying that these relations are satisfied by rates $k^0_{\beta\alpha}$ of the form given by eq. 2.41.

The error of this procedure scales as the square of the distance between neighboring bins[46]. Indeed the rates $k^0_{\beta\alpha}$ are proportional to the inverse of the square distance between neighbouring bins, instead the error in eq. 2.45 scales as $|\Delta\mathbf{s}|^4$. At finite grid spacing the accuracy can be improved allowing transitions between non-neighboring bins. This is because higher order relations can be imposed having a larger number of neighbours. Eq. 2.43 involves the probability densities, but can

be easily converted in a relation involving bins probabilities by multiplying all the terms in eq. 2.43 for the volume factor $\prod_i ds_i$.

### 2.6.1 Estimating the diffusion matrix by maximum likelihood

In this section we describe a manner for estimating the $\mathbf{D}$ matrix entering in the rates of eq. 2.40. The diffusion matrix is estimated using the approach of Ref. [46], in which one maximizes the likelihood that a given MD trajectory is generated by a rate equation of the form Eq. 2.40. Computing $\mathbf{D}$ requires first generating at least one MD trajectory without the metadynamics bias. The accuracy of the procedure can be improved, if the relevant metastable states of the system are known, by running several independent MDs starting from these states. Otherwise one can select at random a few conformations along the BE trajectory and use these as the initial conditions for MD. The trajectory (or the set of trajectories) is then mapped at a time lag $\Delta t$ onto the bins $(\alpha(0), \alpha(\Delta t), \alpha(2\Delta t), \cdots)$. Then several KMC trajectories are run with an initial guess for $\mathbf{D}$, starting from the bins visited by the MD trajectory. Using the KMC trajectories one computes the conditional transition probabilities at a time lag $\Delta t$ $p_{\mathbf{D}}(\gamma|\beta)$ among all the pairs of bins $\beta$, $\gamma$ visited by the trajectory. This is evaluated by counting transitions between the bins:

$$p_{\mathbf{D}}(\gamma|\beta) = \frac{n(\gamma(\Delta t)|\beta(0))}{n(\beta)}$$

where $n(\gamma(\Delta t))$ is the number of times the KMC trajectory is found in bin $\gamma$ at time $\Delta t$ being in bin $\beta$ at time zero, and $n(\beta)$ is the number of times the trajectory visits bin $\beta$. This procedure is slightly different from the one used in Ref. [46], where $p_{\mathbf{D}}(\gamma|\beta)$ is calculated by diagonalizing the rate matrix, which in the cases considered in this thesis has a very large size (of the order of $10^5 \times 10^5$). The notation $p_{\mathbf{D}}$ indicates that these probabilities depend parametrically on $\mathbf{D}$.

Using these probabilities one evaluates the logarithm of the likelihood to observe the sequence of bins obtained by MD. This is given by

$$L(\mathbf{D}) = \log \prod_t p_{\mathbf{D}}(\alpha(t+\Delta t)|\alpha(t)) \ . \tag{2.50}$$

$L(\mathbf{D})$ is then maximized as a function of $\mathbf{D}$. This can be done by simulated annealing, starting from an initial guess of $\mathbf{D}$ and iterating until the likelihood reaches a plateau. As outlined in Ref[47], the diffusion matrix found in this way depends in general on the chosen time lag. A common behavior is that by increasing the time lag $\Delta t$ the elements of the diffusion matrix converge to a well defined value. This means that after this $\Delta t$ the dynamics between bins is close to Markovian and is well approximated by the model proposed. As a consequence only transition that occur on a time scale bigger than $\Delta t$ are correctly described by this model.

Applying this procedure the prefactor of the rate Eq. 2.40, which has the form of a jump process among a discrete set of states, is directly optimized. This is a clear

advantage with respect to other methods for computing $\mathbf{D}$, in which a continuous evolution of the collective variables is assumed. Moreover, as the free energies $F_\alpha$ are known, the only variational parameter is $\mathbf{D}$ and comparably short trajectories are sufficient to determine it with a good statistical accuracy. A good accuracy of the free energy is of course crucial in this approach as they enter in Eq. 2.40 as an exponential.

A model constructed in this manner allows studying the evolution of the system on very long time scales using (e.g.) KMC. However, the reliability of the model relies on several approximations. If the free energy is flat, by construction the model gives the correct diffusive behavior but if $F \neq 0$ deviations from this behavior are observed when the bin size is too large. On the other hand, a small bin size can hinder the accuracy of the free energies. Thus, both large and small bin size may alter the quality of the kinetic model due to bad description of the underlying free energy surface or inaccurate sampling. Moreover even if there are no problems related to the bin size, describing the dynamics with Eq. 2.40 amounts to neglecting memory effects. This approximation can be particularly severe if an important variable is not included explicitly in the model. The model is expected to be reasonably accurate if the memory time is much smaller than the typical transition time (usually between metastable sets) that one wants to measure.

In the next chapter we will discuss the validity of these hypothesis in real applications.

CHAPTER

**THREE**

# Ace-$Ala_3$-Nme: a benchmark system

The approach presented in chapter 2 is here illustrated on the Ace-$Ala_3$-Nme peptide (hereafter $Ala_3$). $Ala_3$ is a simple polypeptide that has been extensively used as a benchmark system. Although small, this system shows several protein-like features, such as intramolecular hydrogen bonds and a fragment of $\alpha$-helical structure. Since the system is small, it is possible to characterize carefully its equilibrium and kinetic properties by extended MD simulations. The results obtained from the BE simulations of $Ala_3$ are benchmarked against the ones obtained from a long standard MD simulation ($\sim 2\ \mu s$). For this system the model is capable of reproducing with excellent accuracy the kinetics and thermodynamics observed in the unbiased run.

## 3.1 Computational setup

$Ala_3$ was simulated using BE and MD in explicit solvent using the GROMACS suite of programs[95, 96] and the AMBER03[64] force field. $Ala_3$ was placed in a periodic cubic box containing 1052 TIP3P water [65] molecules. The time step was set to 2 fs and the LINCS [67] algorithm was used to fix the bond lengths of $Ala_3$. The SETTLE algorithm[68] was used to fix angle and bond length of water molecules. Electrostatic and Lennard-Jones interactions were calculated with a cutoff of 1.0 nm. Lennard-Jones interactions are switched off smoothly from 0.9 nm to 1.0 nm. The neighboring list was updated every 5 steps and the cut-off distance for the short-range neighbor list was set to 1.1 nm. The Particle Mesh Ewald method [69, 97] was used to treat long-range electrostatic interactions with a maximum grid spacing for the fast Fourier transform of 0.12 nm and an interpolation order of 4. A constant temperature of 300 K was achieved by coupling the system to a Berendsen thermostat [71] with a characteristic time of 0.1 ps. A constant pressure of 1 bar was achieved by coupling the system to a Berendsen barostat [71] with a characteristic time of 2.5 ps. Several independent MD simulations were performed, with a length varying between $\sim$ 30 ns and $\sim$ 300 ns, for a cumulative time of 1.8 $\mu$s.

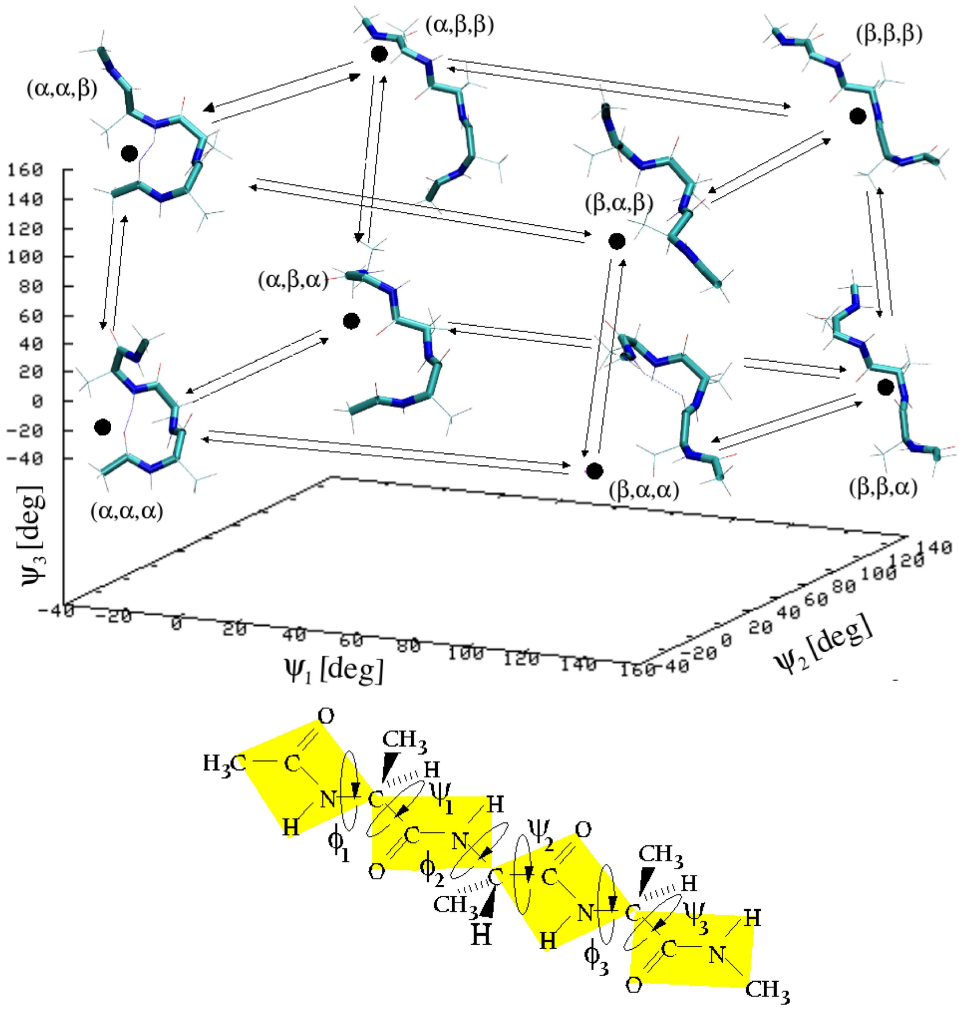


Figure 3.1: **Structures of the attractors for the relevant free energy basins of $Ala_3$ found in the MD and BE simulations.** Inset: Schematic picture of $Ala_3$ test system. The dihedral angles $\phi$ and $\psi$ displayed in the figure are chosen as CVs for the BE simulation. They are labeled with suffix according to their position along the chain.

The conformations of $Ala_3$ are here specified by its six backbone dihedral angles ($\phi_i, \psi_i$, where $i = 1, 2, 3$) (see Fig. 3.1, inset). Following Refs. [98–100], $\phi_2$ and $\psi_2$ (central Ramachandran angles of $Ala_3$) were considered in order to assign the main conformations of the system, denoted by $PP_{II}$ ($\phi_2 \in [-90°, -30°]$, $\psi_2 \in [120°, 180°]$), $\beta$ ($\phi_2 \in [-180°, -120°]$, $\psi_2 \in [120°, 180°]$), $\alpha_R$ ($\phi_2 \in [-90°, -30°]$, $\psi_2 \in [-90°, 0°]$), and $\alpha_L$ ($\phi_2 \in [30°, 90°]$, $\psi_2 \in [0°, 90°]$). Besides the latter conformational states, eight different states were also considered in order to analyze the results of the kinetic model. These are the free energy minima with the three dihedrals $(\psi_1, \psi_2, \psi_3)$ in the $\alpha$ or $\beta$ region of the Ramachandran plane, namely $(\alpha, \alpha, \alpha)$, $(\alpha, \alpha, \beta)$, etc. (see Fig. 3.1).

The system was also simulated using BE [38] exploiting the six dihedral angles (see Fig. 3.1, inset) as CVs. Each CV was biased in a different walker. Hence,

NR= 6, and each walker evolved under the action of a one-dimensional metadynamics potential acting on one of the six CVs. The width and the height of the Gaussians used in metadynamics were 0.1 rad and 0.1 kJ/mol respectively. A new Gaussian was added to the metadynamics potential every 1 ps. Exchanges of the bias potentials between pairs of walkers were attempted every 10 ps. Three independent BE simulations of 30 ns each (one simulation consist of 30 ns for each replica) were carried out in order to check the reproducibility of the results.

## 3.2 Results

**BE simulation of $Ala_3$.** The system was simulated using BE [38] employing the six backbone dihedral angles as CVs for biasing the dynamics. As expected BE improves the sampling of saddle regions (see Fig. 3.2B) and less stable minima (e.g. the $\alpha_L$ region of the Ramachandran angle).

The results of the BE simulation of $Ala_3$ are six one-dimensional free energy profiles (see Fig. 3.4), each a function of one of the six dihedral angles. After approximately 5 ns the free energy profiles do not change significantly anymore (see also Fig. 3.2A and 3.3), except for the fluctuations that are typical of metadynamics.

The profiles extracted from the three independent BE runs do not show sizable differences (root mean square deviation (RMSD) of free energy $\approx$ 0.4 kJ/mol, maximum deviation $\approx$ 1 kJ/mol), and they agree with the MD results within the error bars (RMSD of free energy $\approx$ 0.8 kJ/mol, maximum deviation $\approx$ 2 kJ/mol, see Fig. 3.2B). The profiles obtained applying eq.2.25 averaging on the last 10 ns of a BE simulations are shown in Fig. 3.4.

**Bin-based thermodynamic model**. Even in this simple system the different structures (see Fig. 3.1) are defined by the value of at least two of the six collective variables and thus one-dimensional free energies are not very insightful. In order to estimate the relative probability of the different structures we applied the approach introduced in the chapter 2. The six dimensional space was divided in hypercubes of side $ds$ ("bins"). Due to the high dimensionality of the space the number of bins increases rapidly by decreasing the box side. Reducing $ds$ from 40° to 30° the number of bins that are visited increases from 70,000 to 300,000. On the other hand, for small $ds$ most of the bins are visited only a few times, and this hinders the accuracy of the free energy estimate (see Eq. 2.31). The free energy of each bin was calculated for several choices of the bins size $ds$ applying Eq. 2.29 to the BE simulation data. The free energy profile entering in Eq. 2.29 was calculated using eq.2.25 with $t_F$=5 ns. In order to reduce the error induced by the time dependent fluctuations, the bias potential was averaged independently in the two halves of the interval $[5ns, 30ns]$ (see chapter 2). Only configurations collected after 5 ns in which the two averaged potentials are consistent within $T$ are retained for further analysis. The free energies were evaluated independently from the $\sim 2\ \mu s$ equilibrium MD trajectories by applying the standard thermodynamic relation $F_\alpha = -T \log n_\alpha$,

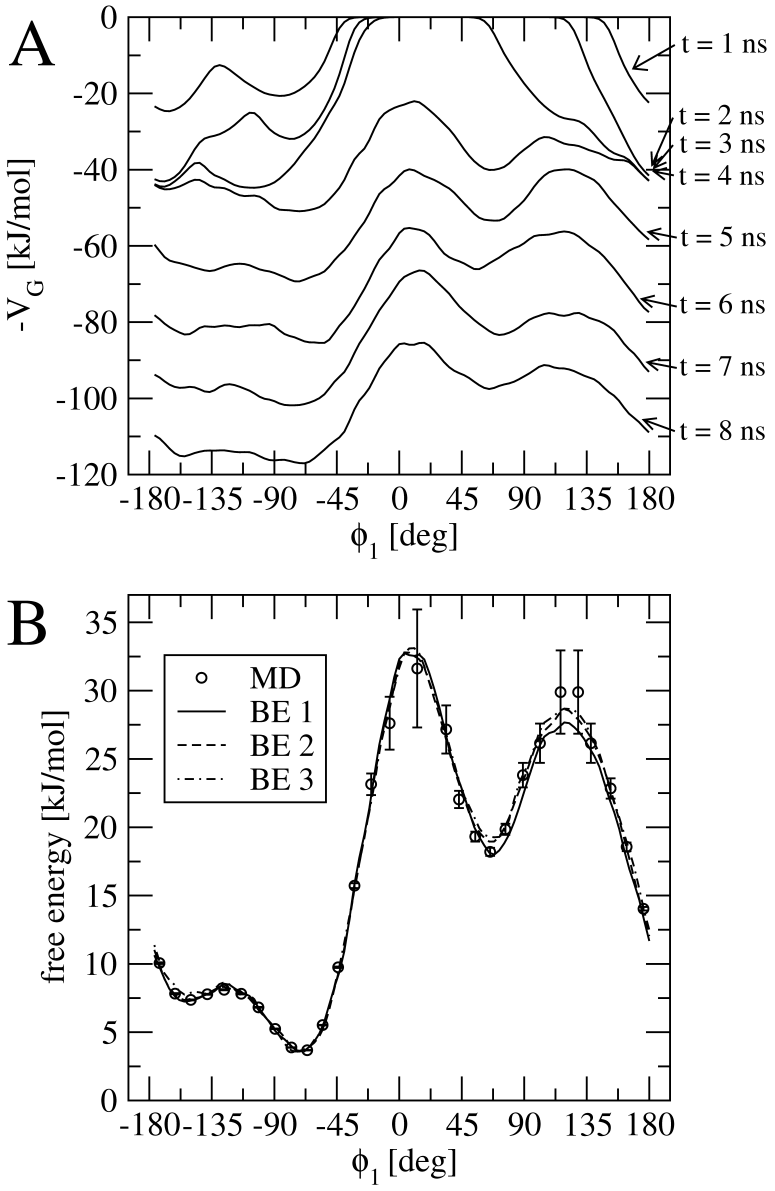


Figure 3.2: **Free energy profiles as a function of $\phi$1 (see Fig. 3.1) for Ala$_3$** Panel A: time evolution of $-V_G(s,t)$ during a BE simulation between 1 and 8 ns; after $\sim$5 ns the bias potential converges and grows parallel to itself. Panel B: Free energy profile from the 1.8 $\mu$s MD simulation compared with the profiles obtained from three independent BE simulations. The 3 BE profiles are obtained by applying eq. 2.25.

where $n_\alpha$ is the population of the bin $\alpha$. In Fig. 3.5, it is shown that the free energies calculated in the two manners correlate very well, especially at low free energy, where MD is accurate. Indeed, the horizontal stripes at high $F$ in Fig. 3.5 correspond to bins that are explored only a small number of times in MD. In Fig. 3.5, inset, it is shown the distribution of the relative error $(F_{BE} - F_{MD})/\sigma_{MD}$ where $F_{MD}$ and $F_{BE}$ are the free energies of the bins computed by MD and BE and $\sigma_{MD}$ is the error on $F_{MD}$ estimated by Eq. 2.31 on the MD trajectory (using

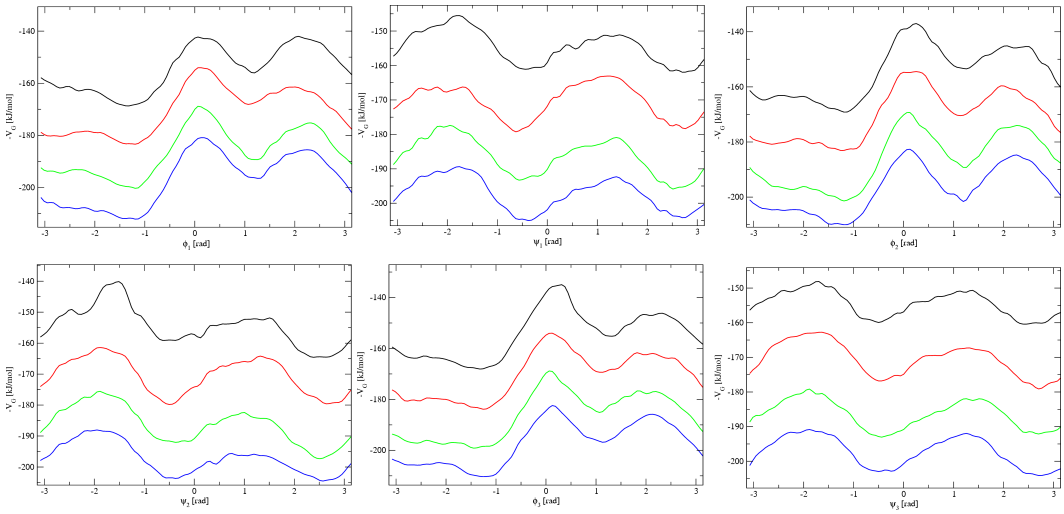

Figure 3.3: **Free energy profiles as a function of time for $Ala_3$ obtained with a 30 ns BE simulation.** $-V_G$ is reported for each backbone dihedral angle at several times after the filling time. Each time is represented with a different color: black (10 ns), red (11 ns), green (12 ns) and blue (13 ns). The parallel growth in time of the metadynamics bias potential is evident from the picture.

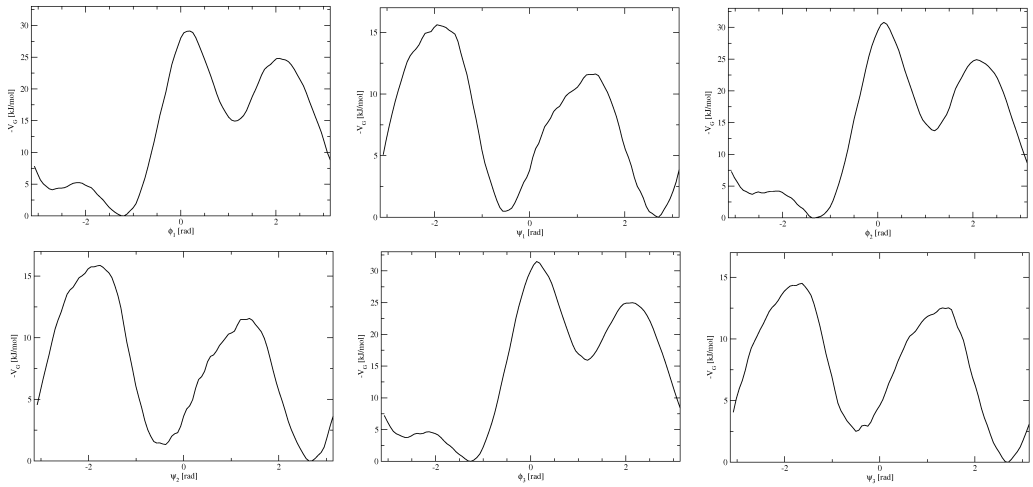

Figure 3.4: **Free energy profiles of $Ala_3$ along the six backbone dihedral angles for $Ala_3$.** The profiles are calculated using eq. 2.25 on the last 10 ns of a 30 ns BE simulation.

$g = 1$ ps). A Gaussian fit to these data (blue line) shows that this relative error has an average value of zero and is normally distributed, indicating that the deviations are unlikely to be systematic and are probably due to inaccurate sampling. If the analysis is repeated for a larger bin size the width of the relative error distribution becomes smaller. In fact, all the bins are visited more often and the free energies

are computed with better accuracy. As already underlined, in normal MD the error is small for low free energy states and large otherwise. In BE the error is instead much more uniform, and the free energy can be computed reliably also for several bins that are not even observed in MD. This property as we will show in the following is essential for constructing a reliable kinetic model of the system.

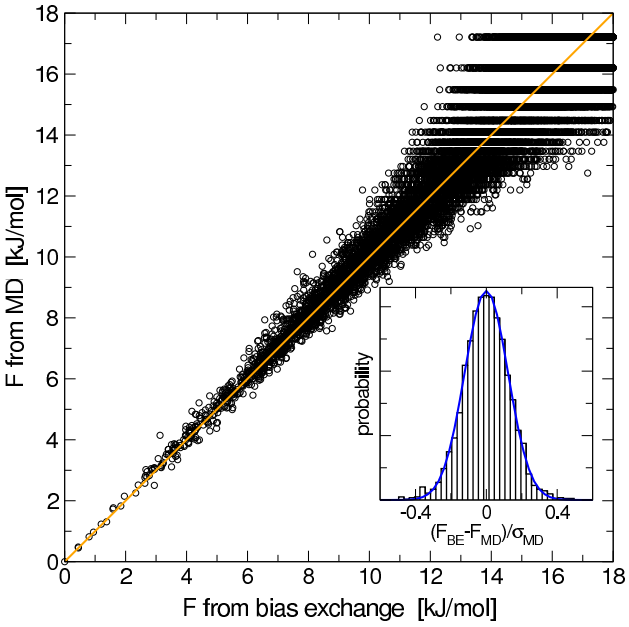


Figure 3.5: **Bins free energies of $Ala_3$ from BE and from MD.** Correlation between the bins free energies calculated using Eq. 2.29 applied on BE simulations data and using the standard thermodynamics relation $F_\alpha = -T \log n_\alpha$ on MD results. A bin size of 30° has been used. In the inset it is shown the distribution of the deviations between the bins free energies calculated from BE and from MD, divided by the estimated error on the MD free energy. A Gaussian fit of the distribution is also shown.

The equilibrium population of each of the $PP_{II}$, $\beta$, $\alpha_R$, and $\alpha_L$regions in the $(\phi_2, \psi_2)$ Ramachandran plot defined in section 3.1 was computed by summing the populations of the bins which are contained inside. The occupation probability calculated from MD and BE simulations is reported in Table 3.1: extended conformations ($PP_{II}$ and $\beta$) are the most populated, the helical $\alpha_R$ state is less populated while $\alpha_L$ has an occupancy lower than 0.1%, in agreement with available experimental data [101–103] and with previous simulations [98–100]. Once again (Table 3.1), the agreement between BE and MD results is very good for all the regions.

**Bin-based kinetic model.** A kinetic model of $Ala_3$ was built according to the procedure introduced in chapter 2. The free energies estimated from the BE simulations were used for constructing the kinetic model according to eq. 2.40. The diffusion matrix entering in eq. 2.41, was calculated by maximum likelihood for several choices of the time lag $\Delta t$ and bin size on MD simulations of length ranging from a few ns to 300 ns.

Table 3.1: **Equilibrium populations of the four main regions in the Ramachandran plot $(\phi_2, \psi_2)$ of $Ala_3$.**

| | $PP_{II}$ | $\beta$ | $\alpha_R$ | $\alpha_L$ |
|---|---|---|---|---|
| MD | 34.3% | 12.6% | 22.0% | 0.050% |
| BE | 32.1% | 12.0% | 22.3% | 0.085% |

The results from BE are compared to those from MD.

Table 3.2: **Diffusion matrix of $Ala_3$.** An MD trajectory of 60 ns is employed, using a time lag of 16 ps and a cubic side of 30°. The elements of the diffusion matrix are expressed in $rad^2/ps$. The position dependence of the matrix and the statistical uncertainty give a total estimated error of 20%.

| | $\phi_1$ | $\psi_1$ | $\phi_2$ | $\psi_2$ | $\phi_3$ | $\psi_3$ |
|---|---|---|---|---|---|---|
| $\phi_1$ | 0.040 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| $\psi_1$ | 0.000 | 0.037 | -0.018 | 0.000 | 0.000 | 0.000 |
| $\phi_2$ | 0.000 | -0.018 | -0.034 | 0.000 | 0.000 | 0.000 |
| $\psi_2$ | 0.000 | 0.000 | 0.000 | 0.034 | -0.014 | 0.000 |
| $\phi_3$ | 0.000 | 0.000 | 0.000 | -0.014 | 0.040 | 0.000 |
| $\psi_3$ | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.038 |

To estimate the accuracy of the kinetic model the mean first passage times (MFPT) for transitions among the four regions in $(\phi_2, \psi_2)$-space $PP_{II}$, $\beta$, $\alpha_R$, and $\alpha_L$ have been calculated both from MD and KMC. Moreover, the MFPT have been calculated also for transitions between the 8 bins corresponding to the 8 free energy minima obtained assigning the three $\psi$ dihedral angles in the $\alpha$ or in the $\beta$ region (see section 3.1 and Fig. 3.1). First, the kinetic model has been constructed for a bin size of 30° and optimizing a position independent $D$ with a time lag $\Delta t = 16$ ps (see table 3.2). The correlation plot between MD and KMC is shown in Fig. 3.6A, where only transitions observed at least 50 times in the MD trajectory are reported. The overall correlation is excellent except for transitions that display a large error bar in the MD simulation. The distribution of the first passage times for well visited transitions involving the central dihedral angles are also shown in Fig. 3.6 (panels B and C), both for MD and KMC. The agreement is excellent especially for the $\alpha_R \rightarrow PP_{II}$ transition, which occurs on a long time scale. All these results show that the rate model is able to reproduce accurately the kinetics of the real system. In order to quantify this accuracy it is useful to consider the slope $S$ of the line fitting the pairs $(\tau_i^{MD}, \tau_i^{KMC})$ of MFPT in Fig. 3.6A, where $i$ denotes a transition, as well as the RMS relative deviation

$$E = \sqrt{\frac{1}{N}\sum_i \left(\frac{\tau_i^{KMC} - S\ \tau_i^{MD}}{S\ \tau_i^{MD}}\right)^2}$$

where the sum runs over the $N$ transitions. $S$ and $E$, which should ideally have the values 1 and 0, have been computed for many different models in order to point out the critical issues that can affect the accuracy of the rate model:

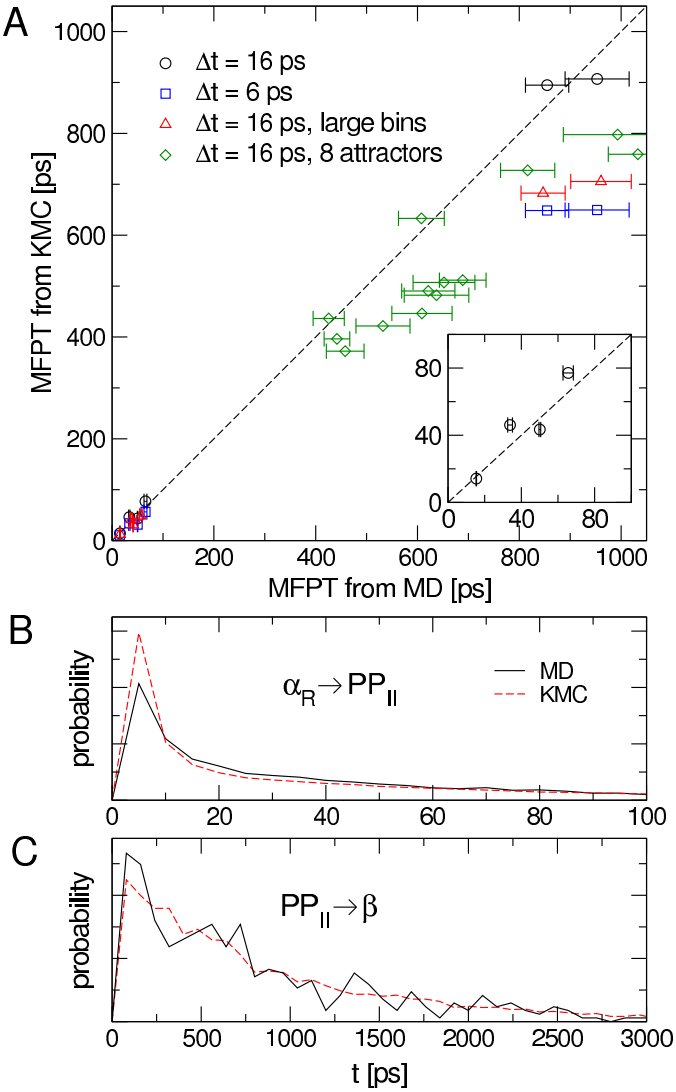


Figure 3.6: **Mean first passage times between the free energy basins of $\mathbf{Ala_3}$.** Panel A: correlation between the MFPT among the four regions in $(\phi_2, \psi_2)$-space $\mathrm{PP_{II}}$, $\beta$, $\alpha_R$, and $\alpha_L$, and among the eight attractors (see text and Fig. 3.1), obtained by MD simulations and by KMC using the kinetic model. The MFPT are calculated as the average time to go from one region to another, without passing through different regions. The error bars due to the statistical error in the MD simulations are also displayed. Large bins have a cubic side of 36°, while when not specified a cubic side of 30° is used. Panel B: distribution of FPTs from $\alpha_R$ to $\mathrm{PP_{II}}$ for MD and the kinetic model. Panel C: distribution of FPTs from $\mathrm{PP_{II}}$ to $\beta$ for MD and the kinetic model. For panel B and C a cubic side of 30° and a time lag of 16 ps was used for calculating the diffusion matrix $D$ (see table 3.2).

- **The time lag $\Delta t$ used to estimate $D$.** A position independent $D$ was optimized for different choices of time lag $\Delta t$ and MD trajectory length. The value of $S$ that is obtained for each $\Delta t$ is reported in Fig. 3.7. For $\Delta t = 16$ ps an error $E = 0.180$ and $S = 0.995$ is obtained, whereas for $\Delta t = 6$ ps $E = 0.188$ and $S = 0.721$, and for $\Delta t = 2$ ps $E = 0.185$ and $S = 0.357$.

This shows that the correct time scale is obtained if the time lag $\Delta t$ is large enough. For very small $\Delta t$ the MD trajectory cannot be approximated by a Markovian model [47].

- **The size of the bins**. Care must be taken in employing a bin size which is small enough to describe accurately the free energy of the system as a function of the CVs. Increasing the bin size from 30° to 36° still leads to reasonable transition times: the standard deviation and the slope become $E = 0.184$ and $S = 0.766$ for $\Delta t = 16$ ps (Fig. 3.6A). If the bin size is further increased to 40° the kinetic model compares badly with MD: $E = 0.682$ and $S = 0.152$. A position independent $D$ was optimized for each bin size using a 300 ns MD trajectory.

- **The length of the MD trajectory used to estimate $D$ by maximizing the likelihood**. The value of $S$ as a function of the length of the MD trajectory is reported in Fig. 3.7. A $\sim 50$ ns MD trajectory is necessary to obtain a $D$ which accurately reproduces the MFPT with $S \approx 1$. Increasing the length of the MD trajectory up to 300 ns does not change significantly $S$, whereas employing a shorter trajectory down to $\sim 10$ ns gives slightly larger errors. Thus changing the length of the MD trajectory between $10 - 300$ ns affects the time scale $S$ much less than the time lag $\Delta t$.

- **The position-dependence of $D$**. The MFPT was calculated using two different diffusion matrices obtained maximizing the likelihood only for the part of the MD trajectory that is close to two different attractors ($\alpha\alpha\alpha$) and ($\alpha\beta\alpha$), always using a time lag $\Delta t = 16$ ps. The difference in the slope $S$ is of the order of 10-20 %. This shows that the error that derives from neglecting the position dependence of $D$ is, at least for this system, smaller than the error due to the choice of the time lag $\Delta t$.

As a general comment, even in the worst cases investigated (short $\Delta t$, short MD trajectory), provided the bins size is not very large, the rate model produces MFPTs that are well correlated with the MD results, as shown by the relatively small value of $E$. The various approximations introduced in deriving the model affect only the proportionality factor, as quantified by $S$, that can be $\sim 0.5$ in the worst case (see Fig. 3.7). If the free energy of bins were estimated from MD and not from BE the correlation in the MFPT would be completely lost (data not shown). This is due to the fact that even in a quite extended MD simulation barriers are not well sampled; instead, in the BE simulation all the relevant bins are explored and the accuracy of the barriers between clusters is remarkably improved.

## 3.3 Discussion

The approach presented in chapter 2 exploits the trajectories of multiple metadynamics simulations for building a thermodynamic and kinetic model of complex

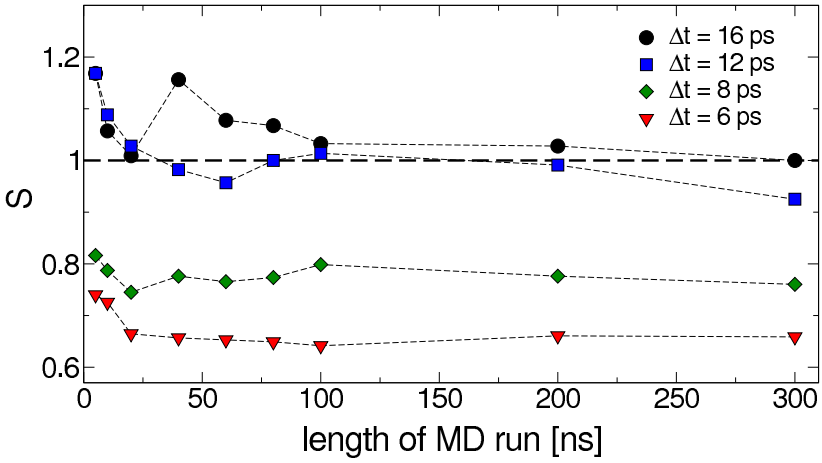


Figure 3.7: **Dependence of the diffusion coefficient of $Ala_3$ on the time lag and the trajectory length.** Dependence of the slope $S$ of the line fitting the pairs of mean first passage times $(\tau_i^{MD}, \tau_i^{KMC})$ (see text and Fig. 3.6A) from the parameters used in the fit of the diffusion matrix $D$: the length of the MD run and the time lag $\Delta t$. For $\Delta t \geq 12$ ps $S$ converges to the optimal value 1 (dashed line). A cubic side of 30° was used.

processes whose description requires a large number of collective variables. The aim of the model is to reproduce the long time scale dynamics of the system and to extract the metastable sets (clusters) of the kinetic process. The model is constructed as follows: in a first step the equilibrium probabilities of a finite set of conformational states, or bins, are determined by a weighted-histogram procedure exploiting the low-dimensional free energies estimated by metadynamics. In a second step an approximated description of the kinetics is obtained estimating the transition rates among the bins. The diffusion matrix entering in the model is estimated by a maximum-likelihood procedure [46] employing relatively short unbiased MD trajectories. The approach was tested on the Ace-$Ala_3$-Nme peptide in explicit solvent using the six backbone dihedral angles as CVs. For this system equilibrium MD trajectories on the microsecond timescale are sufficient to sample the relevant conformational space and were used as a reference to evaluate the accuracy of the kinetic model obtained from the BE results. The bins free energies obtained with the method presented in chapter 2 are in excellent agreement with free energies computed from equilibrium MD. The transition rates among neighboring bins are used to run a long KMC. The mean first passage times among selected states obtained in this way are in agreement with those extracted from the reference MD simulations.

CHAPTER

**FOUR**

# Kinetic model of Trp-cage folding

In this chapter we describe a first application of the approach introduced in chapter 2 and 3 to a realistic model of a protein.

A system that is almost ideal for theoretical investigation is the Trp-cage (TC5b)[104], a designed 20-residue miniprotein that folds rapidly [105] and spontaneously to a globular structure. The NMR structure (1L2Y)[104] reveals a compact hydrophobic core, in which the Trp side chain is buried. The secondary structure elements include a short $\alpha$-helix (residues 2-8), a $3_{10}$-helix (residues 11-14) and a polyproline II helix at the C-terminus. The folding mechanism of this system has been studied with several experimental techniques. Calorimetry, circular dichroism spectroscopy (CD) [106] and fluorescence [105] show a cooperative two-state folding behavior with transition midpoint at approximately 314 K and a relaxation time of 3.1 $\mu$s at 296 K[105]. UV-Resonance Raman [107] reveals a more complex unfolding behavior, with the presence of a compact intermediate that retains an $\alpha$-helical character and in which the hydrophobic core is even more compact. NMR experiments[104, 108] show a substantially cooperative thermal unfolding, but the large negative chemical shift deviations of Pro12-$\delta$3 and Gly11-$\alpha$3 suggest that those residues might pack more tightly as the temperature is raised. Also fluorescence correlation spectroscopy experiments cannot be interpreted in terms of a simple two-state folding and the formation of a molten-globule-like intermediate has been proposed [109].

By atomistic modeling the Trp-cage folding has been studied using several different approaches [110–120]. In particular, with an all-atom explicit-solvent description, the folding of Trp-cage has been studied by replica exchange molecular dynamics (REMD) [118, 121]. Starting from an extended configuration, a structure with a $C_\alpha$ root mean square deviation (RMSD) $< 2$ Å from the NMR reference structure is obtained after 100 ns of simulation on 40 replicas[121]. A relatively high melting temperature of 440 K is predicted. Other studies suggested that, even if Trp-cage is a rather small system, achieving statistical convergence in a REMD simulation may require much longer simulation times [34, 83]. The kinetics of Trp-cage folding was studied, in explicit solvent, by transition path sampling (TPS)

[83] and transition interface sampling (TIS) [122].

After the application of the approach described in chapter 2 to a simple system (see chapter 3) we here apply it for constructing a detailed kinetic and thermodynamic model of a complex process such as the Trp-cage folding.

A model is built that allows describing the folding process, computing the folding rates and the NMR spectra, simulating a T-jump experiment, etc. The scenario that emerges is in good agreement with the available experimental data. By kinetic Monte Carlo(KMC)[123, 124] and Markov cluster analysis(MCL)[125, 126] several metastable sets (clusters) are identified. These states, except for the folded cluster, can be considered misfolded intermediates of the folding process. At 298 K two main clusters are present, with a population of 58% and 25%, respectively. The most populated is the folded state and its structural properties are very close to the NMR ensemble. The second most populated cluster retains a significant amount of secondary structure, but has a $C_\alpha$ RMSD from the native state of approximately 4.4 Å. In this cluster, the Trp is trapped in a hydrophobic pocket and its distance from Pro12 and Gly11 is reduced. The presence of this cluster in the thermal ensemble of the system can explain some anomalies in the temperature behavior observed in NMR [104] and UV-Raman [107] experiments. The structures of the most populated misfolded intermediates are in good agreement with the unfolded states distances reported in Ref.[108]. Using the kinetic model a fluorescence T-jump experiment is also simulated. In agreement with the experimental results [105], a relaxation time of $2.3 \pm 0.7$ $\mu$s is found. This time is primarily determined by the relaxation towards the folded state of a compact molten globule-like structure, which acts as a kinetic trap. Relaxation times among all the other clusters, including transitions between fully unstructured states and the folded state, are all in the sub-microsecond time domain.

## 4.1 Computational setup

The simulations were performed with the GROMACS suite of programs [95, 96] and the AMBER03 force field [64], at a temperature of 298 K. The initial structure (pdb entry 1L2Y) [104] was solvated with 2075 TIP3P [65] water molecules in a $40 \times 40 \times 40$ Å water box. The system was simulated using BE[38]. Five collective variables (CVs) were biased according to the bias exchange scheme discussed in chapter 2[38].

- $CV1$ is the number of $C_\gamma$ contacts; $CV2$ is the number of $C_\alpha$ contacts; $CV3$ is the number of backbone h-bonds. They are defined as

$$CV1,2,3 = \sum_{ij} \frac{1-(r_{ij}/r_c)^8}{1-(r_{ij}/r_c)^{10}} \tag{4.1}$$

  where the sum runs over the appropriate set of atoms (all the $C_\gamma$ for $CV1$, all the $C_\alpha$ for $CV2$ and all the backbone H and O for $CV3$) and $r_c = 5$, 6.5

and 2 Å for $CV1$, $CV2$, and $CV3$ respectively.

- $CV4$ is the fraction of $\psi$ dihedrals belonging to the $\alpha$ region in the Ramachandran plot, defined as

$$CV4 = \sum_{i=1}^{N} \frac{1}{2} \left(1 + \cos\left(\psi_i - 45^o\right)\right) \tag{4.2}$$

  where the sum runs over all residues .

- $CV5$ is the correlation between successive $\psi$ dihedrals, defined as

$$CV5 = \sum_{i=1}^{N-1} \sqrt{1 + \cos^2\left(\psi_i - \psi_{i+1}\right)} \tag{4.3}$$

  where the sum runs over all residues .

All the variables are dimensionless and none of them requires the *a priori* knowledge of the folded state. The Gaussian widths chosen for $CV1$, $CV2$, $CV3$, $CV4$, $CV5$ were $\sigma_1 = 1.0$, $\sigma_2 = 2.0$, $\sigma_3 = 1.0$, $\sigma_4 = 0.4$, and $\sigma_5 = 0.4$, respectively. Simulations were performed with 8 walkers: one for each variable plus two walkers reconstructing a free energy surface in two dimensions: $CV3$-$CV4$ and $CV4$-$CV5$. The last walker, the "neutral walker", is not biased by any metadynamics potential, but is allowed to exchange conformations with the others. A Gaussian of height 0.1 kJ/mol was added every 1 ps to the bias potential for all the walkers except the neutral walker. The total length of the simulations was 50 ns. In Ref. [38] it was shown that the neutral walker statistics is approximately canonical, and all the averages were there computed using only its configurations, while the trajectories of the biased walkers were not used at all. The converged free energy profiles for each walker can be found in Ref. [38]. The MD simulations used for calculating the diffusion matrix and the NMR properties were run with the same computational setup of the BE simulation (except for specified changes in temperature).

## 4.2 Results

The results presented here were obtained analyzing, with the method introduced in chapter 2 the BE trajectory of Trp-cage from Ref. [38].

**Bin-based thermodynamic model.** The set of bins used for constructing the rate model was defined partitioning the five-dimensional CV space in small hypercubes according to the procedure outlined in chapter 2. A convenient choice of the cubic sides was found to be $ds_i = 2\sigma_i$, where $\sigma_i$ is the width of the Gaussian used for CV $i$. With this choice, the number of bins that are explored at least twice is $\sim 10000$. To check the consistency of the model other cubic sides were also attempted. We checked that the CVs we are using do not lump together different conformations: indeed, the $C_\alpha$ RMSD from the bin reference structure is less then 2.5 Å for most of the low free energy bins. We also verified that if a compact

secondary structure element is present in the reference structure of a bin, the same structure element will be present in the overwhelming majority of frames assigned to that bin: high RMSD values are primarily determined by flexible regions that undergo fast rearrangement on the ns time scale.

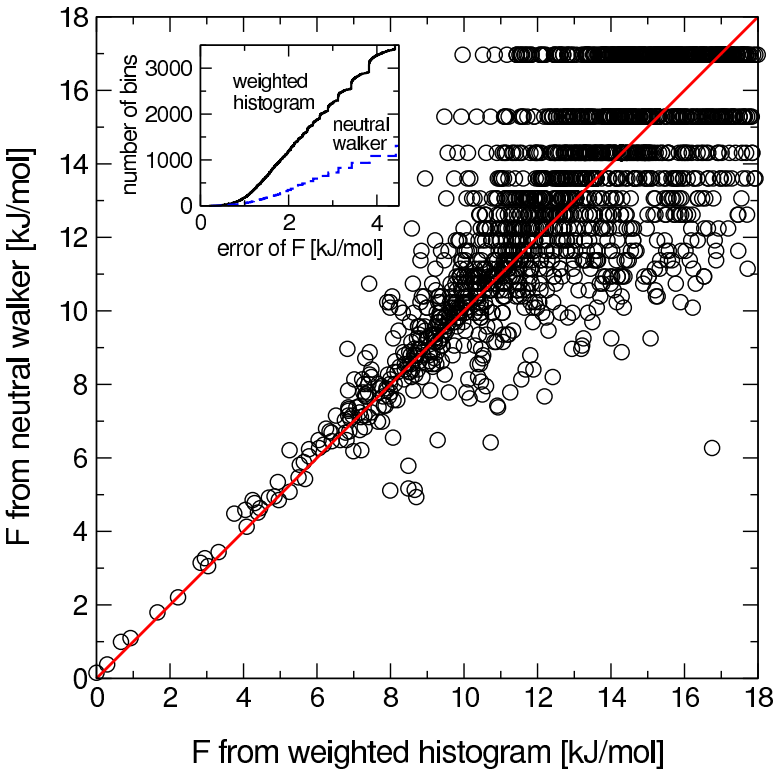


Figure 4.1: **Correlation between free energies of neutral walker and WHAM for Trp-cage.** Correlation between the bins free energy evaluated using the approach described in chapter 2 and using the neutral walker ensemble at T=298 K. Inset: cumulative number of bins with an error smaller than the value reported in abscissas. The error is estimated using eq 2.31. The value of g entering this equation is estimated from the correlation time of the bin occupancies and is equal to 10 ps.

The free energies of the bins were estimated using Eq. 2.29, evaluating the biasing potentials on each of the eight replicas by Eq. 2.25 with $t_F$ =22 ns. In order to reduce the error induced by the time-dependent fluctuations, the bias potential was averaged independently in the two halves of the interval $[22ns, 50ns]$ (see chapter 2). Only configurations collected after 22 ns in which the two averaged potentials are consistent within $2T$ are retained for further analysis. Unlike for the $Ala_3$ system in the case of the Trp-cage an extended ergodic MD simulation is not available, as equilibrating the system would require performing a run of several

Table 4.1: **Diffusion matrix of Trp-cage close to cluster 1 and cluster 5**. A MD trajectory of 80 ns starting from the folded state and remaining close to it was employed for calculating the diffusion matrix of cluster 1, using a time lag of 12 ns. A MD trajectory of 65 ns exploring cluster 5 was employed for calculating the diffusion matrix of cluster 5, using a time lag of 12 ns. Units are 1/ns. The statistical uncertainty on each element has been calculated by dividing the each trajectory in 3 pieces and evaluating the diffusion matrix independently in each part. The simulated T-jump experiment (see section 4.2.3) performed using these diffusion matrix gives a relaxation time of 2322 ns (cluster 1) and 2148 ns (cluster 5) respectively.

| cluster 1 | $CV1$ | $CV2$ | $CV3$ | $CV4$ | $CV5$ |
|---|---|---|---|---|---|
| $CV1$ | $0.867_{\pm 0.180}$ | $-0.060_{\pm 0.001}$ | $-0.026_{\pm 0.002}$ | $0.0009_{\pm 0.0001}$ | $-0.009_{\pm 0.0003}$ |
| $CV2$ | $-0.060_{\pm 0.001}$ | $3.027_{\pm 0.474}$ | $0.660_{\pm 0.155}$ | $0.010_{\pm 0.002}$ | $-0.020_{\pm 0.004}$ |
| $CV3$ | $-0.026_{\pm 0.002}$ | $0.660_{\pm 0.155}$ | $0.343_{\pm 0.045}$ | $0.00040_{\pm 0.00003}$ | $0.002_{\pm 0.004}$ |
| $CV4$ | $0.0009_{\pm 0.0001}$ | $0.010_{\pm 0.002}$ | $0.00040_{\pm 0.00003}$ | $0.073_{\pm 0.003}$ | $-0.005_{\pm 0.002}$ |
| $CV5$ | $-0.0090_{\pm 0.0003}$ | $-0.020_{\pm 0.004}$ | $0.002_{\pm 0.004}$ | $-0.005_{\pm 0.002}$ | $0.028_{\pm 0.003}$ |
| cluster 5 | | | | | |
| $CV1$ | $0.362_{\pm 0.007}$ | $0.156_{\pm 0.002}$ | $0.004_{\pm 0.00006}$ | $0.004_{\pm 0.00005}$ | $-0.011_{\pm 0.0001}$ |
| $CV2$ | $0.156_{\pm 0.002}$ | $2.875_{\pm 0.040}$ | $0.541_{\pm 0.010}$ | $0.0180_{\pm 0.0002}$ | $-0.023_{\pm 0.0003}$ |
| $CV3$ | $0.004_{\pm 0.00006}$ | $0.541_{\pm 0.010}$ | $0.203_{\pm 0.004}$ | $-0.000140_{\pm 0.000002}$ | $-0.0130_{\pm 0.0005}$ |
| $CV4$ | $0.004_{\pm 0.00005}$ | $0.0180_{\pm 0.0002}$ | $-0.000140_{\pm 0.000002}$ | $0.0270_{\pm 0.0003}$ | $0.00151_{\pm 0.00002}$ |
| $CV5$ | $-0.011_{\pm 0.0001}$ | $-0.023_{\pm 0.0003}$ | $-0.0130_{\pm 0.0005}$ | $0.00151_{\pm 0.00002}$ | $0.040_{\pm 0.002}$ |

tens of $\mu s$. Thus, for Trp-cage it is not possible to compare the equilibrium bins free energies with the ones obtained using BE. Instead the free energies estimated with the WHAM-like[13] procedure are compared with the ones obtained using the neutral walker statistics as described in Ref. [38].

The correlation between the two free energies is excellent, especially for bins with low free energy (see also Fig. 4.1). As shown in Ref. [38], the neutral walker reliably reproduces the ensemble generated with normal replica exchange. This shows that the three methods, replica exchange, the neutral walker method and the weighted histogram approach described in the chapter 2, all give consistent results for the statistics of the most populated bins. The errors on the free energies computed using the neutral walker ensemble are large for bins whose occupancy is low and bins of high free energy are sometimes not explored at all. The number of bins whose error is below 4 kJ/mol is approximately 1000 and 3000 for the neutral walker and the weighted histogram procedure, respectively (see also Fig. 4.1, inset). The weighted histogram free energies are systematically very reliable up to $\sim$ 25 kJ/mol. It is worth to note that most of the low free energy bins are visited independently by several walkers (e.g. the lowest free energy bin is visited by all the walkers).

**Bin-based kinetic model.** Like for the $Ala_3$ case, the free energies of the bins were used for estimating the rate for the transitions between all the neighbouring bins according to Eq. 2.40.

Table 4.2: **Diffusion matrix of Trp-cage**. five MD trajectories for a cumulative time of 500 ns are employed, using a time lag of 12 ns. Units are 1/ns.

| | $CV1$ | $CV2$ | $CV3$ | $CV4$ | $CV5$ |
|---|---|---|---|---|---|
| $CV1$ | 0.445 | 0.263 | 0.010 | 0.012 | -0.010 |
| $CV2$ | 0.263 | 2.725 | 0.530 | 0.034 | -0.025 |
| $CV3$ | 0.010 | 0.530 | 0.300 | -0.005 | -0.015 |
| $CV4$ | 0.012 | 0.034 | -0.005 | 0.037 | -0.003 |
| $CV5$ | -0.010 | -0.025 | -0.015 | -0.003 | 0.040 |

The diffusion matrix entering in eq. 2.41 was evaluated using the maximum likelihood approach described in chapter 2 on five MD trajectories for a total time of $\sim$ 500 ns. In order to estimate the variation of $D$ with the protein conformations, the MD trajectories were initiated from structures belonging respectively to the folded state, and clusters 2, 3, 4 and 5 (see below for the definition of the clusters). Optimizing $D$ separately in each cluster leads to a cluster-dependent diffusion matrix. The diffusion matrix of clusters 1 and 2 are shown in table 4.1, all the are can be found in ref [48] (Text S1). However, these variations influence the relevant observables only mildly. Indeed, the folding relaxation times (see section 4.2.3) computed with a cluster-dependent D or with a constant D (calculated using all the MD trajectories at once) are consistent within a standard deviation of $\pm$500 ns (see tables 4.2, 4.1).

This uncertainty is comparable to the one deriving from the error on the bins free energy (see section 4.2.3). The error bars reported for each element of the diffusion matrices indicate that they are well converged with the simulation length. As the uncertainty induced by using different $D$ is small, all the analysis below is performed employing a position independent $D$ obtained by likelihood optimization using all the trajectories at once (see table 4.2).

The maximum likelihood analysis has been repeated sampling the MD trajectory at several different time lags $\Delta t$. Due to important memory effects $D$ becomes approximately independent on the time lag only for $\Delta t > 10 - 12$ ns. The diffusion matrix obtained with $\Delta t = 12$ ns was used for constructing the kinetic model. As a consequence, the rate model is by construction unable to reproduce the kinetics of transitions that occur on a time scale shorter than $\sim$10-20 ns. The value of few elements of the diffusion matrix as a function of the time lag is reported in Fig. 4.2.

### 4.2.1 Metastable sets (clusters) of the Trp-cage rate model

The rate model described in chapter 2 has the form of a generalized rate equation with the rates given by Eq. 2.40. The presence of metastable sets ("clusters") was detected applying the MCL[125, 126] method to the Trp-cage kinetic model. The algorithm requires choosing a parameter $p$ that tunes the granularity of the

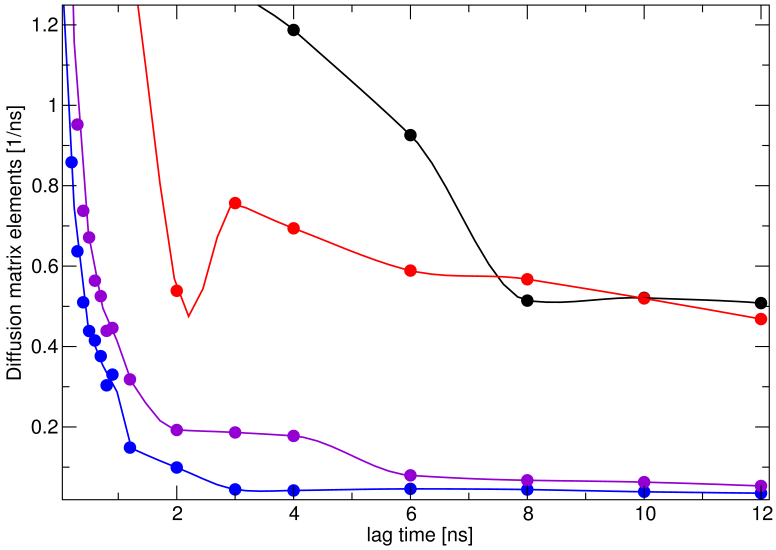


Figure 4.2: **Diffusion matrix of Trp-cage as a function of the time lag.** Few elements of the diffusion matrix are reported. A MD trajectory of $\sim$ 500 ns and the maximum likelihood approach explained in chapter 2 is used for calculating **D** at each time lag. After approximately 8-10 ns the diffusion matrix elements show a converging behaviour.

description: for $p = 1$ only one cluster is detected, while for large $p$ all the bins are assigned to different clusters. Several choices of the $p$ parameter are attempted (in Ref. [125, 126] the value $p = 1.2$ is considered). At 298 K, for $p = 1.13$ only two relevant clusters are found, one with an occupancy of $\approx$ 90% and one of $\approx$ 5%. The RMSD among the structures belonging to the big cluster is very large, indicating that, for this system, $p = 1.13$ is not appropriate. For $p = 1.14$ the large cluster splits in two clusters with populations of $\approx$ 12% and $\approx$ 77%. Still the larger cluster includes qualitatively different structures. At $p = 1.15$ the larger cluster splits further in three, while the other clusters remain approximately unchanged. Increasing further $p$ up to 1.17 does not modify significantly the three most populated clusters, whereas for $p = 1.2$ the system is fragmented in more than 10 clusters. At $p = 1.15$, only 5 significantly populated ($> 1$ %) clusters are found, the two larger ones having a population of $\approx$ 58% and $\approx$ 25% respectively (Table 4.3). The average $C_\alpha$ RMSD between the clusters structures and the NMR ensemble is $\approx$ 1.8 Å for cluster 1 and $>$ 4.4 Å for cluster 2 and the other clusters. Moreover, all the bins with $C_\alpha$ RMSD $<$ 2 Å belong to cluster 1. This allows concluding that MCL analysis using $p = 1.15$ is able to identify a folded cluster with structural properties similar to the NMR ensemble. Its occupancy is of 58 % at 298 K. Remarkably, at this temperature it exists another cluster with non-negligible population (25%) that contains structures that are different from the structural ensemble generated from the NMR data ($C_\alpha$ RMSD = 4.4 Å). In the next section the consequences of the existence of this second cluster in the thermal ensemble at 300 K are discussed. It is worth to note that in the MD simulations used for the calculation of $D$, if the trajectory starts from a structure belonging to

a cluster, it remains there for most of the simulation (few tens of ns). This means that MD simulations are consistent with the description of metastable states given by the MCL algorithm. In Fig. 4.3A, the most populated clusters obtained for $p = 1.15$ are shown using a projection on three variables, the $C_\alpha$ contacts, the $\alpha$-helix fraction, and the correlations between consecutive dihedrals. Each color corresponds to a different cluster, and the lowest free energy bin (attractor) of each cluster is depicted as a sphere of the same color.

The properties of the clusters depicted in Fig.4.3A are summarized in Table 4.3.

Table 4.3: **Selected properties of the Trp-cage clusters represented in Figure 4.3A, at 300 K**. Enthalpies and entropies are expressed with respect to the folded cluster value. The occupancy of each cluster $B$ has been calculated as $P_B = \sum_{\alpha \in B} e^{-F_\alpha/T} / \sum_\alpha e^{-F_\alpha/T}$ where the summation at the numerator is extended to all bins $\alpha$ belonging to the cluster $B$. The observables reported in the table are evaluated using Eq. 2.32, where the summation is extended only to the bins $\alpha$ that belong to a specific cluster. The RMSD is computed as the average RMSD between the cluster structures and all the structures in 1L2Y PDB entry. The number of helical residues has been computed according to Ref. [127] using the program g_helix in the GROMACS distribution.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| % occupancy | $58.3 \pm 0.8$ | $24.6 \pm 0.7$ | $7.0 \pm 0.3$ | $1.2 \pm 0.1$ | $2.8 \pm 0.2$ |
| $\Delta H$ (kJ/mol) | $0.0 \pm 1.9$ | $5.0 \pm 2.6$ | $11.7 \pm 3.8$ | $13.8 \pm 5.3$ | $38.2 \pm 5.3$ |
| $T\Delta S$ (kJ/mol) | $0.0 \pm 1.9$ | $2.9 \pm 2.6$ | $6.5 \pm 3.8$ | $4.1 \pm 5.3$ | $30.7 \pm 5.3$ |
| $C_\alpha$ RMSD (Å) | $1.82 \pm 0.05$ | $4.44 \pm 0.03$ | $6.76 \pm 0.04$ | $5.54 \pm 0.06$ | $6.08 \pm 0.05$ |
| Trp SASA (Å$^2$) | $47.1 \pm 0.6$ | $70.5 \pm 1.0$ | $126.4 \pm 0.7$ | $116.7 \pm 1.0$ | $140.4 \pm 0.8$ |
| Helical residues | $5.31 \pm 0.02$ | $2.91 \pm 0.03$ | $3.86 \pm 0.04$ | $0.66 \pm 0.03$ | $1.70 \pm 0.03$ |

In Fig. 4.5, the hydrophobic contacts and the hydrogen bonds with the Trp6 are shown schematically for each attractor. Selected proton distances are also displayed for the three most populated clusters. A good agreement with the NMR unfolded state distances reported in Ref.[108] is found. Cluster 1, as already anticipated, resembles very closely the NMR structure. More details will be provided in the following section. Cluster 2 has a $C_\alpha$ RMSD of $\sim 4.4$ Å with respect to the NMR structure, but it retains at least part of the native $\alpha$-helix. The Trp SASA in this cluster is $70.5 \pm 1$ Å$^2$, which compares with the value of $47.1 \pm 0.6$ Å$^2$ observed in the folded cluster. This indicates that Trp is shielded from the solvent also in cluster 2. Arg16 forms a $\pi$-stacking with Tyr3 (see Fig.4.3A) while Trp6 is in contact with Pro12, Pro18, Gly11 and the aliphatic chain of Arg16 (see Fig. 4.5). As outlined in Fig. 4.5, except for the Arg16 H$\beta$2-Trp6 H$\eta$2 distance, the cluster 2 attractor(reference structure) shows Pro12 H$\gamma$2-Trp6 H$\eta$2 and Arg16 H$\beta$3-Trp6 H$\eta$2 distances shorter than those in the folded cluster. The nearest hyperpolarized[108] Trp6 proton can be different in each cluster (e.g. in cluster 1 the Arg16 H$\beta$2-Trp6 H$\epsilon$1 distance is shorter than Arg16 H$\beta$2-Trp6 H$\eta$2). These distances are in very

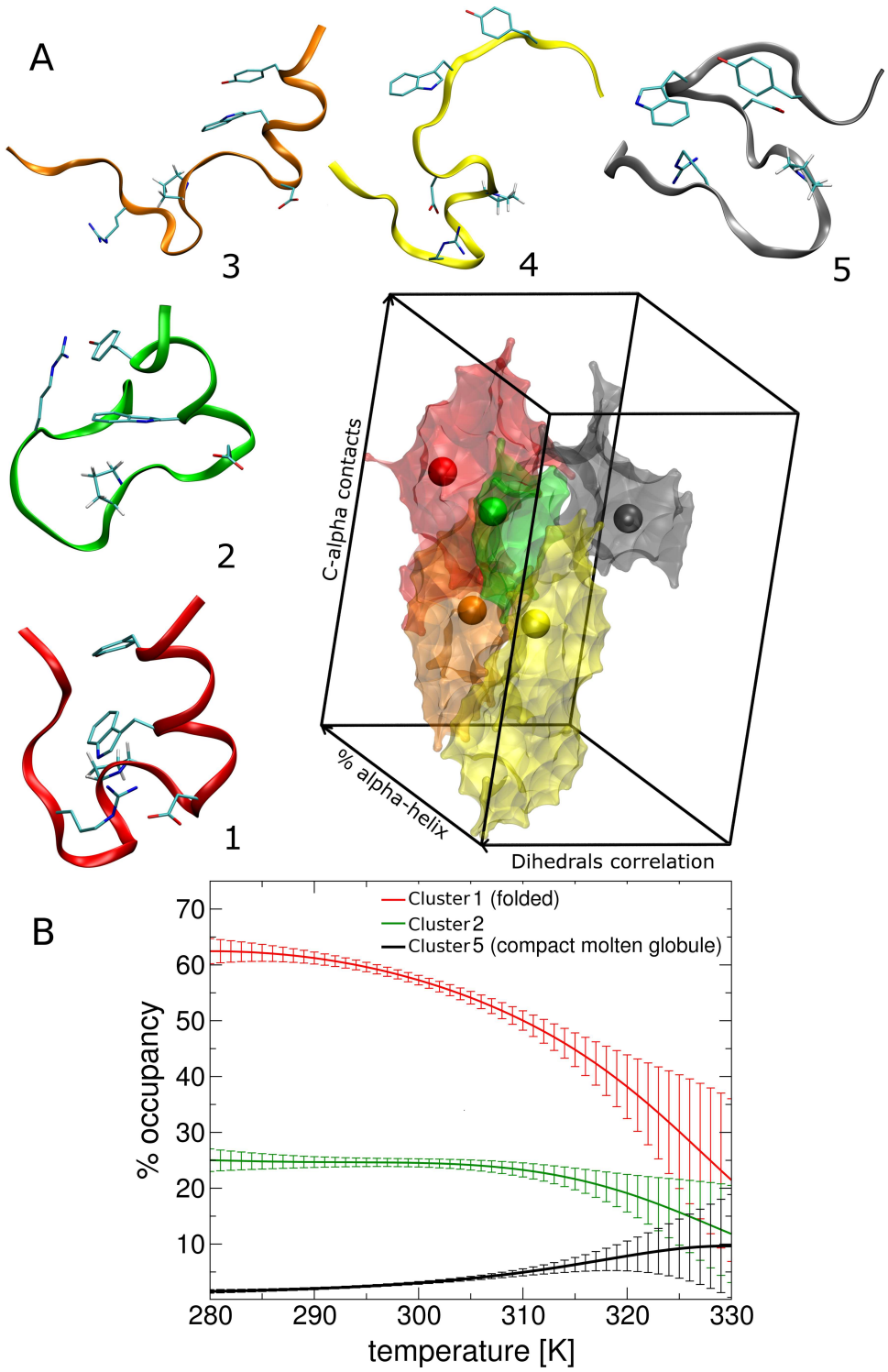


Figure 4.3: **Metastable kinetic clusters of Trp-cage.** Panel A: metastable sets (clusters) detected by MCL method using $p = 1.15$. The colored spheres correspond to the lowest free energy bins of each cluster. The corresponding structures are shown with the same color code. Panel B: occupancy as a function of temperature of cluster 1, 2, and 5.

good agreement with those found in the NMR experiments[108] for the unfolded state. This cluster resembles the intermediate observed in a 100 ns implicit solvent simulation (Ref. [111]. Cluster 3 (orange) still contains a short $\alpha$-helix. The $C_\alpha$

contacts are reduced with respect to the folded cluster and the Trp is partially solvent exposed. The reference structure of cluster 3 is similar to the state I of Ref. [83] and to the intermediate structure found in Ref. [118], with the difference that the Asp9-Arg16 salt bridge in cluster 3 is formed only in a fraction of the bins belonging to the cluster. This may indicate that the salt bridge is rather unstable. The Leu7 H$\delta$2-Trp6 H$\epsilon$3 distance in the cluster 3 attractor is shorter than that in the folded state. Also in this case the distance compare well with the NMR experiments value[108]. This imply that the presence of cluster 2 and cluster 3 (the two most populated misfolded clusters) is consistent with the unfolded state ensemble information reported in Ref.[108]. The other clusters show only a small residual secondary content and can be generically referred to as "unfolded states". The attractor of cluster 4 is stabilized by the formation of the Asp9-Arg16 salt bridge. The bins belonging to cluster 5 are mostly compact molten globule structures characterized by the presence of several hydrophobic and $C_\alpha$ contacts (even more than in the native state) but small secondary content (see Fig. 4.3A and Fig. 4.4).

In the most stable bin of this cluster Trp6 is in contact with Pro17 and Pro18 residues (see Fig. 4.5). In Fig. 4.3B the occupancies of cluster 1, 2, and 5 are plotted as a function of temperature. As expected the folded cluster (cluster 1) increases its occupancy as the temperature decreases. Its population is 50 % at 310 K, a temperature that is consistent with the experimental melting point of 317 K[106, 107]. The error on the occupancies becomes large at $T > 325$ K, indicating that the temperature extrapolation based on Eq. 2.33 is unreliable after this temperature. The occupancy of cluster 5 is almost negligible at 300 K (2.8 %), but it grows significantly with temperature(see Fig. 4.3B). The importance of this will become clear when the kinetic properties of the system will be discussed. The helical content decreases only slowly with temperature, consistently with REMD results in explicit solvent[121]. On the average, only $\sim 1$ $\alpha$-helical residue melts between 290 and 320 K.

### 4.2.2 NMR Properties of Trp-cage

In order to characterize in more detail the nature of the clusters described in the previous section, it is useful to consider their NMR properties. This has been done using semiempirical chemical shift calculations for each protein configuration explored by the BE simulation. This kind of calculations are very powerful as they enable the accurate prediction of chemical shift for proteins and can be used in combination with atomistic simulations for protein structure determination [129]. In this work the protons chemical shift deviations (CSD) and ring current shifts (RCS) of a specific configuration were estimated using the SHIFTS program[130] version 4.1. As only cluster 1 and 2 are compact and show a significant content of secondary structure, the investigation is here restricted mainly to these two clusters. The CSD and RCS calculated for the full ensemble of bins (or for a specific cluster), were evaluated first averaging in each bin and then averaging the result using Eq.

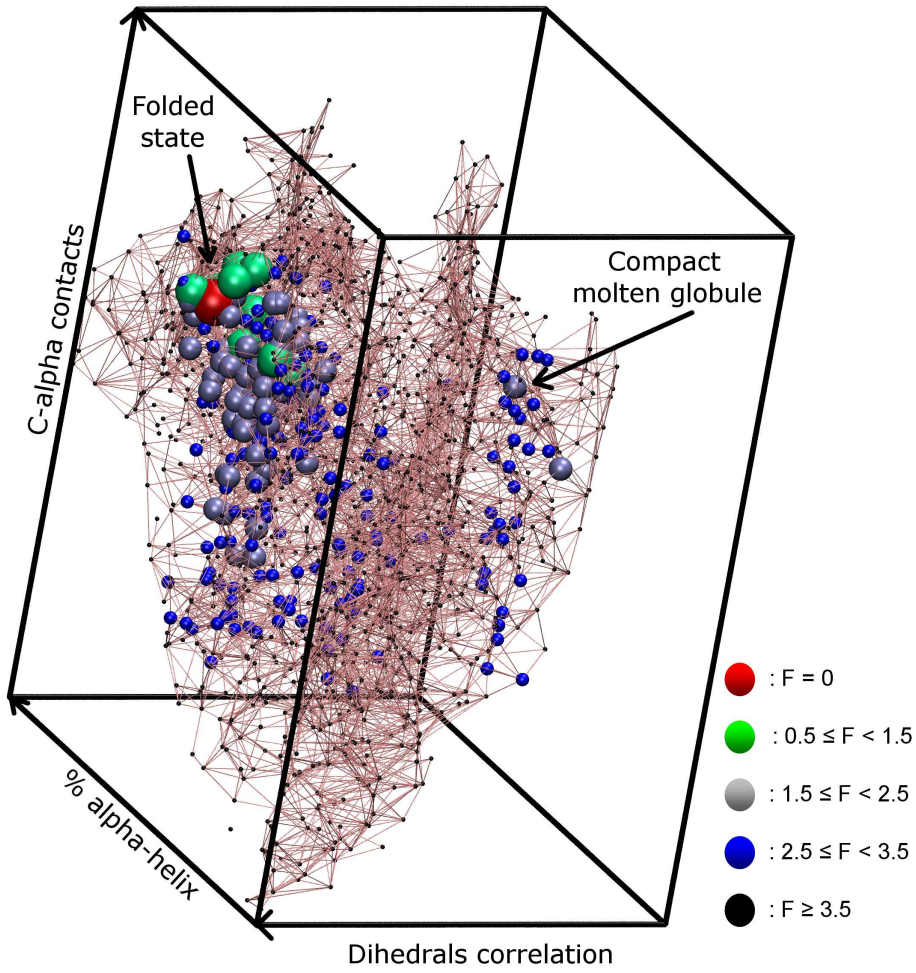


Figure 4.4: **Bins network topology at T=298 K projected on three dimensions: $C_\alpha$ contacts, dihedral correlations and $\alpha$-helix fraction.** Each bin is represented as a sphere whose dimension and color is associated with the free energy (kcal/mol). The location of the folded state and the molten globule (cluster 5) lowest free energy bins are indicated in the figure.

2.32 for all the bins (for all the bins belonging to a specific cluster, see above). The RCS temperature derivatives were calculated by finite difference in the temperature interval $298-303$ K. A 20 ns MD simulation starting from the NMR structure[104] at 282 K was also used for calculating NMR properties. The variation of the $\alpha$ protons RCS with the temperature was calculated by applying Eq. 2.32 and 2.33. In Fig. 4.6A the $\alpha$ protons CSDs of cluster 1 are compared with the experimental results (full circles). The correlation between theoretical and experimental NMR CSDs is rather good ($R^2 = 0.96$), while cluster 2 shows a much smaller correlation with experiments, especially for protons that have negative CSDs. The correlation with NMR data is even smaller for all the other clusters. This confirms that the cluster classification deriving from Markov cluster analysis accurately discriminates between the folded state (cluster 1), an unfolded state with several native-like features (cluster 2), and all the rest. The correlation with experiments is retained using in the average the full ensemble of bin ($R^2 = 0.95$).

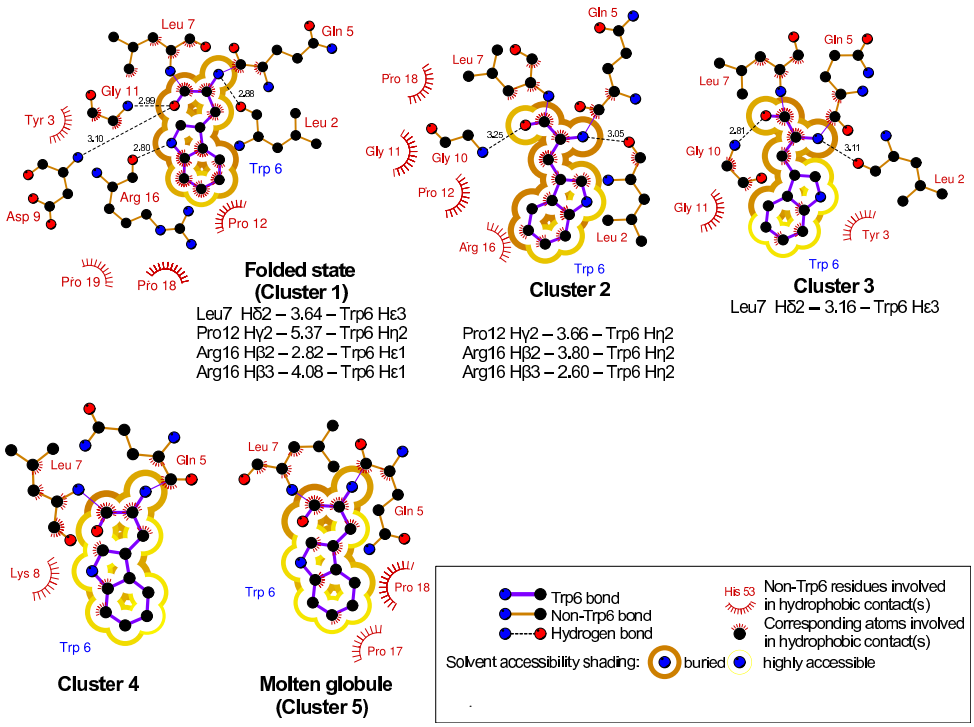


Figure 4.5: **Trp6 interactions in the clusters reference structures of Trp-cage.** Hydrophobic contacts within 3.9 Å and hydrogen bonds(Å) are displayed. The distances(Å) between Leu7, Pro12, Arg16 and Trp6 selected protons are shown for the 3 most populated clusters. The corresponding values can be compared with the unfolded state NOE contact distances reported in Ref.[108]. The nearest hyperpolarized Trp6 protons in the NMR experiment are selected for measuring distances. Short Ile4-Trp6 proton distances[108] (4-5 Å) are not reported in the figure since they are found mostly in open random-coil like structures and in some more compact cluster with population $< 1\%$. This figure was generated using the program LIGPLOT [128]

Even if correlation is good, it has to be noted that the proportionality factor between theoretical and experimental CSDs is 0.46 in the full ensemble of bins and 0.6 in cluster 1. To investigate the origin of the variations in the proportionality factor two 20 ns equilibrium MD simulations have been performed, at 282 K (experimental temperature) and at 300 K, starting from the NMR structure and with the same computational setup used in the BE simulation. At both temperatures the proportionality factor with experimental CSDs is 0.8 instead of 1, therefore 0.8 has to be considered the reference value for our computational setup. The optimal proportionality factor of 0.8 is obtained if the CSDs are computed on the lowest free energy bin of cluster 1. The slope difference between 0.6 (cluster 1) and 0.8 may be ascribed to small inconsistencies between the ensemble of structures generated with BE and by an unbiased MD starting from the NMR structure. The further slope variation when the calculation is extended to the full ensemble of bins is most likely a consequence of calculating NMR properties at 298 K instead of at the experimental temperature of 282 K where the population of cluster 1 is larger.

Using a similar procedure (see chapter 2) RCS and its temperature derivative

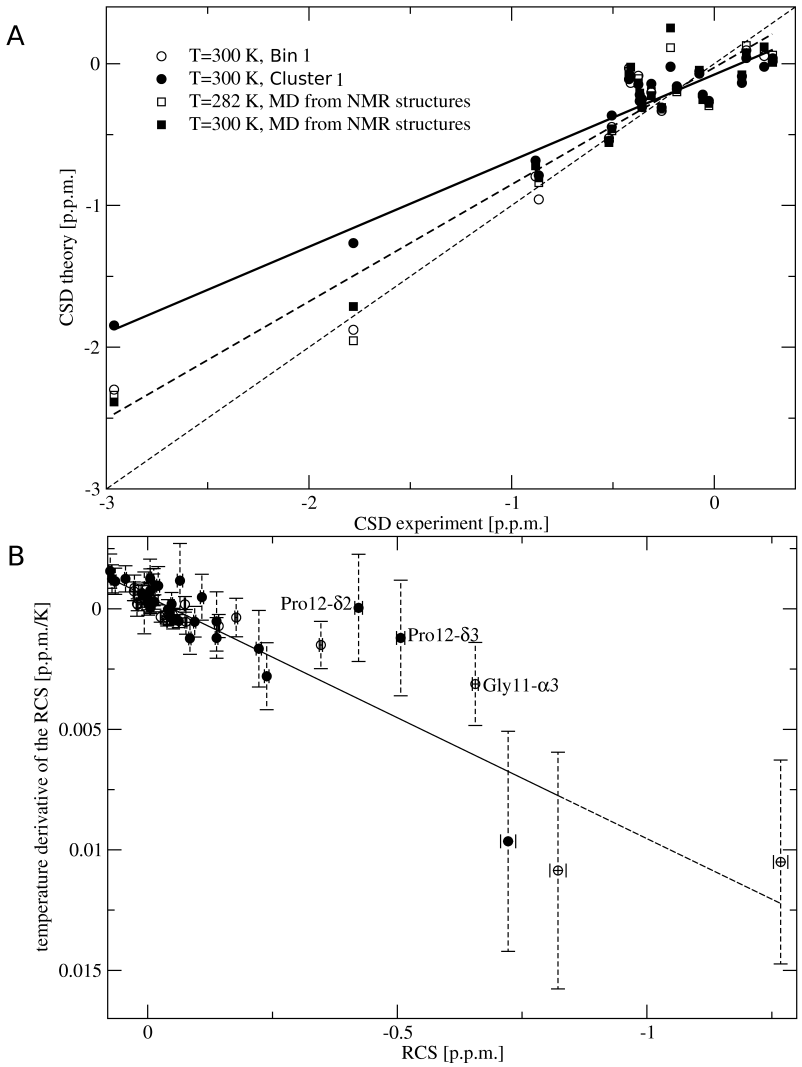


Figure 4.6: **Simulated NMR chemical shift deviations and ring current shifts in Trp-cage.** Panel A: correlation between experimental and calculated $\alpha$ protons CSD for the cluster 1 (black circles), the lowest free energy bin (empty circles), and the ensemble obtained from a simulation started from the NMR structure at 282 K (black squares) and 300 K (empty squares). The continuous and dashed lines are obtained from a linear regression on the black circles and the squares, respectively. The thin dashed line corresponds to a proportionality factor of 1 between experiment and theory. Panel B: correlation between protons ring current shift temperature derivative and the corresponding ring current shift value evaluated at 298 K. Results are shown for $\alpha$ protons (empty circles) and side chain protons (black circles). Ring current shift temperature derivative is calculated as a finite difference between 298 and 303 K using the chemical shift temperature extrapolation obtained using Eq. 2.32 and 2.33.

were also computed. It is worth to note that most of the large CSD are due to the Trp RCS[104]. The protons whose RCS is large are also those whose RCS depends more strongly on $T$, in excellent agreement with the experimental data[104]. The $\alpha$ protons RCS temperature derivatives as a function of the RCS are plotted in Fig. 4.6B. The results are plotted as a function of the RCS estimated at 298 K. The comparison is performed at 298 K and not at the experimental temperature

of 282 K in order to avoid error propagation that is unavoidable if Eq. 2.33 is used for extrapolating the results for a large temperature difference. Despite of this, the two observables correlate linearly ($R^2 = 0.94$ for the $\alpha$-protons), consistently with experiments[104]. Side chain protons in the C-terminal part of the protein fall on the same correlation line, also in agreement with the experiments[104]. A few protons deviate significantly from this linear behavior. The most significant deviation are observed for Pro12-$\delta$2, Pro12-$\delta$3, and Gly11-$\alpha$3, the last two being also reported experimentally[104]. The RCS of Pro12-$\delta$3 and Pro12-$\delta$2 is large, while their RCS derivative is almost zero. The cluster decomposition proposed here can be used to elucidate the presence of these outliers. In fact, the RCS of Pro12-$\delta$3 is $-0.53 \pm 0.01$ p.p.m. and $-0.97 \pm 0.02$ p.p.m in cluster 1 and 2 respectively, while other protons (except Pro12-$\delta$2 and Gly11-$\alpha$3) have RCS which are less negative in cluster 2 than in cluster 1 or similar in the two clusters. The RCS of Gly11-$\alpha$3 has a similar value in both clusters. This significant difference derives from the fact that Pro12-$\delta$3 and Pro12-$\delta$2 in cluster 2 are much closer to Trp than in cluster 1. Since, increasing the temperature, the relative population of cluster 2 and 1 changes (see Fig. 4.3B), the RCS of Pro12-$\delta$2, Pro12-$\delta$3 and Gly11-$\alpha$3 changes with temperature less than the RCS of other protons. In view of these results, the anomalous behavior of Pro12-$\delta$3 and Gly11-$\alpha$3 observed experimentally can be considered a signature of the presence of cluster 2 in the thermal ensemble of Trp-cage.

### 4.2.3 Dynamical properties: simulated Trp SASA T-jump experiment

The fluorescence relaxation after a temperature jump (T-jump) was used in Ref. [105] to infer information on the Trp cage folding kinetics. The fluorescence properties of the system are here estimated by computing the Trp solvent accessible surface area (SASA), which is known to correlate with fluorescence[131]. The Trp SASA was calculated for each bin averaging over all the configurations belonging to a bin using the program g_sas in the GROMACS distribution[132]. The Trp SASA relaxation after a temperature jump (T-jump) was estimated using the rate model. The T-jump experiment was mimicked generating 1,000,000 initial bins from an equilibrium distribution at 291 K. The bins free energies at 291 K used for generating the distribution were evaluated applying Eq. 2.33. Starting from each initial bin a KMC[123, 124] trajectory of 100 $\mu$s was run at 298 K. The Trp SASA was then calculated as a function of time averaging over this ensemble. The influence that the error on the free energies and on the enthalpies has on the results has been checked generating several kinetic models in which $F_\alpha$ and $H_\alpha$ were defined adding to the original values a random number drawn from a Gaussian distribution with standard deviation given by the error interval. A simulated Trp SASA T-jump experiment was repeated for each model. The error on the relaxation time was estimated from the standard deviation of the measures on the different models.

The result shows a smooth decay to an asymptotic value on the time scale of the microseconds.

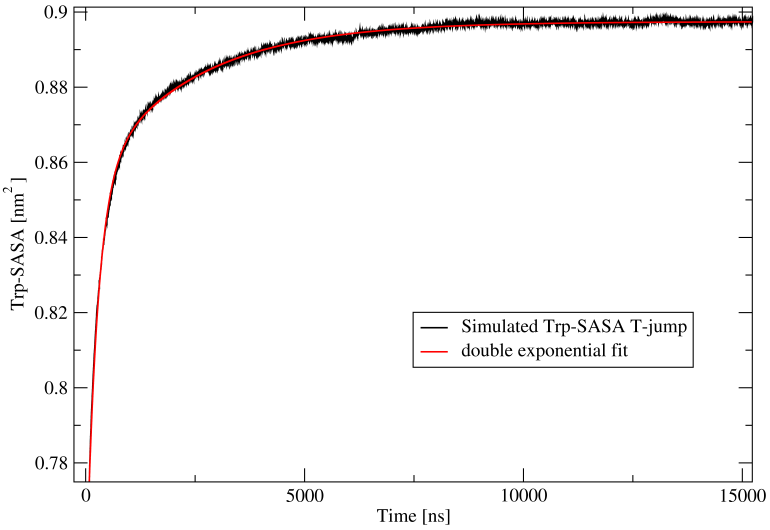


Figure 4.7: **Simulated Trp-SASA T-jump of Trp-cage** Simulated TRP SASA evolution as a function of time at 298 K starting from an initial distribution at 291 K (black line). The red line is a double exponential fit to the data. The two time constants of fit are $\tau 1$= 248 ns, $\tau 2$= 2313 ns. The diffusion matrix entering in the kinetic model was calculated using several MD simulations for a cumulative time of $\sim$ 500 ns. A time lag of 12 ns was used in the maximum likelihood approach for calculating **D**.

A double exponential decay model describes very accurately the data ($R^2$ = 0.9986, see Fig. 4.7). The two time constants are $\tau_1 = 248$, and $\tau_2 = 2313$ ns. The large gap between the first and the second time constant is a strong indication of two-state behavior. The value of $\tau_2$ is in agreement with the experimental relaxation time of 3.1 $\mu$s for the florescence T-jump[105]. This shows that the rate model is capable of reproducing accurately the dynamics of the real system, at least for what concerns the relaxation of fluorescence. The microscopic rearrangements that determine $\tau_2$ will be discussed in detail in the next section. The influence that the error on the free energies and on the enthalpies has on the results is $\sim$ 500 ns. The error deriving from neglecting the position dependence of $D$ is $\sim$ 500 ns (see section 4.2, paragraph bin-based kinetic model and tables 4.2, 4.1). Thus the overall error on the relaxation time is $\sqrt{(500^2 + 500^2)} \sim 700$ ns. Including the correction suggested in Ref.[133] to take into account the unphysical viscosity of TIP3P water[134] the relaxation time is $\tau_2 = 3763 \pm 1200$ ns, still in fair agreement with experiments.

### 4.2.4 Trp-cage folding dynamics

The rate model constructed for the Trp cage was shown in the previous section to be in fair agreement with the experimental two state folding kinetics[105]. Nevertheless MCL shows the presence of 5 relevant metastable states that are compatible (at

least three of them) with other experimental observations [104, 107, 108]. Here we use this model to investigate the folding mechanism of the Trp cage in order to reconcile the two state kinetics with the presence 5 relevant clusters.

The characteristic times of the system are related to the eigenvalues of the rate constant matrix. Consistently with what is found for the Trp SASA relaxation, the second largest eigenvalue corresponds to a characteristic time of 2447 ns. The third eigenvalue corresponds to 434 ns, with a gap of 2013 ns from the first, consistently with a two state behavior[105]. The second eigenvector has large positive components in cluster 1 and 2 and large negative components in cluster 5. This suggests that the longest relaxation time of the system is associated to a transition between these states. In order to analyze more quantitatively this issue, the rates for the transitions between the clusters found by Markov cluster analysis were extracted from a very long KMC simulation ($\tau_{KMC} = 1.5$ seconds). For two clusters A and B with occupancy $P_A$ and $P_B$, the rate constant to go from A to B was calculated counting the number of times $N_{AB}$ that a trajectory goes from A to B without passing from any other cluster during the KMC simulation. The rate to go from A to B was estimated as $k_{AB} = N_{AB}/(P_A \cdot \tau_{KMC})$. To minimize the number of recrossing, the KMC trajectory is assumed to visit a cluster any time it visits any bin belonging to the group of lowest free energy bins containing 70% of the cluster population. Bins that do not fall in this definition were considered as transition states. The transition rates obtained in this manner are represented in Fig. 4.8. For clarity, all the clusters whose occupancy is below 1% are omitted from the figure. The equilibration between cluster 1 and 2 is rather fast and transition times to cluster 3 are also in the sub-microsecond domain, but when the system reaches cluster 5 on average $\sim 2$ $\mu$s are necessary to return to the folded cluster. The folding pathways schematized in figure are consistent with the two routes proposed by Ref. [83], except for the transitions involving cluster 5. The folding pathway initiating from cluster 4 and passing from cluster 3 is characterized by the early formation of an $\alpha$-helix and resembles the pathway passing from state I in Ref. [83]. The pathway passing from cluster 2 is instead characterized by the formation of several hydrophobic contacts, while the $\alpha$ content remains on average lower. This resembles the pathway passing from state L in Ref. [83]. If the molten-globule state (cluster 5) is neglected the folding and unfolding rates are compatible with those reported in Ref.[122], considering the difference in the force field.

## 4.3 Discussion

Trp-cage is a designed miniprotein that, due to its small size and fast folding rate, has been the object of several theoretical investigations. Here this system is analyzed with a new method, introduced in chapter 2, that allows deriving a kinetic model of the system by analyzing a set of biased MD trajectories. The model shows the presence of several metastable states (clusters). The most populated one can be classified as the folded state. The second most populated cluster has a

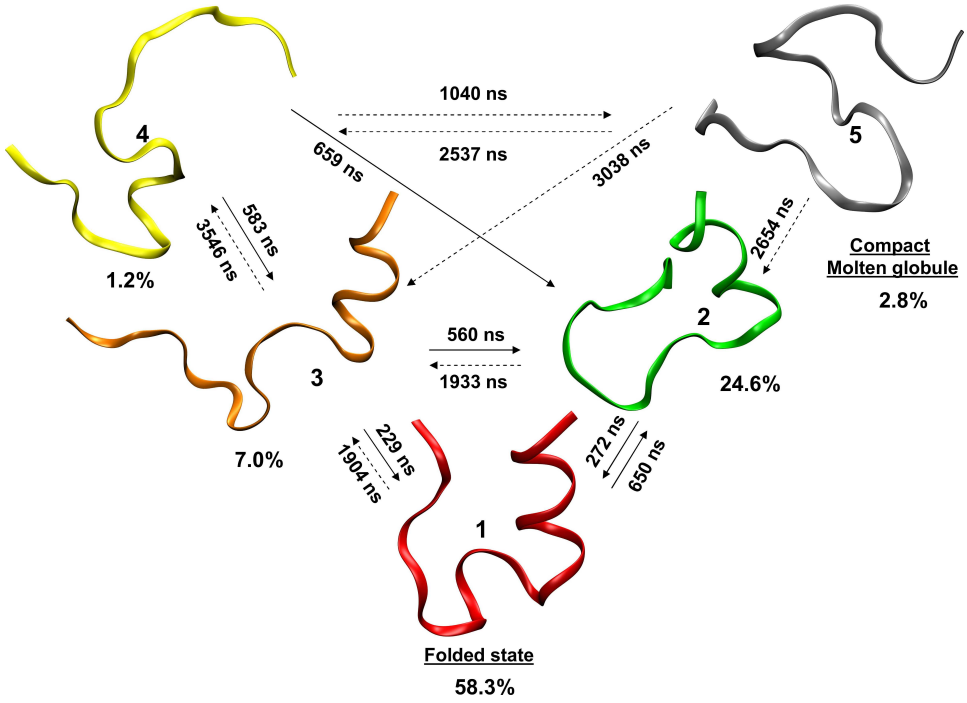


Figure 4.8: **Schematic representation of the Trp-cage folding dynamics.** Times (inverse of rates) for the transitions between the relevant clusters are shown on the arrows. The uncertainty on each transition time due to both the error on the free energies and the position-dependence of $D$ is at most 40%. Only the clusters whose population is higher than 1% are shown. Continuous arrows correspond to direct transitions between clusters that occur on a time smaller than 1$\mu$s. Dashed arrows correspond instead to transition that occur on a time larger than 1$\mu$s or taking place through other intermediate low-populated clusters, not represented in the Figure.

$C_\alpha$ RMSD of $\sim$ 4.4 Å from the NMR structure and retains part of its secondary structure (see Fig.4.3A). In this cluster the Trp is more strongly packed between Gly11 and Pro12 than in the NMR structure and its population relative to cluster 1 increases with temperature (see Fig.4.3B). This can explain the anomalous behavior of the temperature dependence of the CSD of Pro12-$\delta$3 hydrogen atom observed both experimentally[104] and in the simulated NMR experiment (see Fig.4.6B). The cluster 2 and cluster 3 reference structures are consistent with experimental unfolded state distances[108] (see Fig.4.5). The presence of these two clusters is also in agreement with the strengthening of proline(s)-Trp excitonic interactions with temperature and the broad $\alpha$-helix melting observed in Ref.[107].

In spite of the presence of several intermediates both the simulated T-jump experiment (see Fig. 4.7) and the spectrum of the kinetic matrix associated with the rate model are consistent with a two state kinetics[105]. The calculated time constant of the folding process is $\sim 2.3 \pm 0.7$ $\mu$s (or $\sim 3.8 \pm 1.2$ $\mu$s including the correction of Ref. [133]) in fair agreement with the experimental relaxation time[105]. To investigate the folding dynamics using the kinetic model we derived a folding mechanism which involves the detected intermediates (see Fig.4.8). Starting from

open structures, the folding process can follow two main routes. One of them consists in an earlier formation of the N-terminal $\alpha$-helix (cluster 3) followed by the hydrophobic collapse, while the other involves first the formation of hydrophobic contacts with less helical content (cluster 2) and then the completion of both secondary and tertiary structure. This is in agreement with the pathways found in Ref. [83]. The time required to undergo these transitions is in the sub-microsecond time domain, which is less than the slowest relaxation time found in the simulated T-jump experiment and more consistent with the third eigenvalue of the kinetic matrix. Indeed, the folding mechanism (see Fig.4.8) shows that, if Trp-cage reaches the molten globule state, more than 2 $\mu$s are necessary to reach the folded state. This implies that the experimental folding time is ultimately determined by the slow equilibration between the first two clusters and the compact molten globule state that acts as a kinetic trap. In this state no secondary structure element is present, but a hydrophobic core with several tertiary contacts is formed. In Ref.[135] the Pro12Trp mutation brings to an increased stability of the folded state and a faster folding time of $\sim 1$ $\mu$s. This seems to be in agreement with the folding mechanism presented here, since the mutation would strongly stabilize cluster 1 and cluster 2 but not the molten globule cluster. A possible way to assess experimentally the presence of the molten globule could be a mutation of Pro17 to a more polar residue (e.g. Asn) or a chemical modification of this residue as the lower rigidity associated to the absence of the Pro17 ring could destabilize the folded state[136]. In fact in the attractor of cluster 5 Pro17 shows a strong interaction with Trp6, and this interaction does not play a key role in other relevant clusters (see Fig. 4.5).

CHAPTER

**FIVE**

# The Folding Free Energy Landscape of Insulin chain B

In this chapter we present a fully atomistic model of the structural transitions and possible folding pathways of insulin.

Insulin is an important hormone that interacts with the insulin receptor and this regulates the entrance of glucose in the cells. High sugar levels in the blood are a result of reduced secretion or activity of insulin, which can have detrimental effects on the human metabolism[137]. The insulin monomer is composed of two chains, A and B, containing 21 and 30 amino acids, respectively. The monomer contains three disulfide bonds, one is an intra-A chain disulfide A6-A11 and two inter-AB chain disulfides, A7-B7 and A20-B19, which clamp the A chain helices at the end of the central B chain helix. At micromolar concentrations insulin forms dimers, while in the pancreas it is stored as a hexamer in the presence of zinc ions. Upon entrance into the serum, the hexamer dissociates and binds to its receptor as a monomer[138].

There are many structures currently available of insulin and the general binding mode of the insulin-receptor complex is known[139]. The secondary structure features of chain B of insulin are commonly defined as the N-terminus (residues 1 to 8), central $\alpha$-helix (residues 9 to 19), a characteristic type-I $\beta$-turn (residues 20 to 23) and an extended C-terminus (residues 24 to 30). There are several known conformational states for the N-terminus of chain B.[140–144] The principal conformations have been designated as the R-state and T-state,[145] although additional conformations have also been identified.[146] The T-state is associated with insulin's activity and is believed to be representing the monomeric solution state.[147] In the T-state, chain B consists of the $\alpha$-helix (9 to 19) and an extended N- and C-termini regions.[148, 149] In the R-state, all N-terminal residues form $\alpha$-helix, joining with the central helix.[150, 151] Another known variation is the "freyed" or Rf-state, where the $\alpha$-helix is only present for some of the N-terminal residues (4 to 8) in addition to the central helix. A schematic showing each of the described states is presented in fig. 5.1.

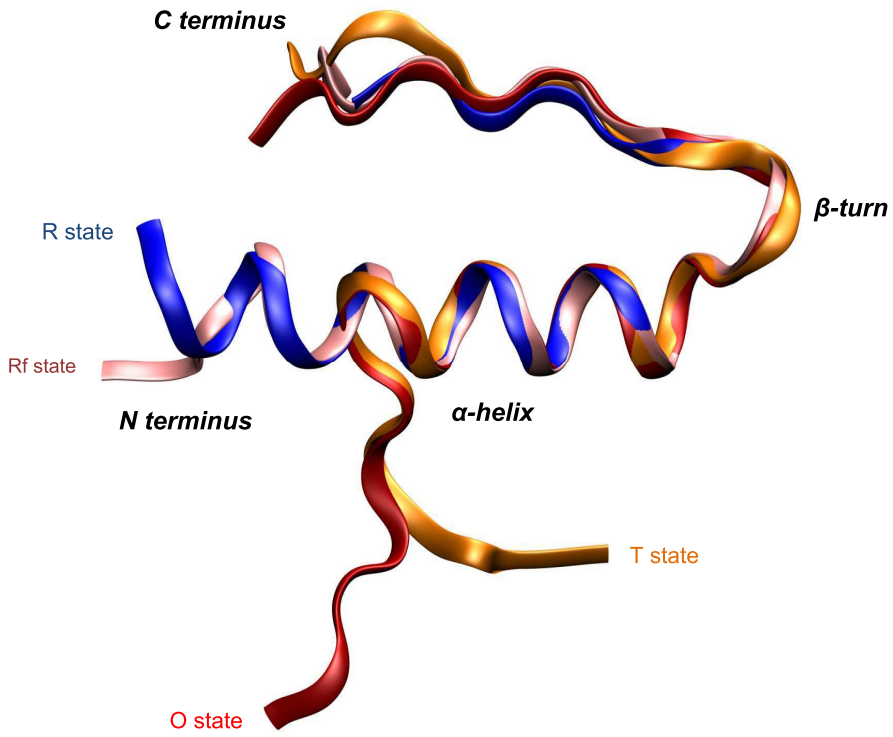


Figure 5.1: **Insulin chain B experimental structures.** An overlay of X-ray crystallographic structures of insulin chain B, showing experimentally observed states: Rf state (1ZNI), R state (1EV3), O state (1B9E) and T state (1LPH).

Chain B of insulin is believed to retain much of its structure independently of chain A.[52–54] Recent study by Budi et al. confirmed this property as they looked at the effect of thermal and chemical stress on isolated insulin chain B and complete insulin monomer[152, 153]. Structure-activity studies of insulin indicate the C-terminus of chain B as integral to receptor information,[55–57] and also suggest that the conformation of the C-terminus is influenced by the structure of the N-terminus.[140, 141, 154] The inherent flexibility of the N- and C-termini regions of chain B of monomeric insulin was observed both by experimental and theoretical studies. Various investigations of insulin,[155, 156] including that on a preliminary crystallographic structure of the native insulin monomer at low pH,[157] and also on a solution structure of isolated chain B determined by NMR spectroscopy,[52, 54] confirmed both termini's mobility. Previous theoretical simulations, molecular dynamics (MD) using the GROMOS 37c forcefield showed conformational flexibility of insulin chain B, reporting a high degree of movement in aqueous solution of both monomer and dimer,[158] similar to the simulations performed by Zoete and coworkers.[159] The effect of electric field on the flexibility of the chain was recently investigated, where it was found that oscillating field had more disruptive effect, while static field stabilized the secondary structure of chain B.[160, 161] Investigation performed by Legge et al.[162] yielded information about the structure and dynamics of insulin with respect to its biological behaviour by performing multiple MD simulations with CHARMM27 forcefield in explicit solvent and ambient conditions. Their work highlighted the importance of packing interactions for the

conformational behaviour of chain B of insulin, specifically structures stabilized by localized hydrophobic interactions. Although insulin has been the subject of a large number of experimental and computational investigations, studies of the monomeric or isolated chain B structure and dynamics in solution have been limited by the monomers' susceptibility to self-associate into oligomers. To the best of our knowledge there is no published study that has successfully simulated the folding of isolated chain B of insulin in explicit solvent.

In this chapter, BE and the accompanying analysis methods introduced in chapter 2 were used to investigate the folding mechanism of insulin, in particular the folding pathways of its chain B in explicit solvent. The structural characterization and biological relevance of the conformations revealed by the simulations is described in the results section of this chapter.

## 5.1 Computational setup

In this study we utilized the Gromacs suite of programs[163] modified to perform bias exchange metadynamics. The AMBER03 forcefield[64] was used for all calculations as it has been previously shown to give a good representation of the experimentally observed behaviour of chain B of insulin.[164] The time step for all the simulations was set to 2 fs. Atom based cutoff of 8 Å was used for nonbonded van der Waals interactions. The Particle Mesh Ewald (PME) summation method[165] was applied to treat long-range electrostatic interactions. All bond lengths were constrained to their equilibrium value with the LINCS[67] algorithm. Constant temperature was achieved by coupling the system to a Nose-Hoover thermostat[70] with a characteristic frequency of 1 ps. Constant pressure was achieved by coupling the system to a Berendsen barostat[71] with a relaxation time of 4 ps.

The starting structure for this study was an extended conformation sampled in a previous work on the effect of electric field[160] on the conformation of porcine insulin (PDB entry 1ZNI[150]). The protein was enclosed in a periodic box of 46 Å 70 Å 46 Å size, then solvated with 4220 TIP3P[65] water molecules, corresponding to water density of $\sim$1.0 g/cm$^3$. The positive charge of the protein was neutralized by adding two Cl- counterions. The whole system was energy minimized to remove steric clashes using the steepest descent algorithm after which 200 ps of NPT molecular dynamics at 298 K and 1 atm were performed to equilibrate the protein and solvent.

For the BE simulations seven generalized reaction coordinates were applied, none of which require *a priori* knowledge of the folded state. A *neutral* walker was also implemented, which is not biased by any metadynamics potential, i.e. evolves as a classical MD simulation, but is allowed to exchange conformations with the other replicas. The neutral walker statistics are approximately canonical as was shown in reference.[38]

### 5.1.1 Collective Variables

The seven collective variables are used in this study for both BE simulation and subsequent analysis:

- $N_{hb}$ is the number of backbone H-bonds ;$N_{C\gamma}$ is the number of $C_\gamma$ contacts and $N_{sb}$ is the number of salt bridges. They are defined as

$$N_{hb,C\gamma,sb} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{1 - \frac{r_{ij}^8}{r_0^8}}{1 - \frac{r_{ij}^{10}}{r_0^{10}}} \tag{5.1}$$

  where $r_{ij}$ is the distance between atoms $i$ and $j$. The distance constraint $r_0$, for $N_{C\gamma}$,$N_{sb}$ and $N_{hb}$ was set to 5.0, 4.5 and 2.0 Å, respectively.

- $\Psi_{\alpha 1}$ is the $\alpha$ fraction of the $\psi$ backbone dihedral angles applied to the 1st half ($^2$VNQHLCGSHLVEAL$^{15}$) of the protein; $\Psi_{\alpha 2}$ is the $\alpha$ fraction of the $\psi$ backbone dihedral angles applied to the 2nd half ($^{16}$YLVCGERGFFYTPK$^{29}$) of the protein. They are defined as

$$\Psi_{\alpha 1,2} = \sum_{i=1}^{N} \frac{1}{2} \left[1 + \cos\left(\psi_i - \psi_0\right)\right] \tag{5.2}$$

  where $i$ runs over all the residues belonging respectively to the 1st half of the protein ($\Psi_{\alpha 1}$) and the 2nd half of the protein ($\Psi_{\alpha 2}$). $\psi_0$ = - 50

- $\Psi_{corr1}$ is the dihedral correlation applied to the 1st half; $\Phi_{corr2}$ is the dihedral correlation applied to the 2nd half of the protein. They are defined as

$$\Psi_{\text{corr } 1,2} = \sum_{i=2}^{N} \sqrt{\left[1 + \cos^2\left(\psi_i - \psi_{i-1}\right)\right]} \tag{5.3}$$

  where $i$ runs over all the residues belonging respectively to the 1st half of the protein ($\Psi_{corr1}$) and the 2nd half of the protein ($\Phi_{corr2}$).

The terminal residues were excluded from these CVs to permit their natural flexibility.

The number of hydrogen bonds in $N_{hb}$ were calculated based on contacts between the $H_N$ and O backbone atoms of the protein. The potential salt bridge pairs in $N_{sb}$ were considered between the C of the carboxylic groups and the N$\zeta$ of Lysine and the C$\zeta$ of Arginine.

The rationale for choosing these CVs is that they are relevant for the description of possible free energy barriers between peptide conformations. The free energy barrier associated with the formation or disruption of H-bonds is described by the $N_{hb}$ variable. The backbone conformational changes are affiliated with the $\Psi_{\alpha 1}$, $\Psi_{\alpha 2}$, $\Psi_{corr1}$ and $\Psi_{corr2}$ variables. The $N_{C\gamma}$ variable describes barriers associated with the formation of hydrophobic clusters, and the salt bridge variable describes

barriers associated with the formation of salt bridges. These variables are similar to those used in chapter 4 for studyng the Trp-cage folding.

Each simulation was performed with 8 replicas, one *neutral* and one for each collective variable. Exchanges between each replica were allowed every 20 ps of MD simulation. Gaussian potentials of height 0.1 kJ $mol^{-1}$ were added to the time-dependent potential every 500 steps (1 ps) during the whole MD simulation. The width of the Gaussians, which ultimately determines the resolution of the free energy reconstruction, for each collective variable $N_{hb}$, $N_{C\gamma}$, $N_{sb}$, $\Psi_{\alpha 1}$, $\Psi_{\alpha 2}$, $\Psi_{corr1}$ and $\Psi_{corr2}$ was chosen to be 2.0, 2.0, 0.5, 0.8, 0.8, 0.5 and 0.5, respectively. The rate and accuracy of exploration of the free energy surface depends on the chosen width and height of the Gaussians, in the same manner as in the ordinary metadynamics. Each replica was evolved for 96 ns, producing accumulated total of 768 ns.

### 5.1.2 Molecular docking

In this work we have used molecular docking to predict how insulin chain A and insulin chain B may interact (see sec. 5.2.4). The HADDOCK program[166, 167] was used to perform molecular docking simulations. This tool uses bioinformatics and experimental information to define distance restrains that are then used along the docking of two up to 6 biomolecules. The used procedure for docking with HADDOCK is composed of two steps: 1) randomization of orientations and rigid body energy minimization (EM); 2) semirigid simulated annealing (SA) in torsion angle space. From one step to the other, the structures are ranked in terms of HADDOCK scoring, and the best ones may proceed to the next step. This scoring is a weighted sum of electrostatic, van der Waals (energetics obtained from a modified version of the OPLS force-field[168]), desolvation energy, buried surface area and terms that take into account the artificially added restraints.

In the first stage the two molecules are separated by 25 Å and rotated randomly around their center of mass. Then cycles (in this work we used 20) of rigid body EM are performed, where both molecules are rotated, followed by two cycles of rotational and translational rigid body minimization.

The best structures in scoring terms, (tipically 200 structures) will proceed to stage 2, a semi-flexible simulated annealing in torsion angle space. This semi-flexible annealing consists of several stages: a) high temperature rigid body search (2000K, 500 steps) b) Rigid body SA (cooling step from 2000K to 500K, 500 steps) c) Semi-flexible SA with flexible side-chains at the interface (from 1000K to 50K, 1500 steps) d) Semi-flexible SA with both backbone and side-chains flexible interface (from 500K to 50K, 1500 steps).

Finally, the docking solution are clustered (based on pairwise backbone RMSD at the interface) and sorted based on HADDOCK score.

## 5.2 Results

With the computational setup described above, we performed BE simulation of chain B of insulin at 298 K, starting from an extended conformation using 8 replicas. The the data produced by this simulations are analyzed following the approach presented in chapter 2

### 5.2.1 Bin-based thermodynamic model

Cluster analysis was performed on the statistics accumulated in the last 40 ns of the simulation and ˜17000 bins were found using a cubic side $ds_j=2\sigma_j$. The population of each bin was assigned by the WHAM procedure described in chapter 2 and their free energies were determined. Several values of filling time were examined, such as 30 ns, 35 ns and 40 ns. The results from these filling times were in good agreement, where all of the most populated bins showed similar population, with variation below T. The filling time of 35 ns was used for the rest of the analysis.

The kinetic model was applied to construct the transition matrix between bins used in the MCL method. The MCL method enabled the determination of the clusters of the system. The bins contained within a cluster are structurally similar. The transitions between bins that belong to the same cluster are faster compared to the transitions occurring between the clusters. The parameter *p,* used in the MCL method controls the height of the energy barriers for clustering. The MCL calculation was performed with different values of the parameter $p = 1.08$, 1.12 and 1.14. Each $p$ was also applied to the free energies derived at the different filling times. We found that the population of the clusters obtained from the various filling times is consistent and all most populated clusters are maintained.

With $p = 1.08$ three clusters were obtained with population of 70%, 25.5% and 4.5%, respectively. As they are obtained with a low $p$ value a high free energy barrier is expected between them. The most populated cluster (molten-globule 1) contains bins that show several salt bridges between Glu13, Glu21, Lys29, Arg22 and the terminal residues. The second most populated cluster (molten-globule 2) is stabilized primarily by hydrophobic contacts and the last cluster contains mostly structures of the native-fold nature. For $p = 1.12$, molten-globules 1 and 2 split into several sub-clusters for a total of 11 clusters with population above 1%. At $p = 1.14$ the most populated clusters were not fully divided so we chose the results at $p = 1.08$ and $p = 1.12$ for structural characterization and analysis.

The low free energy bins for each of the three clusters obtained for $p = 1.08$ are shown in fig. 5.2.

The most populated cluster contains structures mainly governed by electrostatic interactions, where the charged sidechains of the most stable bin (see fig. 5.2a) are closely interacting with the terminal regions of the protein. In the low free energy bins of molten-globule 2 (fig. 5.2b,c), residues Phe24 and Phe25 are packed into a compact hydrophobic core (less than 1 kcal/mol from the lowest free energy bin). This implies that these two residues play an important role in the stabilization of

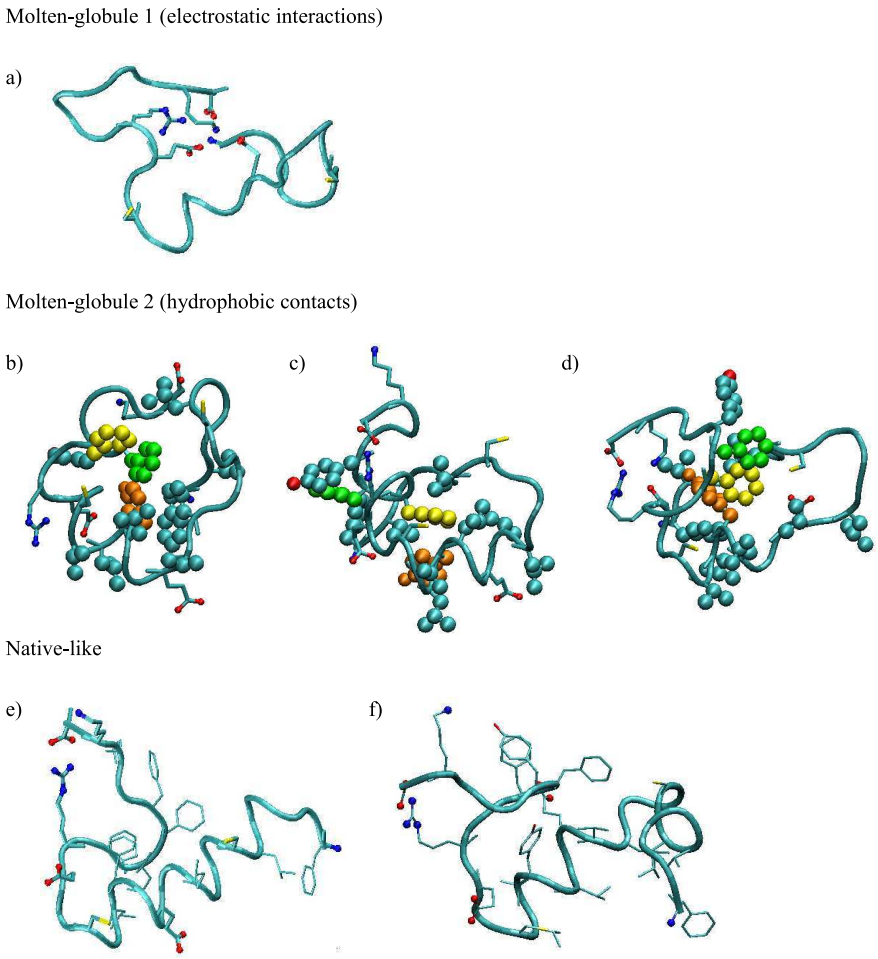


Figure 5.2: **Low free energy bins representative structures of the three clusters determined using p = 1.08.** Molten-globule 1 a) Most stable bin of the cluster, where all charged residues interact with the N and C termini of the protein, giving rise to strong electrostatic contacts. Molten-globule 2 b) Most stable bin of the cluster, where Phe25 (green) is buried in a hydrophobic pocket, while Phe24 (yellow) is partially exposed to the solvent. c) This structure is only 0.9 kcal/mol from the most stable bin and shows the presence of a hydrophobic core in which Phe24 (yellow) is present. Phe25 (green) is partially solvent exposed in this cluster. d) In this bin the Tyr16 (orange) residue and partially Phe24 (yellow) are buried in a hydrophobic pocket. The hydrophilic part of Tyr16 is directed toward the Arg11, Gln21, and the terminal salt bridges. e) Lowest free energy bin in the folded state cluster. f) Bin present between the native-like and molten-globule 2 cluster having the N-terminal $\alpha$-helix unfolded.

the cluster. In the less stable bin Tyr16 can be found shielded from the solvent (see fig. 5.2d). Part of the central $\alpha$-helix is still retained in the structures within the molten-globule 2 cluster. Possible role of the above mentioned residues in the folding mechanism of chain B is discussed later. The most stable bin of the native-like cluster resembles the Rf-state of chain B (see fig. 5.1, fig. 5.2e, fig. 5.2.1). The N-terminal $\alpha$-helix can be found unfolded in some less stable bins (see fig. 5.2f) belonging to the same cluster or near the border of the native-like and molten-globule 2 clusters. This structure resembles states T and O of chain B of insulin shown in fig. 5.1. Detailed structural comparison with experimental data was performed on the conformations contained in the folded state cluster and the

results are presented below.

Thermodynamic properties of the three clusters are listed in table 5.1.

Table 5.1: **Thermodynamic properties of the clusters found using the MCL algorithm with p= 1.08.** All thermodynamic properties were calculated with respect to the native-like clusters.

| | Molten globule 1 | Molten globule 2 | Native-like |
|---|---|---|---|
| %occupancy | 70.0 ± 8.0 | 25.5 ± 1.0 | 4.5 ± 1.0 |
| ΔH (kcal/mol) | 1.4 ± 0.3 | 7.0 ± 1.0 | 0.0 ± 0.8 |
| ΔS (kcal/mol K) | 0.010 ± 0.0015 | 0.027 ± 0.004 | 0.0 ± 0.003 |

All observables and error evaluations are calculated according to the method described in chapter 2. Both molten-globules 1 and 2 show an enthalpic penalty with respect to the native-like cluster, although for molten-globule 1 the difference is only slightly larger than the standard error. The two molten-globules are entropically stable. As expected the cluster with the largest entropy is molten-globule 2, since it has the highest content of hydrophobic contacts.

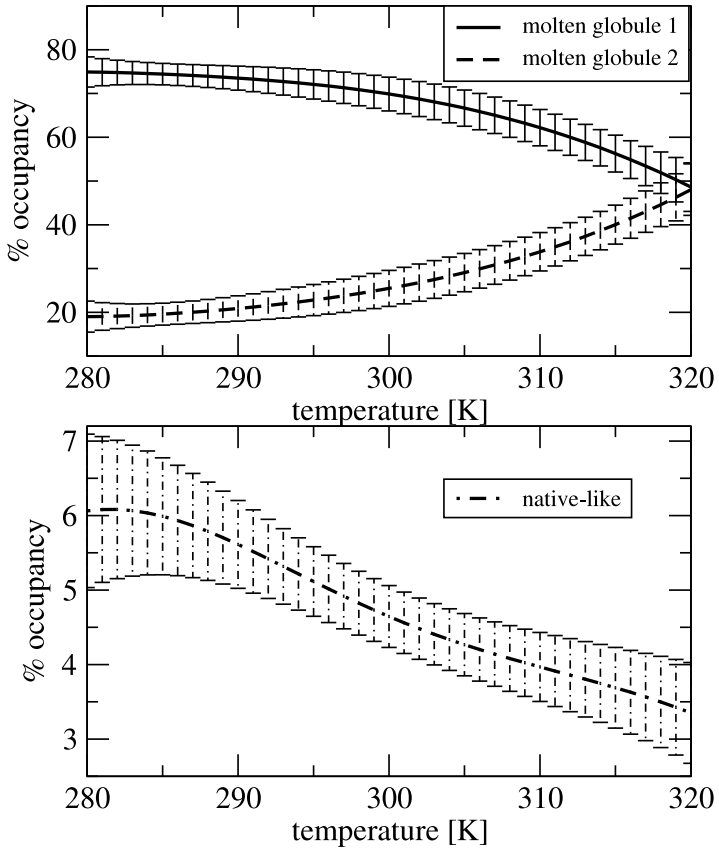


Figure 5.3: **Occupancy (%)of the three clusters found for p = 1.08 as a function of temperature.** A simple linear extrapolation from the calculated bins enthalpy and entropy at 300 K have been used to obtain the clusters occupancy versus temperature.

In fig. 5.3 the occupancy of each cluster as a function of temperature is reported.

According to its entropy content, the population of molten-globule 2 increases as temperature is raised. The native-like cluster occupancy does not vary strongly with the temperature and the two molten-globules show complementary behaviour. The presence of this type of conformations is supported by NMR studies performed by Hua et al.[169, 170] which found local differences between the solution structures of insulin and the crystal structure, suggesting that the biologically active form of insulin is a molten-globule.

Fig. 5.4 shows the eight most populated clusters obtained for $p = 1.12$, represented in three dimensions, as contours around the most populated state: the C-gamma contacts, the $\alpha$-dihedral fraction in the 1$^{st}$ half of the protein ($^{2}$VNQHLCGSHLVEAL$^{15}$) and the $\alpha$-dihedral fraction in the 2$^{nd}$ half of the protein ($^{16}$YLVCGERGFFYTPK$^{29}$). Each color corresponds to a different cluster. The lowest free energy bin of each cluster is represented as a sphere of the same color and its 3D structure is depicted in the order of increasing energy.

Although the most populated bin in our simulations is not the native state of chain B, we sampled structures resembling the R and Rf-state presented in fig. 5.1.

It is evident from fig. 5.4 that there is a large energy barrier i.e. a kinetic gap, between the folded state cluster and the other partially folded clusters. The structural stability of the folded state could depend on factors, such as the binding to chain A. Specifically, insulin's three disulfide bridges (A6-A11, A7-B7, and A20-B19) play a critical role in the protein synthesis, structure and stability. To investigate the effect of non-native disulfide pairing on insulin's structure and biological activity, Hua et al.[171] prepared by direct chemical synthesis two insulin isomers having disulfide bonds between the following pairs: (1) A7-A11, A6-A7, A20-B19 and (2) A6-A7, A11-B7, A20-B19. Using CD and NMR spectroscopy they found that the engineered isomers have less helical content compared to native insulin. Their thermodynamic studies by CD-detected guanidine denaturation demonstrated that their non-native disulfide paired isomers are markedly less stable than the native insulin, suggesting that this instability is in qualitative agreement with the isomers' lower $\alpha$-helix content.

### 5.2.2 Structural analysis of the folded state cluster

Structural analysis was performed on the folded structures identified in cluster 5 of fig. 5.4 which exhibit conformational elements typical for the X-ray crystallographic states of chain B shown in fig. 5.1. The equilibrium averages of the observables considered were calculated using the free energies obtained using eq. 2.32

To the best of our knowledge the structure of isolated chain B of porcine insulin has not been determined. Most of the published work has been performed on engineered insulin monomer[172] and NMR studies have been presented on mutated[54] and oxidized[52] isolated chain B. A comparison of the structures contained in our simulated folded state cluster have been performed with the solution NMR structure of isolated chain B of insulin,[52] X-ray crystallographic structure[150] and

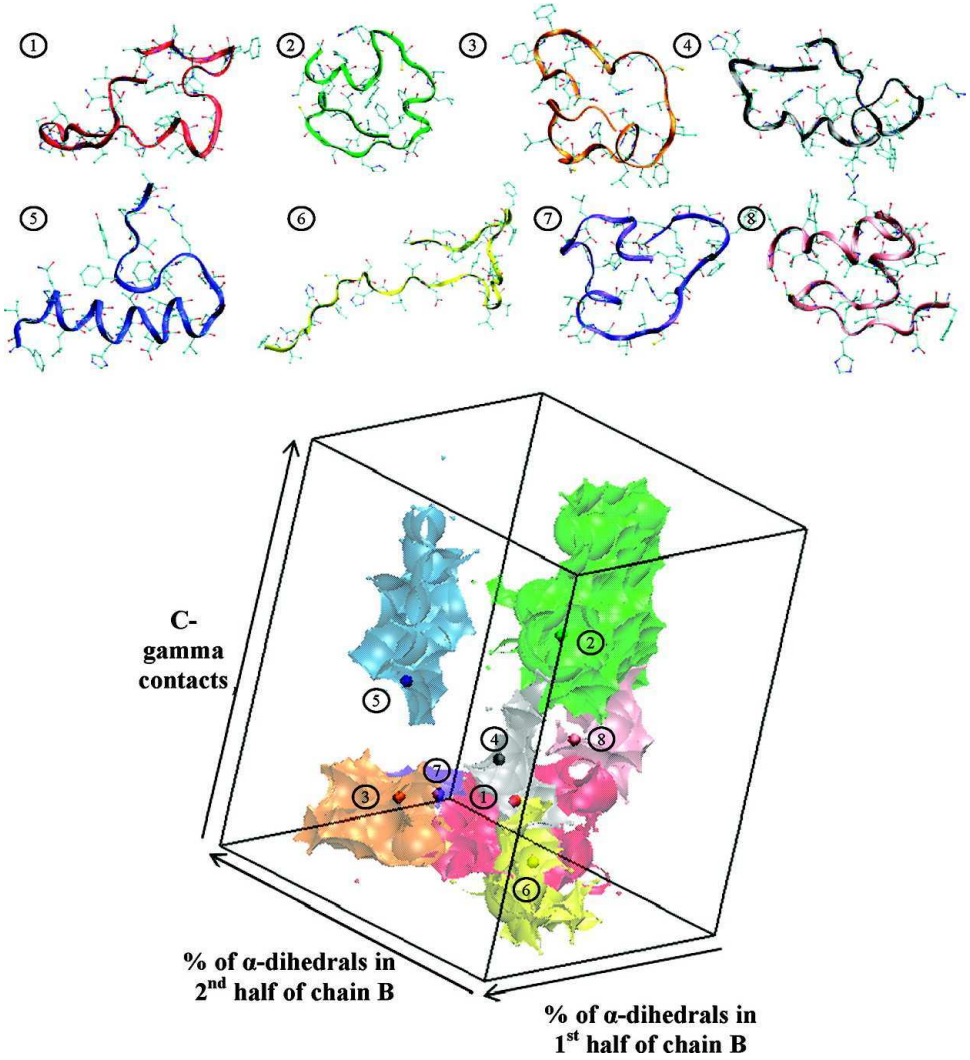


Figure 5.4: **Metastable sets (bin clusters) detected by MCL using p = 1.12.** The clusters are shown as colored contours, with the colored sphere corresponding to the lowest free energy bin of each cluster. The respective structures are presented above with the same color code and ordered based on their free energy (1 is the lowest, 8 is the highest free energy).

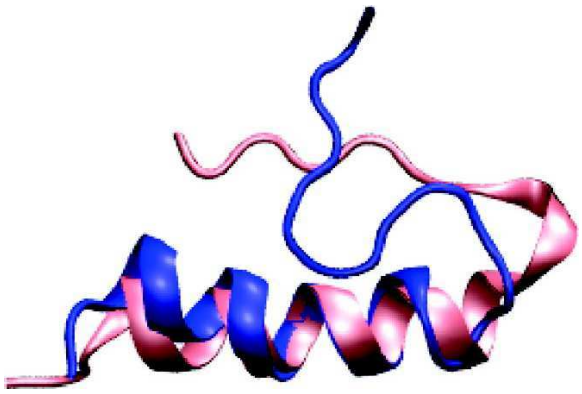

Figure 5.5: **The lowest energy structure (blue) from the folded state cluster superimposed with the X-ray crystallographic structure of chain B, PDB code 1ZNI (pink).**

previous molecular dynamics study on chain B which reproduced experimental conformations.[162, 164] Conserved structural features of chain B were identified

and their RMSD was calculated from the crystallographic Rf-state. The interproton distance violations of the folded structures based on nuclear Overhauser enhancement (NOE) distance restraint data were also analysed in detail.

The data derived from the NMR structure of bovine insulin chain B in solution provide a useful benchmark for evaluating the conformations sampled during the BE simulations. Hawkins et al.[52] derived NOE constraints from a 250 ms NOESY spectrum of the oxidized chain B at 500 MHz, 300 K and pH of 2.2 to 2.5. As coordinates for the structure were not available, NOE distance restraints were used for validation of the conformations obtained.

In the folded state cluster, on average 13 out of 309 restraints were violated, with only 6 severely over their upper limit (violation $> 1.0$ Å). A mean interproton distance violation of 0.13 Å was also calculated. Most violations were found at the C and N termini regions. The same NOE distance restraints, as implemented in ref [164], were used to validate the structures sampled. The total number of violations in the two individual simulations was 13 and 9 respectively, with 7 and 6 violations considered severe. Several studies have alluded to the difficulty in the structural definition of the C-termini region of chain B due to its inherent flexibility.[172] Moreover, studies have shown that receptor binding must be accompanied by a major conformation change in the carboxyl terminus of chain B.[173]

To further inspect the conformations from the folded state cluster, we calculated the average equilibrium backbone RMSD of the important structural elements of chain B, such as the $\alpha$-helical region (residue Ser9 to Cys19) and $\beta$-turn region (residue Gly20 to Gly23). The central helix of chain B plays a key role in the insulin's activity, while the conservation of the $\beta$-turn between residues Gly20, Glu21, Arg22 and Gly23, plays an important part in the folding and conformation stability.[174] The Gly20 to Gly23 turn is integral to the chain B secondary structure because it enables the C-terminal $\beta$-strand to pack against the central $\alpha$-helix. Although the turn is more flexible than these adjoining structural elements, its pattern of hydrogen bonds and dihedral angles is essentially identical among multiple crystal forms.[140–144] Therefore, RMSD calculations were performed from the crystallized Rf-state of chain B to investigate the stability of the helix and $\beta$-turn within the folded state cluster. An overlay of the helical region of the lowest energy bin from the folded state cluster (structure 5, fig. 5.4) and the crystal structure of chain B, taken from the PDB code 1ZNI is depicted in fig. 5.4.

The average RMSD over the helical region in the folded state cluster is 2.2 Å, where that for the lowest energy state is 1.7 Å. This result is illustrated by the observed alignment between the lowest energy state and the crystal structure, shown in fig. 5.4. We also found excellent agreement with the RMSD calculated from the classical MD calculations performed in ref [164] on the folded crystal structure, where an average RMSD of ˜ 1.9 Å was obtained for 2 independent 50 ns simulations.[164] The turn region between residues 20 and 23 was calculated to have an average RMSD of 1.8 Å, with the lowest energy bin having deviation of 1.3 Å. Overall, these results are in agreement with experimental and theoretical data

obtained by a different approach.

The key elements in directing insulin-receptor interactions and in formation of insulin dimers are the well conserved Phenylalanine residues at B24 and B25. Mutational analysis has been extensively applied to investigate the importance of these residues in the hormone insulin (see references in [175]). The conservation of secondary structures and support in intermolecular association are two distinct roles associated with the benzyl side chains of Phe24 and Phe25.[175–178] It has been suggested that Phe24 interacts with the hydrophobic core of insulin, specifically with B12, B15 and the A20-B19 disulfide bond. Furthermore, Phe24 side chains have shown to benefit the stability of the $\beta$-turn at B20-B23,[173] however it may not interact directly with the receptor, unlike Phe25 which is more exposed to the solvent and is easily accessible for receptor interaction. We support these findings by the observed conformations of Phe24 and Phe25 in the folded state (structure 5, fig. 5.4). A turn is formed at the location of the two aromatics as a result of a salt bridge formed between Ala30 and Arg22, resulting in the Phe25 to be slightly solvent exposed. Aromatic ring interaction between Phe25 and His5 is preserving the partial packing of the C-terminal $\beta$-strand against the $\alpha$-helix. Furthermore, residue Phe24 is strongly interacting with the $\alpha$-helix and reduces the flexibility of the $\beta$-turn. These observations are in agreement with the proposed roles of Phe24,25 and give insight into the possible structural transformation C-terminus adopts upon folding.

### 5.2.3 Folding pathway of chain B of insulin

The dynamics of the system was investigated by applying the Kinetic Monte Carlo (KMC)[124] method to compute the transition between the clusters found by the Markov cluster analysis. The method described chapter 2 was applied. This enabled a construction of a reduced rate model in which only transitions between the clusters are considered. The rate constants for this model are given by the inverse of the transition times. For example, by taking two clusters A and B with occupancy, $P_A$ and $P_B$, the rate constant to go from A to B is calculated considering the number of times ($N_{AB}$) a trajectory goes from A to B without passing any other clusters during a long KMC simulation ($\tau_{KMC}$=1.5 seconds). In this way the actual time of transition to go from A to B is estimated as $k_{AB}=N_{AB}/(P_A \cdot \tau_{KMC})$. To minimize the number of recrossing only the stable bins of each clusters were considered, i.e. ˜70% of the cluster population. The kinetic scheme of the insulin chain B folding is outlined in fig. 5.6. The clusters are organized in a similar arrangement to their contour representation in fig. 5.4. The results show that the folded cluster is connected directly only to the molten-globule 2 cluster, while this cluster is connected with that of the molten-globule 1, thus forming an overall linear pattern (see fig. 5.6).

Transitions between these clusters occur in a few thousands of nanoseconds, suggesting that the residence time of the three clusters is of the order of several

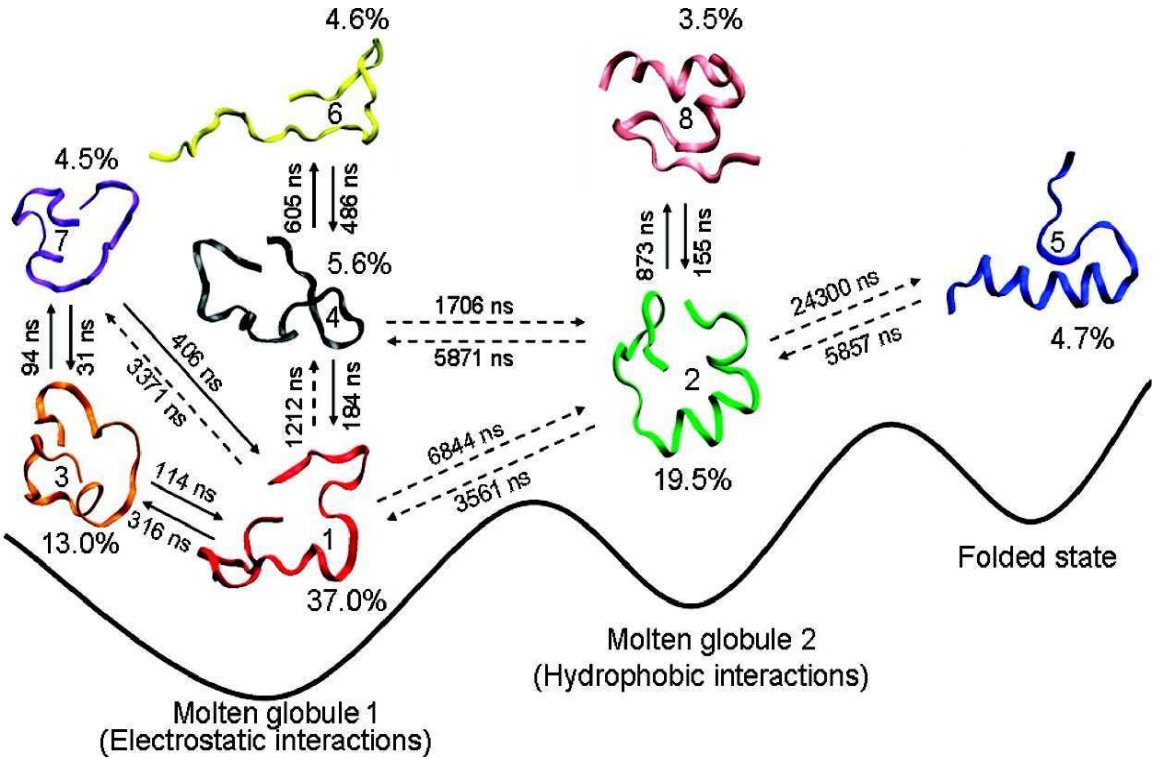


Figure 5.6: **Schematic representation of the clusters dynamics.** Transitions between the clusters are represented by an arrow along with the corresponding transition time. Transition that occur on a time scale longer than 1 s are represented as dashed arrows.

microseconds. Transitions between sub-clusters (obtained with $p = 1.12$) inside each well (obtained with $p = 1.08$) occur in tens or hundreds of nanoseconds. The transition region between the folded state and the molten-globule 2 is made of states that contain unfolded N-terminal helical region, similar to the T and O state of chain B (fig. 5.1, fig. 5.2f). This indicates that the first step of going from the folded state to the nearest molten-globule state is by unfolding the N-terminal $\alpha$-helix (fig. 5.2e,f). In the second step the N-terminal region interacts with the remaining helical elements forming a compact intermediate in which the Tyr16 is buried in a hydrophobic pocket (fig. 5.2d). All these transitions involve the high free energy bins (e.g. 3-5 kcal/mol from the most stable bin). The last step involves a partial exposure of Tyr16 to allow Phe24 to be packed via hydrophobic contacts (see fig. 5.2c), which results in a gain in free energy. The passage from molten-globule 2 to 1 seems to involve the intermediate structure presented in fig. 5.3d. In fact, the state displayed shows a salt bridge formation at the chains terminal regions. The last step involves the expulsion of the Tyr16 from the pocket forming the electrostatic core (fig. 5.2f). It is worth to note that many of the bins associated with the insulin chain B folding could not exist in the presence of disulfide pairing with chain A as this would result in a loss of flexibility (see section below).

### 5.2.4 Docking with insulin chain A

To analyze the ability of different insulin chain B conformations to covalently bind insulin chain A, molecular docking simulations were performed between several conformations of chain A and the reference structures of the relevant bins of insulin chain B. The insulin chain A conformations were extracted from several (NMR and X-ray) PDB entries. For the insulin chain B the relevant bins reference structures (having free energy lower than 6 kcal/mol with respect to the most stable bin of the reference cluster) for each cluster were selected. We used the Haddock program to perform all the docking simulations as described in section 5.1.2. The SG insulin A -SG insulin B distance was restrained to an upper bound of 5.2 Å to allow a good initial docking pose for disulfide bond formation. We used the standard procedure of docking with Haddock package[166, 167] (see sec. 5.1.2). The docking solutions are clustered[92] based on pairwise backbone RMSD at the interface. Cut-off for clustering was chosen by analyzing if: 1) the number of clustered structures was more than the 50% of the docking solutions; 2) the best structure (in term of HADDOCK score) was included in one of the clusters. From this simulations we found that the lowest free energy bin together with most of the relevant bins of the folded cluster are in optimal conformation for the two disulfide bridge formation with chain A.

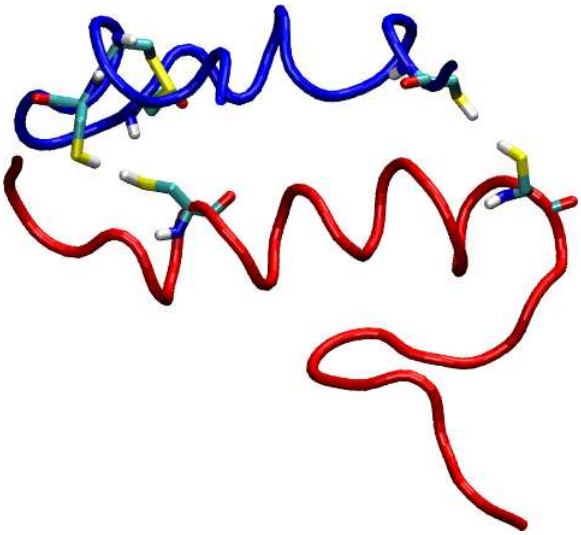

Figure 5.7: **The best docking pose between insilin chain A and the insulin chain B folded state cluster.** This structure corresponds to the highest score conformation of the highest score cluster found using the Haddock package. Chain A is colored in blue, instead chain B in red.

In Fig. 5.7 it is shown the best docking pose found between insulin chain A and the insulin chain B folded state cluster. From the picture it is clear that the sulfur atoms of both chains can get close to each other without steric repulsion. The insulin chain B structure in Fig. 5.7 corresponds also to the lowest free energy bin of the folded cluster.

The docking simulations show also that a few conformations of the others chain B clusters might form disulfide bond with insulin chain A but they correspond

mostly to high free energy bins. This indicates that the two disulfide bridge formation with chain A would stabilize the folded state cluster with respect to the others.

## 5.3 Discussion

Explicit solvent bias exchange metadynamics simulations were performed to effectively sample the conformational space available to chain B of insulin and to shed some light on the complex structural transitions this important protein undergoes upon folding. To exploit the statistics accumulated using this powerful technique, the bin-based analysis described in chapter 2 was used. This allowed constructing a model describing the complex conformational transitions chain B experiences. The model suggested the existence of three metastable clusters separated by large free energy barriers. The two most populated clusters had structures with "molten-globule" characteristics, one being governed by electrostatic interactions and another by mainly hydrophobic contacts. This finding is supported by experimental studies which suggested this type of conformation to be biologically active.[169, 170] The third cluster comprised conformations with folded structural elements, resembling the known crystallographic states of chain B ($\alpha$-helix, $\beta$-turn and flexible termini). The folded state cluster contained physiologically important features, such as: a well conserved $\alpha$-helical content, a $\beta$-turn stabilized by interaction between Phe24 and the hydrophobic core of the $\alpha$-helix; structural transformation of the C-terminus, favourable for possible binding to the insulin receptor.

The kinetics of the system can be qualitatively described by a three state model for the folding pathway of insulin chain B. Starting from an extended structure, at first the protein is governed by electrostatic interactions (molten-globule 1, fig. 5.2a). A progressive building of hydrophobic core is initiated by the burial of the Tyr16, followed by further packing of Phe24 and Phe25 (molten-globule 2, fig. 5.2b,c), resulting in stable compact structures. Furthermore, the hydrogen bonding interactions between the buried backbone groups commence the formation of an $\alpha$-helix at the core of the protein. An unfolded N-terminal region is found in the structures at the border of molten-globule 2 and the folded basin, suggesting that the last stage of the folding of chain B is the complete formation of the $\alpha$-helix. The transformation from molten-globule 2 to a folded state requires crossing of a high barrier. Tens of microseconds are required to make this transition. The calculated transition times gave further insight into the dynamics between the three wells, suggesting that the residence time of the three wells is of the order of several microseconds. Docking simulations suggest that the binding with chain A stabilize mostly the folded cluster.

CHAPTER

**SIX**

# Substrate binding mechanism of HIV-1 protease from explicit-solvent atomistic simulations

In this chapter we show that the technique described in chapter 2 can be used also for studying binding of drugs. In particular, substrate binding to HIV protease. Since more than 20 years the Human Immunodeficiency Virus Type-1 Protease (HIV-1 PR) is one of the main targets of anti-AIDS drug design. HIV-1 PR cuts polyproteins to smaller fragments, and is essential for the virus life cycle. Its inactivation leads to non-infectious viral particles [179]. Many inhibitors have been developed to block the aspartic protease active site, and several of them are used in medical treatment. Most FDA-approved HIV-1 PR inhibitors are peptidomimetic, i.e. they mimic the structure of a fragment of the natural substrate, competing with its binding [180, 181].

The experimental literature about the structure and function of HIV-1 PR is vast. X-ray crystal structures of HIV-1 PR in bound and unbound forms have been reported, revealing a $C_2$ symmetric homodimer with a large binding pocket covered by two Gly-rich $\beta$-hairpins (flaps). The bound and unbound forms of the enzyme show sizable structural differences: the complex has "closed" flaps, i.e. in contact with the ligand [180]. On the other hand the free enzyme can also adopt a "semi-open" conformation with the flaps shifted up from the active site and only partially in contact with each other [180]. The residues that do not belong to the flaps show smaller displacements upon binding. The flexibility of the flaps is confirmed by solution NMR [182, 183] and fluorescence experiments [184, 185]. In the unliganded protease, the semi-open form is considered predominant, but there are also indications of the presence of a scarcely populated truly open form with separated flap tips [182, 183]. Arguably, an extended opening of the flaps is necessary to allow binding of the viral polyproteins, due to the large size of the substrate. It is not clear if the binding of peptidomimetic inhibitors is a one-step or a two-step process [184, 185]: in the latter case, binding would proceed through the

fast formation of a collision complex, followed by a slower conformational change to a tighter complex [184].

HIV-1 PR has been also largely investigated by computational methods. In several works the binding affinity of different inhibitors has been computed by molecular docking based on scoring functions [186, 187] or on the more predictive Molecular Mechanics Poisson-Boltzmann Surface Area model (MM/PBSA) [188, 189]. Although computationally expedient and very useful for drug design, these approaches might be too approximate to capture the correct physical chemistry of the binding process. In particular, they typically do not account for the flexibility of the whole protein and for hydration. Water molecules can make hydrogen bonds simultaneously with both the enzyme and the ligand. This is believed to lower sizably both the enthalpy and entropy of binding [190]. In particular, a water molecule (usually named W301) is hydrogen-bonded between the flaps tips and the ligand in most experimental complexes with peptidomimetic inhibitors [180]. In order to address the role of water and the flexibility of the system appropriately [191] one can use explicit-solvent molecular dynamics (MD). The flap dynamics of the free enzyme has been simulated by MD observing spontaneous opening and closing events [192, 193]. Interestingly, full opening of the flaps has been suggested to be unnecessary for dissociation of Saquinavir from protease mutants [194]. Coarse grained [195, 196] and implicit solvent MD [195, 197] have further provided insight on the binding of ligands to the proteases. In particular, in Ref. [195] it is suggested that a cyclic-urea inhibitor can bind without full opening of the flaps. Unfortunately, to date the high computational cost has limited the timescale of explicit-solvent MD to 100 ns. This is not sufficient to observe binding and unbinding processes. Thus an enhanced sampling technique is required to obtain a realistic picture of the all process.

For systems of the complexity of HIV-1 PR, several variables may simultaneously play an important role for binding, e.g. the opening of the flaps, the distance between the ligand and the cavity, the number of hydrogen bonds or hydrophobic contacts with the flaps or with the cavity, the number of interfacial water molecules, etc [191]. Here we adopted BE [38], to obtain a comprehensive picture of the substrate binding and unbinding mechanism of wild-type HIV-1 PR. The calculations are validated against experimental thermodynamic and kinetic data. We focus on the substrate Thr-Ile-Met-Met-Gln-Arg (p2-NC cleavage site of the gag-pol viral polyprotein). The choice of this ligand is motivated by the fact that most of the inhibitors in clinical use are peptidomimetic, i.e. they are based on such natural substrate, and they might share with it several features.

Our simulation shows several binding events, starting from the ligand outside the enzyme and ending in the Michaelis complex, which is predicted to be the lowest free-energy minimum. A quantitative kinetic model of the association and dissociation process obtained here using the approach reported in chapter 2 has allowed us to calculate binding free energies and rate constants that turn out to be in agreement with available experimental data. Hydration and flap fluctuations

turn out to play a key role for the binding. Remarkably, the main binding pathway of the small substrate does not involve full opening of the flaps, which is instead expected to occur, for topological reasons, when the enzyme binds the long viral polyprotein *in vivo*. Several mutations that have been reported to bring drug resistance involve residues forming important contacts with the substrate in the molecular recognition process. Because of the similarity between the substrate and some FDA-approved drugs it can be speculated that these mutations would affect at a smaller extent the binding of the polyprotein expressed by the virus, as this binding would take place through a different pathway. Thus, mutations have high chances of influencing differently the binding kinetics of small ligands like drugs and of the natural substrate.

## 6.1 Computational setup

MD simulations have been performed with the AMBER03 force field [64] for the enzyme and substrate (hereafter termed SUB), and the TIP3P [65] model for water. The initial atomic model has been obtained starting from the experimental coordinates of the HIV-1 PR/MVT-101 complex (pdb code 4HVP), as described in a previous work [198]. HIV-1 PR and SUB have been solvated by 7710 water molecules in a 269 $nm^3$ orthorhombic periodic box. 6 $Cl^-$ions were added to neutralize the net positive charge. The particle-mesh Ewald method [69, 97] was used for long-range electrostatic with a short-range cutoff of 0.8 nm. A cutoff of 0.8 nm was used for the Lennard-Jones interactions. All bond lengths were constrained to their equilibrium length with the LINCS [67] algorithm. Also the $C_\gamma$(Asp25)-$C_\gamma$(Asp25') distance was constrained to 0.34 nm. The time step for the MD simulation was 2.0 fs. NPT simulations at 300 K and 1 atm were performed by coupling the system to a Nose-Hoover thermostat [70, 199] and a Berendsen barostat [71], both with relaxation time of 1 ps. After 1.4 ns of equilibration, the barostat was removed and the BE simulation was started. The atomic coordinates were saved every 5 ps, the energy every 0.1 ps.

The indexing 1-99 and 1'-99' is adopted for the two dimers forming HIV-1 PR. Based on experimental [200] and theoretical evidence [201], Asp25 has been taken deprotonated and Asp25' monoprotonated. The substrate is indexed as Thr(P3)-Ile(P2)-Met(P1)-Met(P1')-Gln(P2')-Arg(P3'), the scissile bond being P1-P1'. As a reference experimental structure for the HIV-1 PR/SUB complex, the crystallographic positions of the complex between the inactive D25N protease and ATIM-MQRG substrate [202] (pdb code 1KJ7) are considered. The crystallographic water molecule located in the cavity under the flap tips and above SUB is called W301 following the usual terminology.

The BE approach allows biasing simultaneously several collective variables (CVs). The following set of 7 CVs has been selected as putative reaction coordinates to explore the binding mechanism. The CVs are explicit functions of the atomic coordinates:

- $S_A$ is the sum of hydrophobic sidechain carbons contacts between ligand and flaps. $S_B$ is identical but counts contacts between the ligand and the part of cavity not belonging to flaps. They are defined as

$$S_{A,B} = \sum_i \sum_j C(R_{ij})$$

with the sums running on the appropriate sets of atoms. Contacts are defined by the following function, which switches smoothly from 0 to 1 for distances below a threshold $R_0$:

$$C(R) = [1 - (R/R_0)^n]/[1 - (R/R_0)^m]$$

where $R$ is the distance between two atoms, $R_0 = 0.3$ nm, $n = 8$, and $m = 12$.

- $S_C$ is defined like $S_A$ and $S_B$, with the sum running over the possible H-bonds among O and H atoms in the ligand and in the cavity.

- $S_D$ is the distance between the center of the scissile peptidic bond of the ligand and the center of the $C_\gamma$(Asp25)-$C_\gamma$(Asp25') atom pair in the catalytic dyad.

- $S_E$ is the distance between the center of the $C_\alpha$s in the left flap tip (residues 48-53) and the center of those in the right flap tip (residues 48'-53').

- $S_F$ counts the number of water molecules bridging between the ligand and the cavity:

$$S_F = \sum_i \sum_j \sum_w C(R_{iw}) \cdot C(R_{wj})$$

where the sums over $i$ and $j$ run over O/H atoms (corresponding to native H bonds) in the ligand and in the enzyme, respectively, while the sum over $w$ runs over all O atoms belonging to water molecules.

- $S_G$ is the distance between the center of the $C_\alpha$s of the ligand and the center of the $C_\alpha$s of residues 24, 26, 27, 24', 26', 27', located in the middle of the enzyme and close to the Asp25 - Asp25' catalytic dyad.

It has to be stressed that variable $S_C$ provides only an approximate count of the H bonds, since a more rigorous definition would include the angles formed by atoms and a more sharp switching function $C(R)$. Similarly, the number of water molecules bridging through H bonds between ligand and cavity is only approximately proportional to $S_F$. However the present definitions are more suitable to be used as differentiable collective variables. The CVs were saved every 0.1 ps.

To reduce the computational cost walls have been put on variables $S_D$ and $S_E$ preventing them to reach values larger than 1.9 nm. Moreover, the ligand was restricted to a cone of angle 45° with axis equal to the $C_2$ symmetry axis of HIV-1 PR. Control simulations have been also performed beyond the restrictions above, in order to check the convergence of the results.

The parameters adopted in the BE simulations are the following: Gaussian height 0.05 kcal/mol, Gaussian widths equal to 0.5 for $S_{A,B,C,F}$ and 0.02 nm for $S_{D,E,G}$, deposition of a Gaussian every 1 ps, exchanges of bias attempted every 2 ps. These parameters have been optimized in order to obtain a fast exploration of the conformations while still retaining a good accuracy in the reconstructed free energy.

BE simulations were performed biasing each of the 7 CVs on a different replica (plus one replica without any bias) for a total of 45 ns per replica. We initialized the simulation with the substrate completely outside the cavity and misoriented with respect to the complex, namely with Arg(P3') pointing towards the cavity. After a few ns, in several replicas the substrate approaches the cavity and starts to thread inside the pocket starting from Thr(P3). After 10 ns in one replica SUB entered the binding pocket, and after 35 ns 4 out of 8 replicas fully accomplished the binding process by reaching the structure of the Michaelis complex (6.2 and state B1 in 6.1).

During the simulation, a total of 4 and 10 independent binding and unbinding events were observed. This is sufficient to ensure an accurate and reproducible description of the binding/unbinding process and allows harvesting enough statistics for constructing an accurate rate model.

To improve the statistics on the explored states, a simulation has been also performed replicating 4 times each replica, for a total of 32 replicas and 40 ns each. This redundancy speeds up the convergence of the reconstructed free-energy profiles. The trajectories from both the 8-replicas and 32-replicas simulations, for a total simulation time of 1.6 $\mu$s, have been employed to construct a kinetic model, as described in the following section.

### 6.1.1 Construction of the kinetic model

Starting from the BE simulations data, a thermodynamic and kinetic model of the binding process has been constructed, applying the methodology of chapter 2. A careful analysis shows that in order to describe accurately the thermodynamics and kinetics of the binding process it is sufficient to consider variables C, D, E, and F, as the others are correlated to these. First, the BE trajectories have been analyzed by subdividing the CV space of these four variables in a hypercubic grid of 3208 bins with sides in the four directions $R_C = 1$, $R_D = 0.1$ Å, $R_E = 0.1$ Å, $R_F = 1$. Molecular structures within each bin differ by C$\alpha$-RMSD $< 2$ Å (substrate plus enzyme cavity), indicating that this choice of variables appropriately discriminates among all the relevant structures. The equilibrium free energy of each bin has been computed by the weighted-histogram technique[13] reported in chapter 2. It has been verified that multiplying by 1.5 or dividing by 1.5 the side of the hypercubes has no relevant qualitative effect on the description of the system, while it deteriorates the accuracy of the thermodynamic and kinetic models. Similarly, the analysis has been repeated adding other variables and deriving the kinetic model in 5 or 6 dimensions. Also this leads to larger errors but no qualitative changes in the

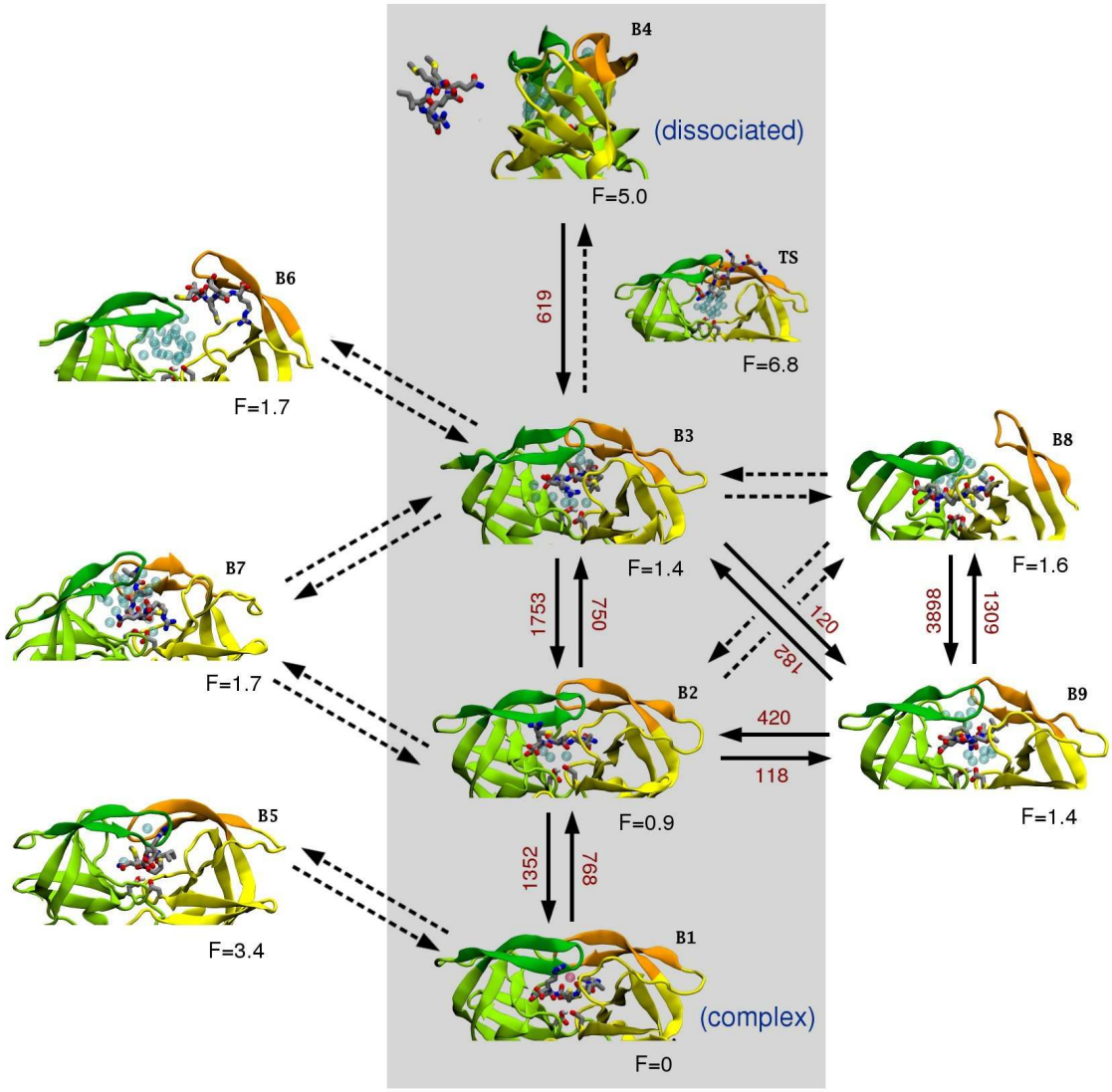


Figure 6.1: Representative structures of the free energy clusters (B1-B4) involved in the main binding/unbinding mechanism (shaded box), plus the transition state (TS). The clusters B6-B9, which are not involved in the main binding pathway, are also displayed. Water molecules inside the enzyme cavity are displayed as blue spheres, except for W301 which is a red sphere. The free energy of each state, in kcal/mol, is reported below the corresponding structure. Arrows are labeled with the corresponding transition rates ($ms^{-1}$) when larger than 100 $ms^{-1}$ (solid line). Transitions with a smaller associated rate are depicted as dashed arrows.

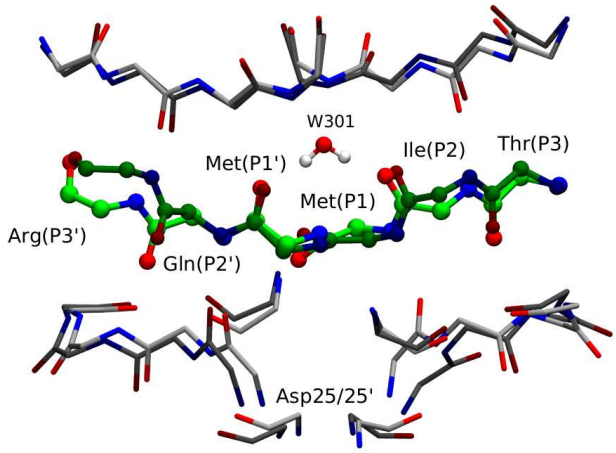


Figure 6.2: Comparison of the lowest free-energy structure from simulations (B1 in 6.1 and 6.1, dark colors) with the experimental structure of the complex between the inactive D25N protease and substrate ATIMMQRG (pdb code 1KJ7, light colors) [202]. Backbone atoms are displayed. The substrate is shown in green. W301 is the crystallographic water molecule bridging between flap tips and the substrate.

overall picture. The error on the bin free energies, estimated according to standard weighted-histogram analysis as detailed in chapter2, is at most 0.5 kcal/mol.

The kinetic model has been constructed following the procedure of chapter 2 [48]. The diffusion matrix $D_{ij}$ in the space of CVs C, D, E, and F. $D$ is estimated from a 10 ns MD simulation started in HIV-1 PR/SUB complex by maximizing the likelihood of the trajectory within the kinetic model [46]. For a time lag of 100 ps the following values are obtained: $D_{CC} = 0.27$, $D_{DD} = 0.0011$, $D_{EE} = 0.0011$, $D_{FF} = 0.35$, $D_{CD} = -0.0033$, $D_{CE} = -0.0013$, $D_{CF} = -0.073$, $D_{DE} \approx 0$, $D_{DF} = 0.0034$, and $D_{EF} = -0.0021$. By enlarging the time lag to 200 ps the coefficients vary by less than 16%, indicating that on this timescale a Markovian behavior is attained [48]. The position dependence of $D$ has been investigated by computing $D$ also from 10 ns MD trajectories started in each of the other metastable states along the main binding pathway (states B2, B3, and B4 in 6.1). The maximum variation in the diagonal elements of the diffusion matrix is only 30%, which leads to similar variations in the relaxation times of the system. By comparison, the maximum estimated error on the bin free energies is 0.5 kcal/mol, which leads to almost one order of magnitude larger variations on the computed transition rates. Therefore the position dependence of $D$ is neglected.

The metastable states (clusters) of the network of bins have been found by the Markov cluster analysis (MCL) algorithm [126] using $p = 1.2$. This allows identifying 9 kinetic clusters (free-energy basins), corresponding to metastable states of the enzyme-ligand system, which together include 99.1% of the equilibrium population. These states have been characterized by the atomic structures corresponding to the kinetic attractors (6.1, 6.1).

Within each of the clusters B1, B2, and B3, atomic structures differ by $C_\alpha$ RMSD of maximum 2.5 Å. Using a smaller MCL parameter $p = 1.15$ brings into

Table 6.1: Description of the kinetic clusters (local free-energy minima) explored by bias-exchange metadynamics simulations

| Basin | $\Delta G$ (kcal/mol) | SUB | Flaps | Water in cavity | Comments |
|---|---|---|---|---|---|
| B1 | 0.0 | in | closed | 1 above SUB (W301) | crystallographic structure of MVT-101 |
| B2 | 0.9 | in | closed | 3-4 under SUB and laterally | |
| B3 | 1.4 | in | closed | 7-10 all around SUB | |
| B4 | 5.0 | out | closed | SUB fully solvated | |
| (TS) | 6.8 | out | closed | SUB almost fully solvated | H-bonds Arg(P3')-Glu35, Thr(P3)-Gly48', Met(P1')-Gly49', Met(P1')-Gly51' |
| B5 | 3.4 | in | closed | 1 between flap tips | similar to B1, but W301 moved between flap tips |
| B6 | 1.7 | half in | one up | SUB almost fully solvated | SUB between flaps and loop P79-T80-P81-V82 |
| B7 | 1.7 | in | open laterally | $>$ 10 above SUB and laterally | |
| B8 | 1.6 | in | one up | $>$ 10 above SUB and laterally | SUB conformation similar to B1 |
| B9 | 1.4 | in | one up | SUB almost fully solvated | |

coalescence the structurally distinct clusters B1, B2, and B9 in 6.1 (which differ for the amount of water in the enzyme cavity or for the opening of the flaps). A larger $p = 1.3$ leads to the fragmentation of cluster B2, which is structurally homogeneous ($C_\alpha$ RMSD $\leq 2$ Å).

To allow comparison with kinetic experiments, the long-time scale dynamics of the system has been modeled on the network of bins by generating a kinetic monte carlo (KMC) [123] trajectory of 100 s. The trajectory, starting from the complex, performs $\approx 6000$ transitions between the complex and the dissociated state. From the analysis of the trajectory, the most probable binding/unbinding pathways between selected states are identified as the lowest free-energy paths. Transition rate constants between a pair $a$ and $b$ of bins are computed as $k_{a\to b} = N_{a\to b}/P_a t_{tot}$, where $N_{a\to b}$ is the number of transitions, $P_a$ is the probability of state $a$, and $t_{tot}$ is the duration of the trajectory. This procedure gives directly the dissociation rate constant (units $s^{-1}$) taking $a$ as the bin corresponding to the experimental complex (lowest free energy in cluster B1), and $b$ as the bin corresponding to the dissociated pair (lowest free energy in cluster B4). To compare with experimental data at standard conditions, the association rate constant (units $M^{-1}s^{-1}$) is obtained by scaling $k_{b\to a}$ with the ratio between the standard concentration (1 M) and the simulated one (0.2 M), which is equivalent to scaling the probability of the dissociated state by the inverse of the ratio. The transition rates have been reduced by a factor 2.26 to correct for the self-diffusion coefficient of TIP3P water which is larger than the experimental one [203].

### 6.1.2 Poisson-Boltzmann calculations

To check the effect of the finite simulation box on the free energies computed from BE, (linearized) Poisson-Boltzmann calculations have been performed on several HIV-1 PR/SUB structures using the program apbs [204] . The following parameters have been used: grid spacing 0.45 Å, ion exclusion radius 2.2 Å, solute dielectric constant 2, continuum solvent dielectric constant 78.5, boundary conditions based on focusing, solute surface defined by a probe sphere of radius 1.4 Å.

The free energy correction (at zero ionic strength) to bring the ligand from state B4 (see Fig. 6.1) to infinity is $< k_B T$, which confirms that SUB does not sizably interact with HIV-1 PR in this state.

In the same manner, it has been checked by a Poisson-Boltzmann calculation that insertion of one $Cl^-$ ion in the cavity in absence of SUB requires 4.0 to 12.0 kcal/mol at ionic strength in the range $0 - 0.1$ M, due to the negative polarization of the catalytic cavity. Indeed the $Cl^-$ counterions never enter the enzyme cavity during the BE simulations.

## 6.2 Results

### 6.2.1 Binding and unbinding processes

From the BE trajectories a kinetic model has been constructed based on the weighted-histogram approach described in chapter 2. The calculated lowest free-energy path passes through the following states (Fig. 6.1, table 6.1):

- **B4**. SUB is solvated without any contact with the enzyme; the flaps are quite closed (see below).

- **TS**. SUB outside the cavity, perpendicular with respect to the orientation in the complex, with Thr(P3) close to the cavity. H-bonds are formed first with Asp30', then with Gly48', Gly49', and Gly51' in one flap and with Glu35 (salt bridge with Arg(P3')) on the loop in front of the cavity (Fig. 6.3).

- **B3**. The cavity, enlarged by a moderate displacement of the flaps, allows SUB to enter (starting from Thr(P3)) together with a solvation shell.

- **B2**. The water molecules bridging between SUB and the flap tips are expelled, tightening the cavity.

- **B1**. The water molecules bridging between SUB and the catalytic dyad are expelled. A water molecule (W301) is introduced between SUB and Ile50-Ile50' on the flap tips. The experimental complex is formed (SUB+cavity all-atom root mean square deviation (RMSD) 1.6 Å compared to experimental structure [202], backbone RMSD = 0.9 Å. Cavity is defined as residues within 4.5 Å from SUB. See Fig. 6.2).

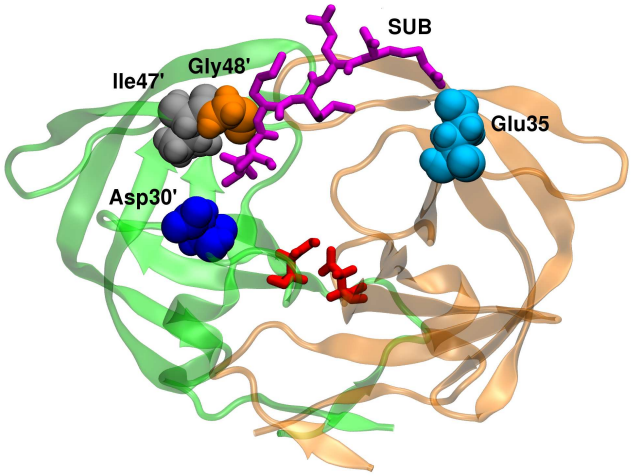


Figure 6.3: The protease residues forming H-bonds with the substrate in either the transition state or the earliest stage of binding (between B4 and TS in Fig. 6.1) are labeled. The catalytic dyad Asp25-Asp25' is shown in red.

The most populated state (46.9%) is the HIV-1 PR/SUB complex B1 (Fig. 6.1). This result is a prediction of the experimental structure and stability of the complex. The equilibrium population of the cluster corresponding to the enzyme separated from the ligand (B4 in Fig. 6.1) is instead almost negligible.

The values taken by the reaction coordinates along the pathway are reported in Fig. 6.4.

During the binding process, the flap tips undergo moderate displacements, without fully opening: the distance between their tips ($S_E$) fluctuates between 0.55 nm (bound complex B1) and 0.83 nm (transition state TS, Fig. 6.4 and Figure S1 in Supporting Information), the distance between the Asp dyad and flap tips varies between 1.1-1.8 nm (all distances are referred to $C_\alpha$s). The displacements are asymmetric within the flap pair: one flap lifts above the cavity more than the other while approaching the transition state, interacting directly with the ligand. An asymmetric role of the flap tips upon binding is also suggested from crystallographic data of the complex with the NC-p1 substrate [205]. The most probable binding pathway is reversible: the unbinding pathway does not show sizable differences.

### 6.2.2 Thermodynamics and kinetics of the binding process

Our model of the binding and unbinding processes allows computing all relevant thermodynamic and kinetic parameters, which can be compared with experimental data.

The predicted binding free energy is $\Delta G_b \sim F_{B1} - F_{B4} = -5.0$ kcal/mol, with a statistical error of 0.5 kcal/mol. The value of $\Delta G_b$ has been obtained for a simulated molarity of 0.2 M. Correcting for normal conditions (1 M) gives $\Delta G_b = -6(1)$ kcal/mol. Corrections due to the finite size of the simulation box are instead small: indeed by a Poisson-Boltzmann calculation it can be shown that in state B4 SUB is practically not interacting with the enzyme. The final estimate

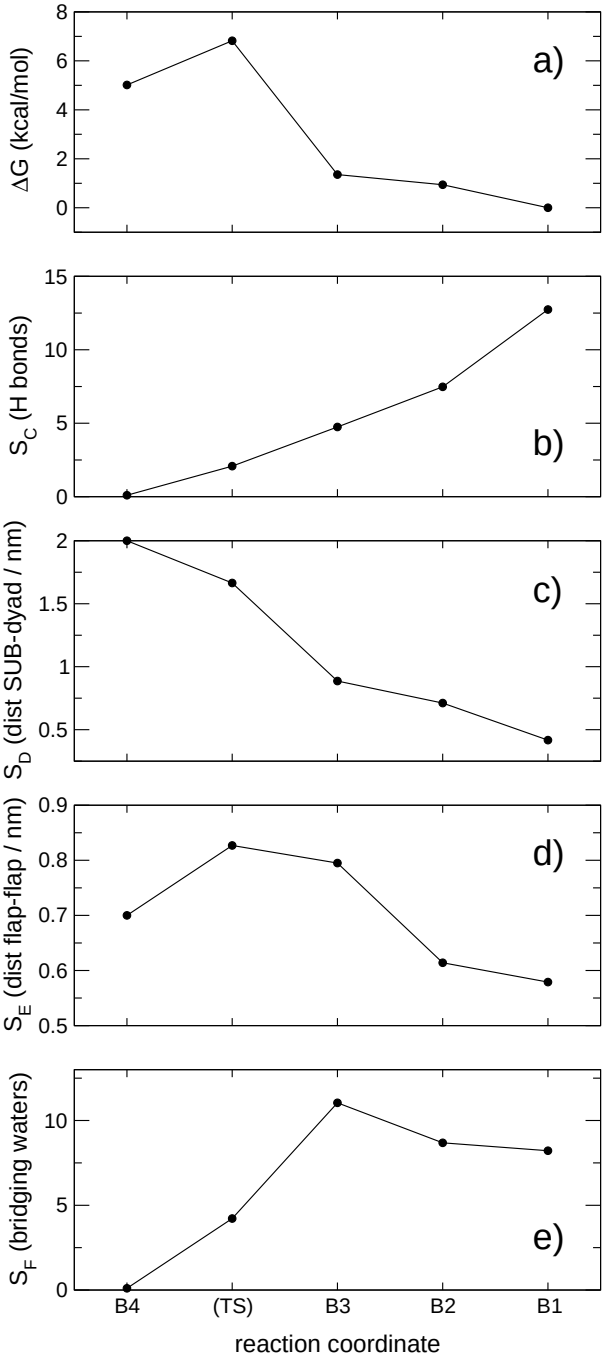


Figure 6.4: Variation of the free energy (panel a) and of selected CVs (panels b-e) along the main pathway for binding and unbinding.

$\Delta G_b = -6(1)$ kcal/mol compares well with $\Delta G_b^{exp} = -8.1$ kcal/mol measured on the peptidomimetic inhibitor MVT-101 [206] which is similar to SUB. Comparison with other inhibitors whose thermodynamic data are available [181] is not reported here as their Tanimoto shape and electrostatic similarity [207] with SUB is significantly smaller.

Also the kinetics predicted by the model can be compared with experiments. Fluorometric assay data are available for peptide substrates similar to SUB (= TIM*MQR): values of $k_{cat} = 41$ s$^{-1}$ and $K_M \equiv (k_{off} + k_{cat})/k_{on} = 3.0 \cdot 10^{-3}$ M have been reported for the substrate ATIM*MQRG (pH= 5.5, ionic strength 0.4 M) [208], while $k_{cat} = 6.99$ s$^{-1}$ and $K_M = 2.5 \cdot 10^{-4}$ M have been reported for TNSATIM*MQRGNF (pH= 5.5) [209]. The rate constants for binding and unbinding transitions, $k_{on}$ (B4→B1) and $k_{off}$ (B1→B4), have been computed from a 100 s kinetic monte carlo trajectory and corrected for a 1 M concentration. The result is $k_{on} = 1.26 \cdot 10^6$ M$^{-1}$s$^{-1}$ and $k_{off} = 57.1$ s$^{-1}$. Employing the experimental values of $k_{cat}$ a theoretical $K_M$ in the range $1.4 - 2.0 \cdot 10^{-5}$ M is obtained. The kinetic rate constants $k_{on}$ and $k_{off}$ are directly accessible from biosensor and fluorescence experiments on inhibitors, which are not cleaved by HIV-1 PR. For the inhibitor MVT-101, analog to SUB, $k_{on} = 1.6 \cdot 10^5$ M$^{-1}$s$^{-1}$ and $k_{off} = 0.2 - 0.4$ s$^{-1}$ have been reported (pH= 5.5) [184]. These values are compatible with our theoretical estimates for SUB, considering the different $\Delta G_b$ (-8.1 kcal/mol for MVT-101 and -6.0 kcal/mol for SUB). In fact, FDA-approved peptidomimetic inhibitors of similar size as SUB, binding more effectively than MVT-101, have $\Delta G_b$ between $-12$ and $-15$ kcal/mol [181], and consistently display lower $k_{off}$ values, in the range $10^{-4} - 10^{-3}$ s$^{-1}$, whereas the $k_{on}$ values are similar to MVT-101, in the range $10^5 - 10^6$ M$^{-1}$s$^{-1}$ [210]. It must be also considered that the kinetic constants display a strong dependence on the experimental conditions like pH and ionic strength: in our simulation the protonation of residues corresponds to pH= 7 and the ionic strength is zero.

### 6.2.3 States with extensive flap opening

The dynamics of flaps opening in HIV-1 PR has been extensively studied, due to its possible functional role [182–185, 192, 193]. Indeed, substates of the protease with open flaps are crucial to allow binding of the long viral polyproteins, for simple topological reasons connected to the large size of the substrate. However, our results show that full opening of the flaps is not necessary to bind the smaller SUB ligand. Our model includes also several states with open flaps (B6-B9 in Fig. 6.1 and Fig. 6.1), which are not part of the most probable binding/unbinding pathway of SUB but which may be relevant for the viral polyprotein. E.g. state B9 is similar to B2, but it has a larger distance between flap tips ($1.1 - 1.8$ nm, compared to $0.4 - 1.1$ nm in B2), and it has a free energy 0.7 kcal/mol higher. In states B6 and B7 the flaps act as "tweezers" which trap the ligand in between. The binding rate associated to the pathway in which SUB approaches the enzyme cavity through wide-open flaps is at least two orders of magnitude smaller than that associated

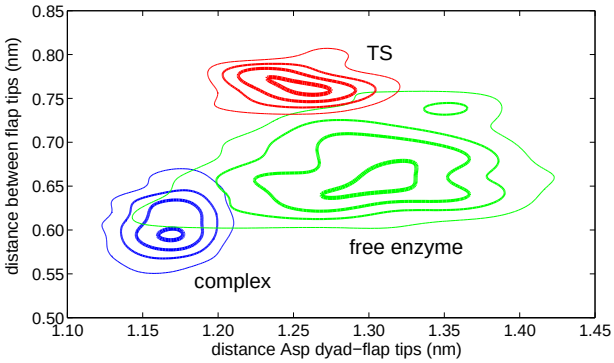


Figure 6.5: Probability distribution of HIV-1 PR conformations in absence of SUB (green), in complex with SUB (blue), and in the transition state (TS) for binding (red). The two variables employed are the distance between Asp dyad and flap tips ($C_\alpha$ Asp25-$C_\alpha$ Gly49) and the distance between flap tips (variable $S_E$, see section 6.1). Equally spaced isoprobability lines are shown.

with the lowest free energy pathway, in which the flaps are closed.

The free energy of HIV-1 PR as a function of the distance between flap tips (CV $S_E$), restricted to configurations with the ligand far from the cavity, shows a preferred distance of 0.6-0.75 nm. 1 kcal/mol is required for opening to more than 1 nm, consistently with Ref. [183, 192, 193]. Instead the complex shows tighter flap tips with a distance of 0.55-0.65 nm (Fig. 6.5 and Figure S1 in Supporting Information).

The associated and dissociated states are also distinguished by a different elevation of the flap tips ($C_\alpha$ of Gly49-Gly49') above the catalytic dyad ($C_\alpha$ of Asp25-Asp25'): 1.1-1.2 nm for the complex and 1.2-1.4 nm for the free enzyme, with the latter displaying also a larger flexibility of the flaps (see Fig. 6.5). These two different conformations compare well respectively to the closed and semi-open structures observed in liganded (e.g. 4HVP) and unliganded (e.g. 1HHP) crystal structures. In agreement with recent solution NMR studies [183] fully open flaps have a negligible population.

### 6.2.4 Hydration inside the enzyme cavity

The crystallographic structure of HIV-1 PR in complex with most peptidomimetic inhibitors shows a water molecule (usually named W301) tetrahedrally hydrogen-bonded between the flap tips and the ligand [180]. W301 is also part of the calculated Michaelis complex (6.2 and state B1 in 6.1). Relocation of this water to a position between the flap tips (stabilized by H-bonds with Gly51, Gly52, Gly49', Ile50', Gly52') is associated with an increase of free energy of 3.4 kcal/mol. This state (B5 in 6.1) is not involved in the main binding pathway. A previous Free Energy Perturbation (FEP) calculation [211] predicted a cost of $3.1 \pm 0.6$ kcal/mol for the displacement of W301 in the HIV-1 PR/KNI-272 complex, while a different FEP calculation on HIV-1 PR/MVT-101[212] displayed a negligible free energy difference. Care should be taken in comparing the free energy value in the latter

calculations with our results due to differences in the setup and in the reference states.

## 6.3 Discussion

The binding and unbinding mechanism of wild-type HIV-1 PR to a model peptide substrate (Thr-Ile-Met-Met-Gln-Arg) has been investigated by molecular dynamics simulations using an all-atom force field with explicit water. To enhance the sampling of the configuration space the bias-exchange metadynamics technique [38] has been employed. The free energy has been computed as a function of 7 reaction coordinates. These have been chosen in an attempt at capturing the most important degrees of freedom associated with binding: the approach of the ligand, the opening of the flaps, the hydrophobic and electrostatic interactions between ligand and enzyme, the amount of interfacial water molecules. Each one of the 7 reaction coordinates has been biased on a different replica of the system, until convergence in the reconstructed free energy projections has been achieved. Starting with the ligand far from the enzyme, several binding events have been observed leading to the Michaelis complex. The substrate slides inside the cavity starting with the Thr head and without sizably opening the flaps. The dissociation and association pathways turn out to be very similar.

The simulation data have been used to build a detailed thermodynamic and kinetic scheme of the binding and unbinding processes following the procedure described in chapter 2. These feature several intermediate states (6.1 and 6.1). The estimated statistical error on the calculated free energies is about 0.5 kcal/mol. The binding free energy has been estimated as -6.0 kcal/mol. This compares well with the experimental value of -8.1 kcal/mol measured with the structurally similar peptidomimetic inhibitor MVT-101. The calculated absolute rate constant for binding is $1.3 \cdot 10^6$ $M^{-1}s^{-1}$ and for unbinding is 57 $s^{-1}$, consistently with experiments on MVT-101 and on peptide substrates similar to SUB.

Flap opening is usually considered to play an important role in the binding of the natural polyprotein substrate. Instead, the extent to which flaps open upon the binding of a small substrate is not clear [213]. Here we found that in the most probable association and dissociation pathways the flaps of the protease do not open sizably and the substrate threads inside the enzyme cavity from the tight lateral channel. When the ligand approaches the enzyme in the early stages of binding, and in the transition state, it forms hydrogen bonds with residues Asp30', Ile47', Gly48', and Glu35, which all lie around the opening of the channel leading to the catalytic cavity (6.3). It is remarkable that mutation of residues in this region causes resistance to some FDA-approved peptidomimetic drugs [214]. We here propose an argument which might help rationalize the origin of such drug resistance, by making the plausible assumption that the binding process observed for SUB is similar to that of other peptidomimetic inhibitors. In fact, the viral polyprotein cleaved in the biological function of the enzyme is expected to bind

through a different pathway. It may approach the cavity from above, after the flaps have opened completely, due to the very large substrate size. Thus, mutations are likely to affect in a different manner the barrier along the two pathways, and may change the relative binding efficiency of drugs and substrate even without affecting their relative affinity. A possible confirmation to our hypothesis would be provided by kinetic experiments measuring the association rate of non-cleavable peptides of increasing length. For longer peptides the preferred binding pathway should switch from the closed flaps- to the open flaps-one, with a resulting significant decrease of the association rate.

A crude estimate of the effect of mutations from a polar residue to an apolar one is provided by alanine scanning[215, 216]. We applied this method on the reference structures of TS, B1, B6, and B8 (the latter two having open flaps, 6.1). Interestingly Glu35, whose mutation to Gly is indicated as generating resistance to a few FDA approved drugs, is found to be an "hot spot"[215, 216] only in TS. This suggests, at a speculative level, that this mutation may play a role in the binding pathway. To quantitatively assess this scenario free-energy calculations of the binding mechanism of inhibitors to drug-resistant variants would be required.

Our results show that, due to sizably different conformations of the flaps, the conformations of HIV-1 PR in absence of the ligand, in the transition state for binding and in the complex have little overlap (see Fig. 6.5). This confirms the appropriateness of docking protocols which account for the flexibility of the receptor by using an ensemble of target enzyme conformations (e.g. relaxed complex scheme [217]).

Individual water molecules at the interface between ligand and enzyme play a pivotal role throughout the binding process, and they constitute a relevant reaction coordinate which enables differentiating the intermediate states explored by the system. Therefore we stress the importance of treating explicitly and accurately the solvent molecules in computational studies of protein-ligand binding.

CHAPTER

**SEVEN**

# Conclusions and perspectives

In in this thesis we have presented an approach aimed at constructing a markovian rate model for complex biomolecular processes starting from a set of biased MD trajectories. We used BE as enhanced sampling technique to simulate complex rare events like protein folding and protein ligand binding. We applied this simulation technique also in another study[39], not reported in this thesis, in which it was possible to predict the effect of a point mutation to the folding of Villin and Advillin. As the method to reconstruct free energies reported in this thesis was still not developed at that time, thermodynamic properties were there calculated as in classical hamiltonian replica exchange by accumulating the statistics of an unbiased walker. The approach presented in this thesis can be considered an improvement of this work and of the one presented in ref.[38]. There the focus was primarily on the capabilities of BE and on the thermodynamic properties of the systems studied. Here we were focused at extracting both *thermodynamic* and *kinetic* properties from BE. This was achieved by constructing a markovian kinetic model from the simulation data similar to the one of refs. [42–47]. As a first step reference structures are extracted through a binning procedure in the space of selected collective variables. The populations of these states are then evaluated exploiting the statistics accumulated during a BE simulation. The use of biased trajectories allows achieving an excellent accuracy also at the transition regions. Rates are estimated assuming a simplified form in which the kinetic prefactor is determined by a diffusion matrix, that is estimated using a maximum likelihood approach using short standard MD trajectories, not necessarily being ergodic.

This approach was applied to a benchmark system for which it was possible to explore the relevant conformations using a standard MD of $\sim 2\mu$s. An excellent correlation for both thermodynamics and kinetic properties was found between the kinetic model and MD. This approach was then applied to the folding of Trp-cage and Insulin chain B and to study the binding mechanism of a little peptide to HIV-1 protease. The results shows that the model allows obtaining a detailed description of thermodynamic and kinetic properties of complex biomolecules in agreement with experimental evidence. Several metastable states were extracted

for all the systems studied that are intermediates in the kinetic mechanism. The meaning of this states can be understood in terms of a separation of time scales: the equilibration inside each state is faster than transitions from a state to another. In the case of folding these represent "misfolded" intermediates, instead for the protein-ligand interaction study they are binding intermediates in which the ligand can be far from the protein, or in a partially binded conformation.

Although with the model presented we could obtain results that are in fair agreement with experiments there are several improvements that can be done to increase the efficiency and the accuracy of the model:

- The CVs and the parameters used for BE simulation can be optimized to obtain a more efficient sampling. In order to address this issue we performed a recent study[218] in which a simplified protein force field was used and the simulation parameter were selectively changed to find the best combination.

- The *binning* procedure to extract the reference states could be improved by constructing a non uniform grid in the CVs space, e.g. by using the iterative approach presented in ref. [42]. Optimizing the bin construction would lead to a smaller number of bins, increasing thus the free energy accuracy and also the time scale at which a markovian behavior is observed.

- The CVs used for the kinetic model construction can be optimized. Also this would reduce the time scale at which a markovian behavior is observed.

- Assuming that the bias alters only the bins populations the diffusion matrices can be calculated with the same methodology using directly continuous BE trajectories. This also could increase the accuracy of the diffusion matrices close to transition state regions.

The methodology presented in this thesis was already used by independetn researchers (see ref. [41]) to extract thermodynamics information on the catalytic mechanism of cis-trans isomerization of Cyclophilin A. Recently[219] the present approach was also applied to investigate the binding of selected drugs to the prion protein.

At the moment we are applying the same technique to study the binding mechanism of Barnase with Barstar. To this purpose explicit solvent BE simulations ($\sim$ 44000 atoms) were performed using CVs similar to one used for the binding of the HIV-1 with the small peptide. Starting from the two moieties completely separated and with a big layer of water in between (Fig. 7.1A), a structure very close to the X-ray complex, was obtained after only 10 ns of simulation and using only six replica. In Fig. 7.1C it is shown the protein RMSD of the residues at the interface respect to the native complex as a function of the simulation time for one of the six walkers. From the picture it can be noted that after $\sim$10 ns the RMSD is smaller than 2Å. In Fig. 7.1B and 7.1D it is shown the lowest free energy structure obtained at the end of the BE simulation (structure in blue) aligned with

the equilibrated X-ray structure (structure in red). As it can be noted the two structures are almost identical.

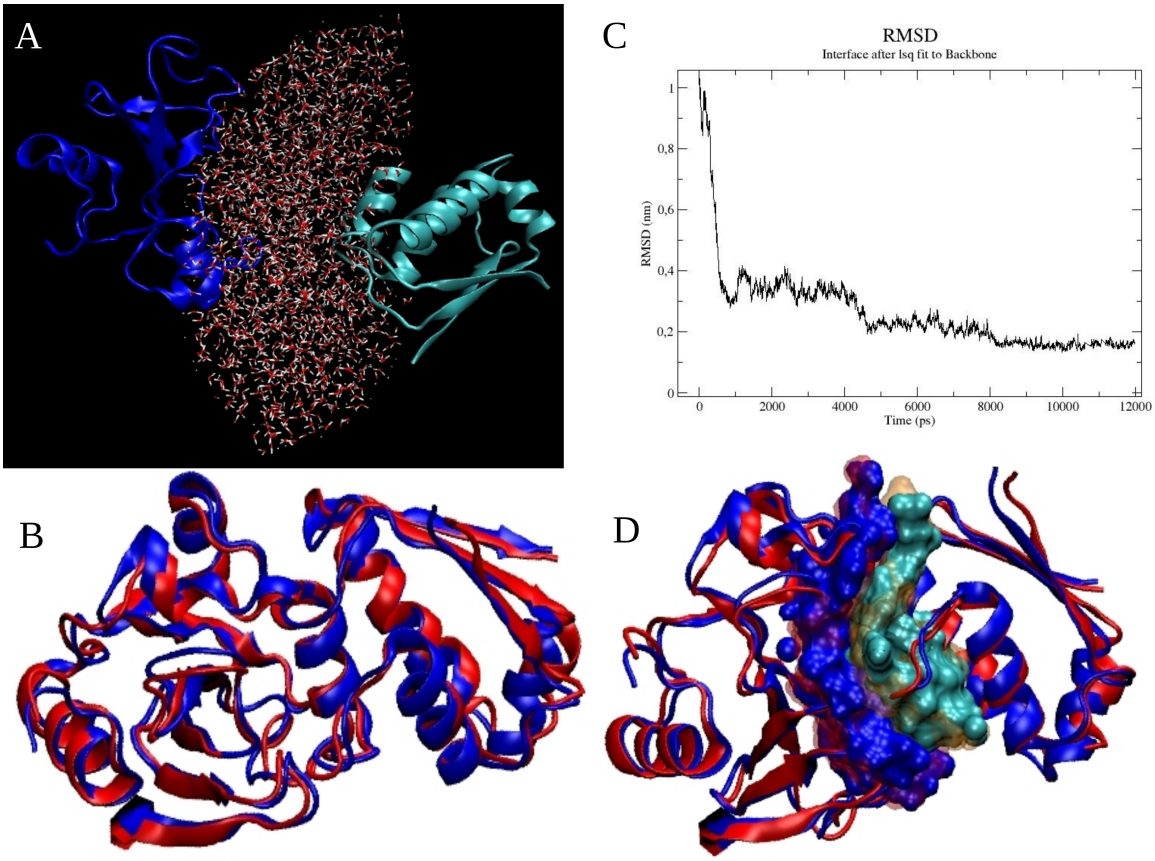


Figure 7.1: **Barnase-Barstar complex.** Panel A: Starting structure of the BE sumulation. A layer of explicit water is evidenced in between the two proteins. Panel B: in blue it is reported the lowest free energy structure obtained from the BE simulation and it is aligned with the equilibrated X-ray structure (PDB entry 1BRS) that is shown in red. Panel C: RMSD (nm) of the residues at the interface respect to the native complex (equilibrated using MD) as a function of the simultion time for one of the 6 walker used in BE that get close to the native complex. Panel D: the interface between the two proteins it is evidenced in the structrures of panel B, the same color code is used.

To gain accuracy in the free energy calculation several independent simulation were run starting from the structures obtained after 12 ns for a cumulative time of $\sim 2\mu$s. The next step will be constructing a kinetic model for the protein-protein interaction hoping it allows undertanding how the two proteins recognize each other, the role of the water and the contribution of specific residues to the binding mechanism.

CHAPTER

**EIGHT**

# List of publications

S. Piana, A. Laio, **F. Marinelli**, M. V. Troys, D. Bourry, C. Ampe, J. C. Martins. Predicting the Effect of a Point Mutation on a Protein Fold: The Villin and Advillin Headpieces and Their Pro62Ala Mutants. (2008) *J. Mol. Biol.*, 375, 460-470.

N. Todorova, **F. Marinelli**, S. Piana, I. Yarovsky. Exploring the Folding Free Energy Landscape of Insulin using Bias Exchange Metadynamics. (2009) *J. Phys. Chem. B*, 113, 3556-3564.

**F. Marinelli**, F. Pietrucci, A. Laio and S. Piana. A kinetic model of Trp-cage folding from multiple biased molecular dynamics simulations. (2009) *PLoS Comput Biol*, 5, e1000452.

F. Pietrucci, **F. Marinelli**, P. Carloni, and A. Laio. Substrate binding mechanism of HIV-1 protease from explicit-solvent atomistic simulations. *J. Am. Chem. Soc.* (2009) 131, 11811-11818.

M. Bonomi, D. Branduardi, G. Bussi, C. Camilloni, D. Provasi, P. Raiteri, D. Donadio, **F. Marinelli**, F. Pietrucci, R. A. Broglia, M. Parrinello. PLUMED: a portable plugin for free energy calculations with molecular dynamics. (2009) *Comp. Phys. Comm.* 180, 1961.

Y. Crespo, **F. Marinelli**, F. Pietrucci, A.Laio. Metadynamics convergence law in a two dimensional Ising model. (2009) *SUBMITTED*.

P. Cossio, **F. Marinelli**, A. Laio,F. Pietrucci. Optimizing the performance of Bias Exchange Metadynamics for Protein Folding. (2009) *SUBMITTED*.

# Bibliography


[1] Arthur M. Lesk. *Introduction to Bioinformatics*, Chapter 5, pages 216–276. Oxford University Press, Oxford, UK, First edition, 2002.

[2] I Halperin, BY Ma, H Wolfson, and R Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins*, 47(4):409–443, 2002.

[3] Nelly Andrusier, Efrat Mashiach, Ruth Nussinov, and Haim J. Wolfson. Principles of flexible protein-protein docking. *Proteins*, 73(2):271–289, 2008.

[4] John L. Klepeis, Kresten Lindorff-Larsen, Ron O. Dror, and David E. Shaw. Long-timescale molecular dynamics simulations of protein structure and function. *Curr. Opin. Struct. Biol.*, 19(2):120–127, APR 2009.

[5] A. D. Mackerell. Empirical force fields for biological macromolecules: Overview and issues. *J. Comput. Chem.*, 25(13):1584–1604, 2004.

[6] J. W. Ponder and D. A. Case. Force fields for protein simulations. *Adv. Protein Chem.*, 66:27+, 2003.

[7] W. L. Jorgensen and J. Tirado-Rives. Potential energy functions for atomic-level simulations of water and organic and biomolecular systems. *Proc. Natl. Acad. Sci. U. S. A.*, 102(19):6665–6670, 2005.

[8] C. A. F. De Oliveira, D. Hamelberg, and J. A. McCammon. Estimating kinetic rates from accelerated molecular dynamics simulations: Alanine dipeptide in explicit solvent as a case study. *J. Chem. Phys.*, 127(17):175105, 2007.

[9] E A Carter, G Ciccotti, J T Hynes, and R Kapral. Constrained reaction coordinate dynamics for the simulation of rare events. *Chem. Phys. Lett.*, 156:472–477, 1989.

[10] P A Bash, U C Singh, F K Brown, R Langridge, and P A Kollman. Free energy calculation by computer simulation. *Science*, 235:574–576, 1987.

[11] G. N. Patey and J. P. Valleau. Monte-carlo method for obtaining interionic potential of mean force in ionic solution. *J. Chem. Phys.*, 63:2334–2339, 1975.

[12] Helmut Grubmüller. Predicting slow structural transitions in macromolecular systems: conformational flooding. *Phys. Rev. E*, 52:2893–2906, 1995.

[13] S. Kumar, J. M. Rosenberg, D. Bouzida, R. H. Swendsen, and P. A. Kollman. Multidimensional free-energy calculations using the weighted histogram analysis method. *J. Comput. Chem.*, 16(11):1339–1350, 1995.

[14] C Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78:2690, 1997.

[15] E Darve and A Pohorille. Calculating free energies using average force. *J. Chem. Phys.*, 115:9169–9183, 2001.

[16] J Gullingsrud, R Braun, and K Schulten. Reconstructing potentials of mean force through time series analysis of steered molecular dynamics simulations. *J. Comp. Phys.*, 151:190–211, 1999.

[17] L Rosso, P Minary, Z.W Zhu, and M.E Tuckerman. On the use of the adiabatic molecular dynamics technique in the calculation of free energy profiles. *J. Comput. Phys.*, 116:4389–4402, 2002.

[18] R Elber and M Karplus. A method for determining reaction paths in large molecules - application to myoglobin. *Chem. Phys. Lett.*, 139(5):375–380, SEP 4 1987.

[19] E. Weinan, W. Q. Ren, and E. Vanden-Eijnden. Finite temperature string method for the study of rare events. *J. Phys. Chem. B*, 109:6688, 2005.

[20] C. Dellago, P. Bolhuis, F. S. Csajka, and D. Chandler. Transition path sampling and the calculation of rate constants. *J. Chem. Phys.*, 108:1964–1977, 1998.

[21] A. K. Faradjian and R. Elber. Computing time scales from reaction coordinates by milestoning. *J. Chem Phys.*, 120:10880–10889, 2004.

[22] R Fletcher and M J D Powell. A rapidly convergent descent method for minimization. *Comput. J.*, 6:163–168, 1963.

[23] G Henkelman and H Jonsson. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *J. Comput. Phys.*, 111(15):7010–7022, OCT 15 1999.

[24] AF Voter. Hyperdynamics: Accelerated molecular dynamics of infrequent events. *Phys. Rev. Lett.*, 78(20):3908–3911, MAY 19 1997.

[25] RA Miron and KA Fichthorn. Multiple-time scale accelerated molecular dynamics: Addressing the small-barrier problem. *Phys. Rev. Lett.*, 93(12), SEP 17 2004.

[26] GT Barkema and N Mousseau. Event-based relaxation of continuous disordered systems. *Phys. Rev. Lett.*, 77(21):4358–4361, NOV 18 1996.

[27] H Merlitz and W Wenzel. Comparison of stochastic optimization methods for receptor-ligand docking. *Chem. Phys. Lett.*, 362:271–277, 2002.

[28] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314:141, 1999.

[29] F Wang and D P Landau. Efficient, multiple-range random walk algorithm to calculate the density of states. *Phys. Rev. Lett.*, 86:2050, 2001.

[30] A. Laio and M. Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U. S. A.*, 99(20):12562–12566, 2002.

[31] Alessandro Laio and Francesco L. Gervasio. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. *Rep. Prog. Phys.*, 71, 2008.

[32] S. Kumar, P. W. Payne, and M. Vasquez. Method for free-energy calculations using iterative techinques. *J. Comp. Chem.*, 17:1269–1275, 1996.

[33] K. Hukushima and K. Nemoto. Exchange Monte Carlo method and application to spin glass simulations. *J. Phys. Soc. Jpn.*, 65:1604, 1996.

[34] D. A. C. Beck, G. W. N. White, and V. Daggett. Exploring the energy landscape of protein folding using replica-exchange and conventional molecular dynamics simulations. *J. Struct. Biol.*, 157(3):514–523, 2007.

[35] S Trebst, M Troyer, and UHE Hansmann. Optimized parallel tempering simulations of proteins. *J. Comput. Phys.*, 124(17):174903, MAY 7 2006.

[36] Yuji Sugita, Akio Kitao, and Yuko Okamoto. Multidimensional replica-exchange method for free-energy calculations. *J. Chem. Phys.*, 113(15):6042–6051, 2000.

[37] Jeremy Curuksu and Martin Zacharias. Enhanced conformational sampling of nucleic acids by a new Hamiltonian replica exchange molecular dynamics approach. *J. Chem. Phys.*, 130(10), MAR 14 2009.

[38] S. Piana and A. Laio. A bias-exchange approach to protein folding. *J. Phys. Chem. B*, 111(17):4553–4559, 2007.

[39] S. Piana, A. Laio, F. Marinelli, M. Van Troys, D. Bourry, C. Ampe, and J. C. Martins. Predicting the effect of a point mutation on a protein fold: The villin and advillin headpieces and their Pro62Ala mutants. *J. Mol. Biol.*, 375(2):460–470, 2008.

[40] Nevena Todorova, Fabrizio Marinelli, Stefano Piana, and Irene Yarovsky. Exploring the Folding Free Energy Landscape of Insulin Using Bias Exchange Metadynamics. *J. Phys. Chem. B*, 113:3556–3564, 2009.

[41] Vanessa Leone, Gianluca Lattanzi, Carla Molteni, and Paolo Carloni. Mechanism of action of cyclophilin a explored by metadynamics simulations. *PLoS Comput. Biol.*, 5:e1000309, 2009.

[42] J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *J. Chem. Phys.*, 126(15):155101, 2007.

[43] A. Fischer, S. Waldhausen, I. Horenko, E. Meerbach, and C. Schuette. Identification of Biomolecular conformations from incomplete torsion angle observations by hidden Markov models. *J. Comput. Chem.*, 28(15):2453–2464, 2007.

[44] G. Jayachandran, V. Vishal, and V. S. Pande. Using massively parallel simulation and Markovian models to study protein folding: Examining the dynamics of the villin headpiece. *J. Chem. Phys.*, 124(16):164902, 2006.

[45] I. Horenko, E. Dittmer, A. Fischer, and C. Schuette. Automated model reduction for complex systems exhibiting metastability. *Multiscale Model. Simul.*, 5(3):802–827, 2006.

[46] G. Hummer. Position-dependent diffusion coefficients and free energies from Bayesian analysis of equilibrium and replica molecular dynamics simulations. *New J. Phys.*, 7:34, 2005.

[47] N.-V. Buchete and G. Hummer. Coarse master equations for peptide folding dynamics. *J. Phys. Chem. B*, 112:6057, 2008.

[48] Fabrizio Marinelli, Fabio Pietrucci, Alessandro Laio, and Stefano Piana. A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations. *PLoS Comput Biol*, 5(8):e1000452, 08 2009.

[49] Michael Shirts and Vijay S. Pande. COMPUTING: Screen Savers of the World Unite! *Science*, 290(5498):1903–1904, 2000.

[50] D. J. Bicout and A. Szabo. Electron transfer reaction dynamics in non-Debye solvents. *J. Chem. Phys.*, 109(6):2325–2338, 1998.

[51] Stefan Auer, Mark A. Miller, Sergei V. Krivov, Christopher M. Dobson, Martin Karplus, and Michele Vendruscolo. Importance of metastable states in the free energy landscapes of polypeptide chains. *Phys. Rev. Lett.*, 99(17), 2007.

[52] B HAWKINS, K CROSS, and D CRAIK. SOLUTION STRUCTURE OF THE B-CHAIN OF INSULIN AS DETERMINED BY H-1-NMR SPECTROSCOPY - COMPARISON WITH THE CRYSTAL-STRUCTURE OF THE INSULIN HEXAMER AND WITH THE SOLUTION STRUCTURE OF THE INSULIN MONOMER. *Int. J. Pept. Protein Res.*, 46(5):424–433, 1995.

[53] ZS Qiao, CY Min, QX Hua, MA Weiss, and YM Feng. In vitro refolding of human proinsulin - Kinetic intermediates, putative disulfide-forming pathway, folding initiation site, and potential role of C-peptide in folding process. *J. Biol. Chem.*, 278(20):17800–17809, 2003.

[54] FY Dupradeau, T Richard, G Le Flem, H Oulyadi, Y Prigent, and JP Monti. A new B-chain mutant of insulin: comparison with the insulin crystal structure and role of sulfonate groups in the B-chain structure. *J. Pept. Res.*, 60(1):56–64, 2002.

[55] RA PULLEN, DG LINDSAY, SP WOOD, IJ TICKLE, TL BLUNDELL, A WOLLMER, G KRAIL, D BRANDENBURG, H ZAHN, J GLIEMANN, and S GAMMELTOFT. RECEPTOR-BINDING REGION OF INSULIN. *Nature*, 259(5542):369–373, 1976.

[56] RG MIRMIRA, SH NAKAGAWA, and HS TAGER. IMPORTANCE OF THE CHARACTER AND CONFIGURATION OF RESIDUES-B24, RESIDUES-B25, AND RESIDUES-B26 IN INSULIN-RECEPTOR INTERACTIONS. *J. Biol. Chem.*, 266(3):1428–1436, 1991.

[57] U DEREWENDA, Z DEREWENDA, EJ DODSON, GG DODSON, X BING, and J MARKUSSEN. X-RAY-ANALYSIS OF THE SINGLE CHAIN-B29-A1 PEPTIDE-LINKED INSULIN MOLECULE - A COMPLETELY INACTIVE ANALOG. *J. Mol. Biol.*, 220(2):425–433, 1991.

[58] Fabio Pietrucci, Fabrizio Marinelli, Paolo Carloni, and Alessandro Laio. Substrate Binding Mechanism of HIV-1 Protease from Explicit-Solvent Atomistic Simulations. *J. Am. Chem. Soc.*, 131:11811–11818, 2009.

[59] R. P. Feynman, R. B. Leighton, and M. Sands. *The Feynman Lectures on Physics, Vol. 1, Chapter 9 ("Newton's Laws of Dynamics").* Addison-Wesley, 1963.

[60] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. jr Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, and P.A. Kollman. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[61] W. F. van Gunsteren, S. R. Billeter, A. A. Eising, P. H. Hünenberger, P. Krüger, A. E. Mark, W. R. P. Scott, and I. G. Tironi. *Biomolecular*

*Simulation: The GROMOS96 Manual and User Guide*. Hochschulverlag AG an der ETH Zürich, Zürich, 1996.

[62] A.D. Jr. MacKerell, B. Brooks, C. L. III Brooks, L. Nilsson, B. Roux, Y. Won, and M. Karplus. *The Encyclopedia of Computational Chemistry. 1. CHARMM: The Energy Function and Its Parameterization with an Overview of the Program*. Chichester: John Wiley Sons., 1998.

[63] W.L. Jorgensen and J. Tirado-Rives. The OPLS Force Field for Proteins. Energy Minimizations for Crystals of Cyclic Peptides and Crambin. *J. Am. Chem. Soc.*, 110:1657–1666, 1988.

[64] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. M. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, J. Caldwell, J. M. Wang, and P. Kollman. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *J. Comput. Chem.*, 24(16):1999–2012, 2003.

[65] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.

[66] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints; Molecular dynamics of n-alkanes. *J. Comp. Phys*, 23:327, 1977.

[67] B. Hess, H. Bekker, H. J. C. Berendsen, and G. E. M. J. Fraaije. Lincs: A linear constraint solver for molecular simulations. *J. Comput. Chem.*, 18:1463, 1997.

[68] S. Miyamoto and P. A. Kollman. An analytical version of the SHAKE and RATTLE algorithms for rigid water models. *J. Comput. Chem.*, 13:952–962, 1992.

[69] T. A. Darden and D. York. Particle mesh ewald - an n.log(n) method for ewald sums in large systems. *J. Chem. Phys.*, 98:10089, 1993.

[70] S. Nose. A molecular-dynamics method for simulations in the canonical ensemble. *Mol. Phys.*, 52:255, 1984.

[71] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gusteren, A. Di Nola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684, 1984.

[72] M Sprik and G Ciccotti. Free energy from constrained molecular dynamics. *J. Chem. Phys.*, 109:7737–7744, 1998.

[73] A.M Ferrenberg and R.W Swendsen. Optimized monte-carlo data-analysis. *Phys. Rev. Lett.*, 61:2635, 1988.

[74] B. Roux. The calculation of the potential of mean force using computer-simulations. *Comput. Phys. Comm.*, 91:275–282, 1995.

[75] GE Crooks. Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *J. Stat. Phys.*, 90(5-6):1481–1487, MAR 1998.

[76] D Rodriguez-Gomez, E Darve, and A Pohorille. Assessing the efficiency of free energy calculation methods. *J. Chem. Phys.*, 120:3563–3578, 2004.

[77] G Henkelman and H. Jnsson. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *J. Chem. Phys.*, 113:9978–9985, 2000.

[78] Luca Maragliano, Alexander Fischer, Eric Vanden-Eijnden, and Giovanni Ciccotti. String method in collective variables: Minimum free energy paths and isocommittor surfaces. *J. Comput. Phys.*, 125(2):024106, JUL 14 2006.

[79] C. Dellago, P.G. Bolhuis, and P.L. Geissler. Transition path sampling. *Adv. Chem. Phys.*, 123:1–78, 2002.

[80] Baron Peters and Bernhardt L. Trout. Obtaining reaction coordinates by likelihood maximization. *J. Comput. Phys.*, 125(5):054108, AUG 7 2006.

[81] T. van Erp, D. Moroni, and P. G. Bolhuis. A novel path sampling method for the calculation of rate constants. *J. Chem. Phys.*, 118:7762, 2003.

[82] RJ Allen, PB Warren, and PR ten Wolde. Sampling rare switching events in biochemical networks. *Phys. Rev. Lett.*, 94(1):018104, JAN 14 2005.

[83] J. Juraszek and P. G. Bolhuis. Sampling the multiple folding mechanisms of Trp-cage in explicit solvent. *Proc. Natl. Acad. Sci. U. S. A.*, 103(43):15859–15864, 2006.

[84] A Kidera N Nakajima, J Higo and H Nakamura. Flexible docking of a ligand peptide to a receptor protein by multicanonical molecular dynamics simulation. *Chem. Phys. Lett.*, 278:297–301, 1997.

[85] D Cvijovic and J Klinowski. Taboo search - an approach to the multiple minima problem. *Science*, 267:664–666, 1995.

[86] T Huber, A.E Torda, and W.F van Gunsteren. Local elevation: a method for improving the searching properties of molecular dynamics simulation. *J. Comput. Aided Mol. Des.*, 8:695–708, 1994.

[87] G Bussi, A Laio, and M Parrinello. Equilibrium free energies from nonequilibrium metadynamics. *Phys. Rev. Lett.*, 96, 2006.

[88] A Laio, A Rodriguez-Fortea, F L Gervasio, M Ceccarelli, and M Parrinello. Assessing the accuracy of metadynamics. *J. Phys. Chem. B.*, 109:6714–6721, 2005.

[89] R Zwanzig. Memory effects in irreversible thermodynamics. *Phys. Rev.*, 124:983–992, 1961.

[90] H.C Ottinger. General projection operator formalism for the dynamics and thermodynamics of complex fluids. *Phys. Rev. E*, 57:1416, 1998.

[91] C.W Gardiner. *Handbook of Stochastic Methods*. Springer, 2004.

[92] X. Daura, K. Gademann, B. Jaun, D. Seebach, W. F. van Gunsteren, and a. E. Mark. Peptide folding: When simulation meets experiment. *Angew. Chem.-Int. Edit.*, 38(1-2):236–240, 1999.

[93] C. Micheletti, A. Laio, and M. Parrinello. Reconstructing the density of states by history-dependent metadynamics. *Phys. Rev. Lett.*, 92:170601, 2004.

[94] Eric Vanden-Eijnden and Fabio A. Tal. Transition state theory: Variational formulation, dynamical corrections, and error estimates. *J Chem Phys*, 123(18):184103, 2005.

[95] E. Lindahl, b. Hess, and D. van der Spoel. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.*, 7(8):306–317, 2001.

[96] H. J. C. Berendsen, D. Van der Spoel, and R. Vandrunen. GROMACS - a message-passing parallel molecular-dynamics implementation. *Comput. Phys. Commun.*, 91(1-3):43–56, 1995.

[97] U. Essman, L. Perera, M. L. Berkowitz, T. A. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh ewald method. *J. Chem. Phys.*, 103:8577, 1995.

[98] V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg, and C. Simmerling. Comparison of multiple amber force fields and development of improved protein backbone parameters. *Proteins*, 65(3):712–725, 2006.

[99] J. Graf, P. H. Nguyen, G. Stock, and h. Schwalbe. Structure and dynamics of the homologous series of alanine peptides: A joint molecular dynamics/NMR study. *J. Am. Chem. Soc.*, 129(5):1179–1189, 2007.

[100] Y. G. Mu, D. S. Kosov, and G. Stock. Conformational dynamics of trialanine in water. 2. Comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments. *J. Phys. Chem. B*, 107(21):5064–5073, 2003.

[101] S. Woutersen and P. Hamm. Structure determination of trialanine in water using polarization sensitive two-dimensional vibrational spectroscopy. *J. Phys. Chem. B*, 104(47):11316–11320, 2000.

[102] R. Schweitzer-Stenner, F. Eker, Q. Huang, and K. Griebenow. Dihedral angles of trialanine in D2O determined by combining FTIR and polarized visible Raman spectroscopy. *J. Am. Chem. Soc.*, 123(39):9628–9633, 2001.

[103] R. Schweitzer-Stenner. Dihedral angles of tripeptides in solution directly determined by polarized Raman and FTIR spectroscopy. *Biophys. J.*, 83(1):523–532, 2002.

[104] J. W. Neidigh, R. M. Fesinmeyer, and N. H. Andersen. Designing a 20-residue protein. *Nat. Struct. Biol.*, 9(6):425–430, 2002.

[105] L. L. Qiu, S. A. Pabit, A. E. Roitberg, and S. J. Hagen. Smaller and faster: The 20-residue Trp-cage protein folds in 4 $\mu$s. *J. Am. Chem. Soc.*, 124(44):12952–12953, 2002.

[106] W. W. Streicher and G. I. Makhatadze. Unfolding thermodynamics of Trp-cage, a 20 residue miniprotein, studied by differential scanning calorimetry and circular dichroism spectroscopy. *Biochemistry*, 46(10):2876–2880, 2007.

[107] Z. Ahmed, I. S. Beta, A. V. Mikhonin, and S. A. Asher. UV-resonance Raman thermal unfolding study of Trp-cage shows that it is not a simple two-state miniprotein. *J. Am. Chem. Soc.*, 127(31):10943–10950, 2005.

[108] K. Hun Mok, Lars T. Kuhn, Martin Goez, Iain J. Day, Jasper C. Lin, Niels H. Andersen, and P. J. Hore. A pre-existing hydrophobic collapse in the unfolded state of an ultrafast folding protein. *Nature*, 447:106–109, 2007.

[109] H. Neuweiler, S. Doose, and M. Sauer. A microscopic view of miniprotein folding: Enhanced folding efficiency through formation of an intermediate. *Proc. Natl. Acad. Sci. U. S. A.*, 102(46):16650–16655, 2005.

[110] C. Simmerling, B. Strockbine, and A. E. Roitberg. All-atom structure prediction and folding simulations of a stable protein. *J. Am. Chem. Soc.*, 124(38):11258–11259, 2002.

[111] S. Chowdhury, M. C. Lee, G. M. Xiong, and Y. Duan. Ab initio folding simulation of the Trp-cage mini-protein approaches NMR resolution. *J. Mol. Biol.*, 327(3):711–717, 2003.

[112] A. Schug, T. Herges, a. Verma, K. H. Lee, and W. Wenzel. Comparison of Stochastic optimization methods for all-atom folding of the Trp-cage protein. *ChemPhysChem*, 6(12):2640–2646, 2005.

[113] A Schug, W Wenzel, and UHE Hansmann. Energy landscape paving simulations of the trp-cage protein. *J. Chem. Phys.*, 122(19):194711, MAY 15 2005.

[114] A Schug, T Herges, and W Wenzel. Reproducible protein folding with the stochastic tunneling method. *Phys. Rev. Lett.*, 91(15):158102–158102, Oct 2003.

[115] M. Ota, M. Ikeguchi, and A. Kidera. Phylogeny of protein-folding trajectories reveals a unique pathway to native structure. *Proc. Natl. Acad. Sci. U. S. A.*, 101(51):17658–17663, 2004.

[116] J. W. Pitera and W. Swope. Understanding folding and design: Replica-exchange simulations of "Trp-cage" fly miniproteins. *Proc. Natl. Acad. Sci. U. S. A.*, 100(13):7587–7592, 2003.

[117] B. Zagrovic and V. Pande. Solvent viscosity dependence of the folding rate of a small protein: Distributed computing study. *J. Comput. Chem.*, 24(12):1432–1436, 2003.

[118] R. H. Zhou. Trp-cage: Folding free energy landscape in explicit water. *Proc. Natl. Acad. Sci. U. S. A.*, 100(23):13280–13285, 2003.

[119] C. D. Snow, B. Zagrovic, and V. S. Pande. The Trp cage: Folding kinetics and unfolded state topology via molecular dynamics simulations. *J. Am. Chem. Soc.*, 124(49):14548–14549, 2002.

[120] Alex Kentsis, Tatyana Gindin, Mihaly Mezei, and Roman Osman. Calculation of the free energy and cooperativity of protein folding. *PLoS ONE*, 2:e446, May 2007.

[121] D. Paschek, H. Nymeyer, and A. E. Garcia. Replica exchange simulation of reversible folding/unfolding of the Trp-cage miniprotein in explicit solvent: On the structure and possible role of internal water. *J. Struct. Biol.*, 157(3):524–533, 2007.

[122] Jarek Juraszek and Peter G. Bolhuis. Rate Constant and Reaction Coordinate of Trp-Cage Folding in Explicit Water. *Biophys. J.*, 95(9):4246–4257, 2008.

[123] A. B. Bortz, M. H. Kalos, and J. L. Lebowitz. New algorithm for monte-carlo simulation of ising spin systems. *J. Comput. Phys.*, 17(1):10–18, 1975.

[124] A. F. Voter. *Introduction to the Kinetic Monte Carlo Method.* In Radiation Effects in Solids, Sickafus, K. E., Kotomin, E. A., Eds.; Springer. NATO Publishing Unit: Dordrecht, The Netherlands, 2005.

[125] A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30(7):1575–1584, 2002.

[126] D. Gfeller, P. De Los Rios, A. Caflisch, and F. Rao. Complex network analysis of free-energy landscapes. *Proc. Natl. Acad. Sci. U. S. A.*, 104(6):1817–1822, 2007.

[127] J. D. Hirst and C. L. Brooks III. Helicity, circular dichroism and molecular dynamics of proteins. *J. Mol. Biol.*, 243:173, 1994.

[128] AC Wallace, RA Laskowski, and JM Thornton. LIGPLOT - A Program to generate schematic diagrams of protein ligand interactions. *Protein Eng.*, 8:127–134, FEB 1995.

[129] Paul Robustelli, Andrea Cavalli, Christopher M. Dobson, Michele Vendruscolo, and Xavier Salvatella. Folding of Small Proteins by Monte Carlo Simulations with Chemical Shift Restraints without the Use of Molecular Fragment Replacement or Structural Homology. *J. Phys. Chem. B*, 113(22):7890–7896, 2009.

[130] X. Xu, S. Moon, and D.A. Case. *SHIFTS Program.* Department Molecular Biology, The Scripps Research Institute, 2005.

[131] H. Roder, K. Maki, and H. Cheng. Early events in protein folding explored by rapid mixing methods. *Chem. Rev.*, 106(5):1836–1861, 2006.

[132] D Eisenberg and AD Mclachlan. Solvation energy in protein folding and binding. *Nature*, 319(6050):199–203, JAN 16 1986.

[133] YM Rhee, EJ Sorin, G Jayachandran, E Lindahl, and VS Pande. Simulations of the role of water in the protein-folding mechanism. *Proc. Natl. Acad. Sci. U. S. A.*, 101:6456–6461, 2004.

[134] MY Shen and KF Freed. Long time dynamics of met-enkephalin: Comparison of explicit and implicit solvent models. *Biophys. J.*, 82:1791–1808, 2002.

[135] MR Bunagan, X Yang, JG Saven, and F Gai. Ultrafast folding of a computationally designed Trp-cage mutant: Trp(2)-cage. *J. Phys. Chem. B*, 110:3759–3763, 2006.

[136] Bipasha Barua, Jasper C. Lin, Victoria D. Williams, Phillip Kummler, Jonathan W. Neidigh, and Niels H. Andersen. The Trp-cage: optimizing the stability of a globular miniprotein. *Protein Eng. Des. Sel.*, 21(3):171–185, MAR 2008.

[137] D. F. Steiner, G. I. Bell, H. S. Tager, and A. H. Rubenstein. *Chemistry and biosynthesis of the islet hormones. In Endocrinology.* In Radiation Effects in Solids, WB Saunders: London, p 1296, 1995.

[138] Guy Dodson and Don Steiner. The role of assembly in insulin's biosynthesis. *Curr. Opin. Struct. Biol.*, 8(2):189 – 194, 1998.

[139] Robert Z.-T. Luo, Daniel R. Beniac, Allan Fernandes, Cecil C. Yip, and F. P. Ottensmeyer. Quaternary Structure of the Insulin-Insulin Receptor Complex. *Science*, 285(5430):1077–1080, 1999.

[140] GD SMITH and GG DODSON. THE STRUCTURE OF A RHOMBOHEDRAL R6 INSULIN HEXAMER THAT BINDS PHENOL. *Biopolymers*, 32(4):441–445, 1992.

[141] E CISZAK and GD SMITH. CRYSTALLOGRAPHIC EVIDENCE FOR DUAL COORDINATION AROUND ZINC IN THE T(3)R(3) HUMAN INSULIN HEXAMER. *Biochemistry*, 33(6):1512–1517, 1994.

[142] U DEREWENDA, Z DEREWENDA, EJ DODSON, GG DODSON, CD REYNOLDS, GD SMITH, C SPARKS, and D SWENSON. PHENOL STABILIZES MORE HELIX IN A NEW SYMMETRICAL ZINC INSULIN HEXAMER. *Nature*, 338(6216):594–596, 1989.

[143] EN BAKER, TL BLUNDELL, JF CUTFIELD, SM CUTFIELD, EJ DODSON, GG DODSON, DMC HODGKIN, RE HUBBARD, NW ISAACS, CD REYNOLDS, K SAKABE, N SAKABE, and NM VIJAYAN. THE STRUCTURE OF 2ZN PIG INSULIN CRYSTALS AT 1.5-A RESOLUTION. *Philos. Trans. R. Soc. Lond. Ser. B-Biol. Sci.*, 319(1195):369–&, JUL 6 1988.

[144] GD SMITH, DC SWENSON, EJ DODSON, GG DODSON, and CD REYNOLDS. STRUCTURAL STABILITY IN THE 4-ZINC HUMAN INSULIN HEXAMER. *Proc. Natl. Acad. Sci. USA.*, 81(22):7093–7097, 1984.

[145] NC KAARSHOLM, HC KO, and MF DUNN. COMPARISON OF SOLUTION STRUCTURAL FLEXIBILITY AND ZINC-BINDING DOMAINS FOR INSULIN, PROINSULIN, AND MINIPROINSULIN. *Biochemistry*, 28(10):4427–4435, 1989.

[146] ZP Yao, ZH Zeng, HM Li, Y Zhang, YM Feng, and DC Wang. Structure of an insulin dimer in an orthorhombic crystal: the structure analysis of a human insulin mutant (B9 Ser -¿ Glu). *Acta Crystallogr. Sect. D-Biol. Crystallogr.*, 55(Part 9):1524–1532, 1999.

[147] QX HUA and MA WEISS. COMPARATIVE 2D-NMR STUDIES OF HUMAN INSULIN AND DES-PENTAPEPTIDE INSULIN - SEQUENTIAL RESONANCE ASSIGNMENT AND IMPLICATIONS FOR PROTEIN DYNAMICS AND RECEPTOR RECOGNITION. *Biochemistry*, 30(22):5505–5515, JUN 4 1991.

[148] MJ ADAMS, TL BLUNDELL, EJ DODSON, GG DODSON, M VIJAYAN, EN BAKER, MM HARDING, DC HODGKIN, B RIMMER, and S SHEAT. STRUCTURE OF RHOMBOHEDRAL 2 ZINC INSULIN CRYSTALS. *Nature*, 224(5218):491–&, 1969.

[149] Insulin Structure Group. *Peking Rev.*, 40:11–16, 1971.

[150] G BENTLEY, E DODSON, G DODSON, D HODGKIN, and D MERCOLA. STRUCTURE OF INSULIN IN 4-ZINC INSULIN. *Nature*, 261(5556):166–168, 1976.

[151] GA BENTLEY, J BRANGE, Z DEREWENDA, EJ DODSON, GG DODSON, J MARKUSSEN, AJ WILKINSON, A WOLLMER, and B XIAO. ROLE OF B13 GLU IN INSULIN ASSEMBLY - THE HEXAMER STRUCTURE OF RECOMBINANT MUTANT (B13 GLU-]GLN) INSULIN. *J. Mol. Biol.*, 228(4):1163–1176, 1992.

[152] A Budi, S Legge, H Treutlein, and I Yarovsky. Effect of external stresses on protein conformation: a computer modelling study. *Eur. Biophys. J. Biophys. Lett.*, 33(2):121–129, 2004.

[153] Akin Budi, F. Sue Legge, Herbert Treutlein, and Irene Yarovsky. Comparative study of insulin chain-b in isolated and monomeric environments under external stress. *J. Phys. Chem. B*, 112(26):7916–7924, 2008.

[154] I PITTMAN and HS TAGER. A SPECTROSCOPIC INVESTIGATION OF THE CONFORMATIONAL DYNAMICS OF INSULIN IN SOLUTION. *Biochemistry*, 34(33):10578–10590, 1995.

[155] U DEREWENDA, Z DEREWENDA, GG DODSON, RE HUBBARD, and F KORBER. MOLECULAR-STRUCTURE OF INSULIN - THE INSULIN MONOMER AND ITS ASSEMBLY. *Br. Med. Bull.*, 45(1):4–18, 1989.

[156] S Ludvigsen, HB Olsen, and NC Kaarsholm. A structural switch in a mutant insulin exposes key residues for receptor binding. *J. Mol. Biol.*, 279(1):1–7, 1998.

[157] YS Zhang, JL Whittingham, JP Turkenburg, EJ Dodson, J Brange, and GG Dodson. Crystallization and preliminary crystallographic investigation of a low-pH native insulin monomer with flexible behaviour. *Acta Crystallogr. Sect. D-Biol. Crystallogr.*, 58(Part 1):186–187, 2002.

[158] M Falconi, MT Cambria, A Cambria, and A Desideri. Structure and stability of the insulin dimer investigated by molecular dynamics simulation. *J. Biomol. Struct. Dyn.*, 18(5):761–772, 2001.

[159] V Zoete, M Meuwly, and M Karplus. A comparison of the dynamic behavior of monomeric and dimeric insulin shows structural rearrangements in the active monomer. *J. Mol. Biol.*, 342(3):913–929, 2004.

[160] A Budi, FS Legge, H Treutlein, and I Yarovsky. Electric field effects on insulin chain-B conformation. *J. Phys. Chem. B*, 109(47):22641–22648, 2005.

[161] Akin Budi, F. Sue Legge, Herbert Treutlein, and Irene Yarovsky. Effect of frequency on insulin response to electric field stress. *J. Phys. Chem. B*, 111(20):5748–5756, 2007.

[162] FS Legge, A Budi, H Treutlein, and I Yarovsky. Protein flexibility: Multiple molecular dynamics simulations of insulin chain B. *Biophys. Chem.*, 119(2):146–157, 2006.

[163] D Van der Spoel, E Lindahl, B Hess, G Groenhof, AE Mark, and HJC Berendsen. GROMACS: Fast, flexible, and free. *J. Comput. Chem.*, 26(16):1701–1718, 2005.

[164] Nevena Todorova, F. Sue Legge, Herbert Treutlein, and Irene Yarovsky. Systematic comparison of empirical forcefields for molecular dynamic simulation of insulin. *J. Phys. Chem. B*, 112(35):11137–11146, 2008.

[165] TE CHEATHAM, JL MILLER, T FOX, TA DARDEN, and PA KOLLMAN. MOLECULAR-DYNAMICS SIMULATIONS ON SOLVATED BIOMOLECULAR SYSTEMS - THE PARTICLE MESH EWALD METHOD LEADS TO STABLE TRAJECTORIES OF DNA, RNA, AND PROTEINS. *J. Am. Chem. Soc.*, 117(14):4193–4194, 1995.

[166] C Dominguez, R Boelens, and AMJJ Bonvin. HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, 125(7):1731–1737, 2003.

[167] Sjoerd J. De Vries, Aalt D. J. van Dijk, Mickael Krzeminski, Mark van Dijk, Aurelien Thureau, Victor Hsu, Tsjerk Wassenaar, and Alexandre M. J. J. Bonvin. HADDOCK versus HADDOCK: New features and performance of HADDOCK2.0 on the CAPRI targets. *Proteins*, 69(4):726–733, 2007.

[168] JP Linge, MA Williams, CAEM Spronk, AMJJ Bonvin, and M Nilges. Refinement of protein structures in explicit solvent. *Proteins*, 50(3):496–506, 2003.

[169] QX HUA, M KOCHOYAN, and MA WEISS. STRUCTURE AND DYNAMICS OF DES-PENTAPEPTIDE-INSULIN IN SOLUTION - THE MOLTEN-GLOBULE HYPOTHESIS. *Proc. Natl. Acad. Sci. U. S. A.*, 89(6):2379–2383, 1992.

[170] QX HUA, JE LADBURY, and MA WEISS. DYNAMICS OF A MONOMERIC INSULIN ANALOG - TESTING THE MOLTEN-GLOBULE HYPOTHESIS. *Biochemistry*, 32(6):1433–1442, FEB 16 1993.

[171] QX Hua, WH Jia, BH Frank, NFB Phillips, and MA Weiss. A protein caught in a kinetic trap: Structures and stabilities of insulin disulfide isomers. *Biochemistry*, 41(50):14700–14715, 2002.

[172] HB Olsen, S Ludvigsen, and NC Kaarsholm. Solution structure of an engineered insulin monomer at neutral pH. *Biochemistry*, 35(27):8836–8845, 1996.

[173] QX HUA, SE SHOELSON, M KOCHOYAN, and MA WEISS. RECEPTOR-BINDING REDEFINED BY A STRUCTURAL SWITCH IN A MUTANT HUMAN INSULIN. *Nature*, 354(6350):238–241, 1991.

[174] Satoe H. Nakagawa, Qing-xin Hua, Shi-Quan Hu, Wenhua Jia, Shuhua Wang, Panayotis G. Katsoyannis, and Michael A. Weiss. Chiral mutagenesis of insulin - Contribution of the B20-B23 beta-turn to activity and stability. *J. Biol. Chem.*, 281(31):22386–22396, AUG 4 2006.

[175] SE SHOELSON, ZX LU, L PARLAUTAN, CS LYNCH, and MA WEISS. MUTATIONS AT THE DIMER, HEXAMER, AND RECEPTOR-BINDING SURFACES OF INSULIN INDEPENDENTLY AFFECT INSULIN INSULIN AND INSULIN-RECEPTOR INTERACTIONS. *Biochemistry*, 31(6):1757–1767, 1992.

[176] Tom Blundell, Guy Dodson, Dorothy Hodgkin, and Dan Mercola. Insulin: The structure in the crystal and its reflection in chemistry and biology by. volume 26 of *Advances in Protein Chemistry*, pages 279 – 286, 286a, 287–402. Academic Press, 1972.

[177] SH NAKAGAWA and HS TAGER. ROLE OF THE PHENYLALANINE-B25 SIDE-CHAIN IN DIRECTING INSULIN INTERACTION WITH ITS RECEPTOR - STERIC AND CONFORMATIONAL EFFECTS. *J. Biol. Chem.*, 261(16):7332–7341, 1986.

[178] SH NAKAGAWA and HS TAGER. ROLE OF THE COOH-TERMINAL B-CHAIN DOMAIN IN INSULIN-RECEPTOR INTERACTIONS - IDENTIFICATION OF PERTURBATIONS INVOLVING THE INSULIN MAIN CHAIN. *J. Biol. Chem.*, 262(25):12054–12058, 1987.

[179] S. C. Pettit, M. D. Moody, R. S. Wehbie, A. H. Kaplan, P. V. Nantermet, C. A. Klein, and R. Swanstrom. The p2 domain of human immunodeficiency virus type 1 Gag regulates sequential proteolytic processing and is required to produce fully infectious virions. *J. Virol.*, 68:8017–8027, 1994.

[180] A. Wlodaver and J. Vondrasek. Inhibitors of hiv-1 protease: a major success of structure-assisted drug design. *Annu. Rev. Biophys. Biomol. Struct.*, 27:249, 1998.

[181] H. Ohtaka and E. Freire. Adaptive inhibitors of the hiv-1 protease. *Prog. Biophys. Mol. Biol.*, 88:193, 2005.

[182] R. Ishima, D. I. Freedberg, Y. X. Wang, J. M. Louis, and D. A. Torchia. Flap opening and dimer-interface flexibility in the free and inhibitor-bound

hiv protease, and their implications for function. *Struct. Fold. Des.*, 7:1047, 1999.

[183] R. Ishima and J. M. Louis. A diverse view of protein dynamics from nmr studies of hiv-1 protease flaps. *Proteins*, 70:1408, 2008.

[184] E. S. Furfine, E. Dsouza, K. J. Ingold, J. J. Leban, T. Spector, and D. J. T. Porter. 2-step binding mechanism for hiv protease inhibitors. *Biochemistry*, 31:7886, 1992.

[185] E. J. Rodriguez, C. Debouck, I. C. Deckman, H. Abusoud, F. M. Raushel, and T. D. Meek. Inhibitor binding to the phe53trp mutant of hiv-1 protease promotes conformational-changes detectable by spectrofluorometry. *Biochemistry*, 32:3557, 1993.

[186] C. Y. Lee, P. K. Yang, W. S. Tzou, and M. J. Hwang. Estimates of relative binding free energies for hiv protease inhibitors using different levels of approximations. *Protein Eng.*, 11:429, 1998.

[187] M. L. Verdonk, G. Chessari, J. C. Cole, M. J. Hartshorn, C. W. Murray, J. W. M. Nissink, R. D. Taylor, and R. Taylor. Modeling water molecules in protein-ligand docking using gold. *J. Med. Chem.*, 48:6504, 2005.

[188] C. Bartels, A. Widmer, and C. Ehrhardt. Absolute free energies of binding of peptide analogs to the hiv-1 protease from molecular dynamics simulations. *J. Comput. Chem.*, 26:1294, 2005.

[189] I. Stoica, S. K. Sadiq, and P. V. Coveney. Rapid and accurate prediction of binding free energies for saquinavir-bound hiv-1 proteases. *J. Am. Chem. Soc.*, 130:2639, 2008.

[190] Z. Li and T. Lazaridis. Water at biomolecular binding interfaces. *Phys. Chem. Chem. Phys.*, 9:573, 2007.

[191] MK Gilson, JA Given, BL Bush, and JA McCammon. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys. J.*, 72:1047–1069, 1997.

[192] KL Meagher and HA Carlson. Solvation influences flap collapse in hiv-1 protease. *Proteins*, 58:119, 2005.

[193] V Hornak, A Okur, RC Rizzo, and C Simmerling. Hiv-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proc. Natl. Acad. Sci. U. S. A.*, 103:915, 2006.

[194] A. K. Sadiq, S. Wan, and P. V. Coveney. Insights into a mutation-assisted lateral drug escape mechanism from the HIV-1 protease active site. *Biochemistry*, 46:14865–14877, 2007.

[195] C. E. Chang, J. Trylska, V. Tozzini, and J. A. McCammon. Binding pathways of ligands to hiv-1 protease: Coarse-grained and atomistic simulations. *Chem. Biol. Drug Des.*, 69:5–13, 2007.

[196] J. Trylska, V. Tozzini, C. E. Chang, and J. A. McCammon. Hiv-1 protease substrate binding and product release pathways explored with coarse-grained molecular dynamics. *Biophys. J.*, 92:4179–4187, 2007.

[197] V. Hornak, A. Okur, R. C. Rizzo, and C. Simmerling. Hiv-1 protease flaps spontaneously close to the correct structure in simulations following manual placement of an inhibitor into the open state. *J. Am. Chem. Soc.*, 128:2812–2813, 2006.

[198] S. Piana, P. Carloni, and M. Parrinello. Role of conformational fluctuations in the enzymatic reaction of hiv-1 protease. *J. Mol. Biol.*, 319:567, 2002.

[199] W. G. Hoover. Canonical dynamics - equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695, 1985.

[200] L. J. Hyland, T. A. Tomaszek, G. D. Roberts, S. A. Carr, V. W. Magaard, H. L. Bryan, S. A. Fakhoury, M. L. Moore, M. D. Minnich, J. S. Culp, R. L. Desjarlais, and T. D. Meek. Human immunodeficiency virus-1 protease .1. initial velocity studies and kinetic characterization of reaction intermediates by o-18 isotope exchange. *Biochemistry*, 30:8441, 1991.

[201] S Piana, D Sebastiani, P Carloni, and M Parrinello. Ab initio molecular dynamics-based assignment of the protonation state of pepstatin a/hiv-1 protease cleavage site. *J. Am. Chem. Soc.*, 123:8730, 2001.

[202] M Prabu-Jeyabalan, E Nalivaika, and CA Schiffer. Substrate shape determines specificity of recognition for HIV-1 protease: Analysis of crystal structures of six substrate complexes. *Structure*, 10:369–381, 2002.

[203] M. W. Mahoney and W. L.Jorgensen. Diffusion constant of the tip5p model of liquid water. *J. Chem. Phys.*, 114:363, 2001.

[204] N. A. Baker, D. Sept, S. Joseph, M. J. Holst, and J. A. McCammon. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. U.S.A.*, 98:10037–10041, 2001.

[205] M Prabu-Jeyabalan, EA Nalivaika, K Romano, and CA Schiffer. Mechanism of substrate recognition by drug-resistant human immunodeficiency virus type 1 protease variants revealed by a novel structural intermediate. *J. Virol.*, 80:3607, 2006.

[206] G. Verkhivker, K. Appelt, S. T. Freer, and J. E. Villafranca. Empirical free-energy calculations of ligand-protein crystallographic complexes .1.

knowledge-based ligand-protein interaction potentials applied to the prediction of human-immunodeficiency-virus-1 protease binding-affinity. *Protein Eng.*, 8:677, 1995.

[207] Oechem, version 1.3.4, openeye scientific software, inc., santa fe, nm, usa, www.eyesopen.com, 2005.

[208] B. Maschera, G. Darby, G. Palu, L. L. Wright, M. Tisdale, R. Myers, E. D. Blair, and E. S. Furfine. Human immunodeficiency virus - mutations in the viral protease that confer resistance to saquinavir increase the dissociation rate constant of the protease-saquinavir complex. *J. Biol. Chem.*, 271:33231, 1996.

[209] H. B. Schock, V. M. Garsky, and L. C. Kuo. Mutational anatomy of an hiv-1 protease variant conferring cross-resistance to protease inhibitors in clinical trials - compensatory modulations of binding and activity. *J. Biol. Chem.*, 271:31957, 1996.

[210] C. F. Shuman, P. O. Markgren, M. Hamalainen, and U. H. Danielson. Elucidation of hiv-1 protease resistance by characterization of interaction kinetics between inhibitors and enzyme variants. *Antiviral Res.*, 58:235, 2003.

[211] D. Hamelberg and J. A. McCammon. Standard free energy of releasing a localized water molecule from the binding pockets of proteins: Double-decoupling method. *J. Am. Chem. Soc.*, 126:7683, 2004.

[212] E. C. B. Johnson, E. Malito, Y. Shen, B. Pentelute, D. Rich, J. Florian, W. Tang, and S. B. H. Kent. Insights from atomic-resolution x-ray structures of chemically synthesized hiv-1 protease in complex with inhibitors. *J. Mol. Biol.*, 373:573, 2007.

[213] V. Hornak and C. Simmerling. Targeting structural flexibility in hiv-1 protease inhibitor binding. *Drug Discov. Today*, 12:132–138, 2007.

[214] R W Shafer. Rationale and uses of a public hiv drug-resistance database. *J. Infect. Dis.*, 194:S51–S58, 2006.

[215] T Kortemme and D Baker. A simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl. Acad. Sci. U. S. A.*, 99:14116–14121, 2002.

[216] T Kortemme, D E Kim, and D Baker. Computational alanine scanning of protein-protein interfaces. *Sci. STKE*, 219:pl2, 2004.

[217] R E Amaro, R Baron, and J A McCammon. An improved relaxed complex scheme for receptor flexibility in computer-aided drug design. *J. Comput. Aided Mol. Des.*, 22:693–705, 2008.

[218] Pilar Cossio, Fabrizio Marinelli, Alessandro Laio, and Fabio Pietrucci. Optimizing the performance of Bias Exchange Metadynamics for Protein Folding . *SUBMITTED*, 2009.

[219] Kranjc Agata, Bongarzone Salvatore, Rossetti Giulia, Biarnes Xevi, Cavalli Andrea, Bolognesi MariaLaura, Roberti Marinella, Legname Giuseppe, and Carloni Paolo. Docking Ligands on Protein Surfaces: The Case Study of Prion Protein. *J. Chem. Theory Comput.*, 5:2565–2573, 2009.