FOLDING, MISFOLDING AND AGGREGATION OF PROTEINS AND PEPTIDES: A MOLECULAR DYNAMICS STUDY

Isabella Daidone

Dottorato di ricerca in scienze chimiche - XVII ciclo

Dipartimento di Chimica
Università degli Studi di Roma "La Sapienza"

Coordinatore: Prof. Pasquale De Santis
Dipartimento di Chimica
Università degli Studi di Roma "La Sapienza"

Supervisore: Prof. Alfredo Di Nola
Dipartimento di Chimica
Università degli Studi di Roma "La Sapienza"

Docenti esaminatori: Prof. Alfredo Di Nola
Dipartimento di Chimica
Università degli Studi di Roma "La Sapienza"

Prof. Antonio Palleschi
Dipartimento di Scienze e Tecnologie Chimiche
Università degli studi di Roma "Tor Vergata"

Prof. Francesco Ramondo
Dipartimento di Chimica, Ingegneria Chimica e Materiali
Università dell'Aquila

This thesis is based on the following publications:

1. Daidone I., Amadei A., Roccatano D. and Di Nola A. *Molecular dynamics simulation of protein folding by essential dynamics sampling: folding landscape of horse heart cytochrome c.* **Biophys. J.** 2003, 85:2865-2871.

2. Daidone I., Roccatano D. and Hayward S. *Investigating the accessibility of the closed domain conformation of citrate synthase using essential dynamics sampling.* **J. Mol. Biol**. 2004, 339:515-525.

3. Daidone I., Simone F., Roccatano D., Broglia R.A., Tiana G., Colombo G. and Di Nola A. *ß-hairpin conformation of fibrillogenic peptides: structure and a-ß transition mechanism revealed by molecular dynamics simulations.* **Proteins** 2004, 57:198-204.

4. Daidone I., Amadei A. and Di Nola A. *Thermodynamic and kinetic characterization of a ß-hairpin peptide in solution: the complete phase space sampling by molecular dynamics simulations in explicit water*. **Proteins** 2004, in press.

5. Colombo G., Daidone I., Gazit E., Amadei A. and Di Nola A. *Molecular dynamics simulation of the aggregation of the core recognition motif of the islet amyloid polypeptide in explicit water.* **Proteins** 2004, in press.

Other related publications:

6. Roccatano D., Daidone I., Ceruso M.A., Bossa C. and Di Nola A. *Selective excitation of native fluctuations during thermal unfolding simulations: horse heart cytochrome c.* **Biophys. J.** 2003, 84:1876-1883.

7. Flöck D., Daidone I. and Di Nola A. *A Molecular Dynamics study of acylphosphatase in aggregation promoting conditions: the influence of TFE/water solvent.* **Biopolymers** 2004, 75:491-496.

8. Bossa C., Amadei A., Daidone I., Anselmi M., Vallone B., Brunori M. and Di Nola A. *Molecular dynamics simulation of sperm whale myoglobin: effects of mutations and trapped CO on the structure and cavities dynamics.* **Biophys. J.** 2004, submitted.

9. Amadei A., D'Abramo M., Ramondo F., Daidone I., D'Alessandro M., Di Nola A. and Aschi M. *Statistical mechanical characterization of the free energy surface and (classical) kinetics for an intramolecular reaction in solution: proton transfer in aqueous malonaldehyde.* 2004, submitted.

10. Daidone I., Di Nola A. and Amadei A. *α–helix and ß-hairpin folding kinetics over weak free energy barriers.* 2004, in preparation.

The most exciting phrase to hear in science, the one that heralds new discoveries, is not "Eureka!" (I found it!) but "That's funny..." *Isaac Asimov*

Contents

CHAPTER

ONE

Introduction

The process by which a linear sequence of amino acids folds into a unique functional three-dimensional protein is one of the most remarkable examples of the effect of natural selection on biological molecules. The main information for protein folding is contained in the amino acid sequence that is subject to evolutionary pressure to adjust folding rates and product stability according to physiological needs. Only correctly folded proteins provide selectivity and diversity in their functions and have long-term stability in crowded biological environments. The failure to fold correctly, or to remain correctly folded, is the origin of a wide variety of pathological conditions, which may lead to aberrant degenerative diseases such as Alzheimer's and Prion diseases. These illnesses are associated with the aggregation of uncorrectly folded, or "misfolded", structures whose growth may lead to amyloid fibrils and plaques formation, that are usually found in the damaged organs and tissues. Questions that still remain open are how proteins are able to find their unique native states in such a robust and fast way and why a few of them, such as the prion protein or the Aβ peptide, under certain, almost unknown, conditions escape the quality-control system of the cell, leading to amyloid diseases. In the present thesis work the complex mechanism of protein folding and the nature of misfolding and its links with disease are explored with computational methods, in particular using molecular dynamics (MD[1]) simulations.

The free energy of the native state of a correctly folded protein is only slightly lower than that of denatured and misfolded states under physiological conditions, thus it is not favoured by a great thermodynamic stabilization. Nevertheless, the total number of possible conformations of a polypeptide chain is so large that it would take an astronomical length of time to find one particular structure by means of a systematic search of the whole conformational space. Hence, it becomes evident that only a very small number of all possible conformations needs to be sampled during the folding

[1]A complete list of abbreviations used in the present thesis work is provided at the end of the manuscript

process, but how this is achieved is still an open question. To explain the way a protein restricts its conformational search, Levinthal[1] postulated the existence of a series of mandatory steps between specific partially folded states toward the native state. This classical view of folding "pathways" has been extended in a "new view", based on statistical mechanics and polymer physics,[2–5] that emphasizes the ensemble nature of protein conformational states. The new view invokes the concept of an energy landscape for each protein, describing the free energy of the polypeptide chain as a function of its conformational degrees of freedom (Figure 1.1). To enable a protein to fold rapidly and efficiently, the landscape should have the shape of a funnel since, in such a way, the conformational space accessible to the polypeptide chain is reduced as the native state is approached. In essence the high degree of disorder of the polypeptide chain is reduced as folding progresses, as the more favourable enthalpy associated with stable native-like interactions can offset the decreasing entropy as the structure becomes more ordered. On such an energy landscape, a polypeptide chain is able to reach the native state by multiple routes, without the imposition of the same pathway to be followed by every molecule. In a similar scenario, misfolding processes may take place as well: one of the local minimum of the landscape could act as a folding trap, irreversibly capturing the polypeptide chain in a misfolded state and eventually leading to amyloidogenesis.

The study of protein folding is now at the stage where theory and experiment can together make rapid progress toward an understanding of this complex process. The structural transitions taking place during folding can be investigated *in vitro* by a variety of experimental techniques, ranging from optical methods, such as UV Circular Dichroism (CD) and time resolved small angle X-ray scattering (SAXS), to NMR spectroscopy,[5] some of which can now even be used to follow the behaviour of single molecules.[7] The advent of continuous-flow rapid-micromixing techniques,[8, 9] rapid laser-induced temperature-jump heating methods,[10] redox-triggered folding experiments,[11] and NMR lineshape analysis [12] is beginning to allow even direct examination of the earliest events in protein folding (down to nanoseconds) and permitting more direct comparisons to theories and models of protein folding.[13]

Theory too has developed both to address the global character of folding thermodynamics and kinetics and to provide microscopic details.[14, 15] A range of theoretical studies, particularly involving computer simulation techniques,[15] have been largely used to address these questions and MD simulations is one of the most used computational methods. The major problem with MD simulations, in particular when equilibrium thermodynamic properties have to be calculated, is due to the conformational sampling efficiency; even in the 1-microsecond simulation of a 36-residue protein,[16] one of the longest simulations so far afforded, the sampled space explored represents a small fraction of the available conformational space. The problem of adequate sampling is also present in systems with a lower complexity, such as peptides, but is more tractable than for large proteins. Experimentally, peptides fold at very fast rates, requiring probing

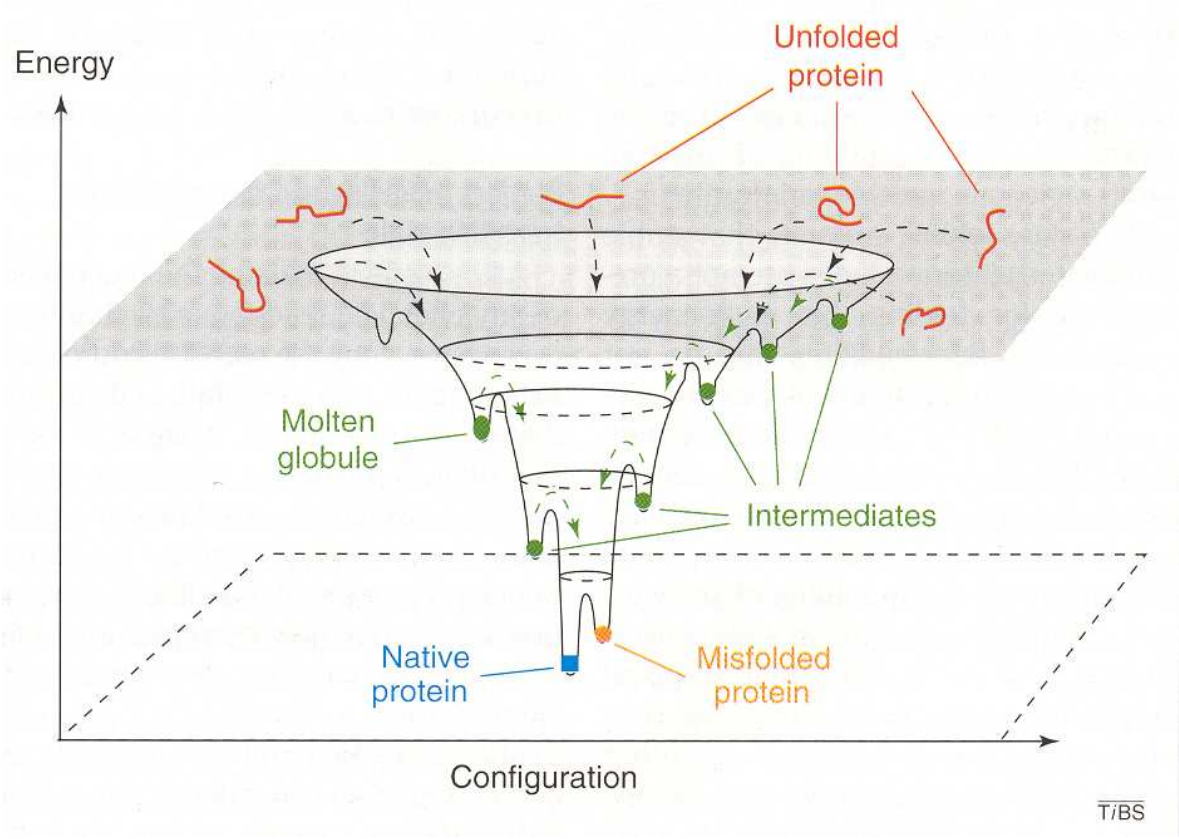


Figure 1.1: Schematic representation of the folding free energy landscape of a protein where the free energy is displayed as a function of the topological arrangements of the atoms. Adapted from Schultz.[6]

on the nanosecond-microsecond time resolution, hence offering a unique opportunity to bridge the gap between theoretical and experimental understanding of protein folding. Only very recently long time scale unbiased MD simulations in the canonical ensemble provided the folding of peptides into α helix[17] or very short β structures.[18]

Apart from the latter exceptions, the development and implementation of new sampling algorithms have become necessary to overcome the limitations of insufficient sampling of the equilibrium thermodynamics and kinetics of folding processes. Conceptually, three categories of techniques can be distinguished: (i) those that simplify the molecular models involved, thus gaining computation time by neglecting details, (ii) those that aim to mimic biological systems as realistically as possible and focus on sophisticated (mathematical) methods to enhance computational efficiency and (iii) those that use thermal unfolding simulations to infer informations on folding, relying on microscopic reversibility. The most relevant methods will be discussed in **Chapter 2**.

For what concerns protein misfolding, the insoluble and massive character of fibrils

rules out the possibility to investigate their formation and their structure at atomic detail with conventional experimental techniques. Using techniques such as electron microscopy and atomic force microscopy,[19–23] the analyses of amyloid deposits show remarkable ultrastructural similarity of fibrils from different sources (e.g. from the pancreas of type II diabetes patients as compared to the brain of Alzheimer's disease patients). Furthermore, X-ray diffraction patterns of several fibrils show a predominant β-sheet structure[24] and it is suggested by spectroscopic techniques that an α to β conformational transition plays a key role in promoting aggregation.[25–29] However, understanding the conformational transitions that trigger aggregation and amyloidogenesis of otherwise soluble proteins and peptides at atomic resolution would be of fundamental relevance for the design of effective therapeutic agents against amyloid related disorders. In such a case the use of computational approaches becomes very useful and precious.

In the present thesis work the transition from an ideal α-helix to a β-hairpin conformation of two well studied amyloidogenic peptides, the 14 residues H1 peptide from prion protein and the Aβ(12–28) fragment from the Aβ(1-42) peptide responsible for Alzheimer disease, is revealed, for the first time, by long time scale, all atom MD simulations in explicit water solvent (**Chapter 3**). Due to the huge time scale afforded by our simulations, we were also able to provide a thermodynamic and kinetic characterization of the folding process of the H1 peptide in water solution (**Chapter 4**). Moreover, the initial self-assembly stages of 26 replicas of another fibrillogenic peptide, the 6 residues core-recognition motif of the type II diabetes associated islet amyloid polypeptide, is studied by MD simulations and forms the subject of **Chapter 5**. These studies provide a description of the molecular determinants involved in fibril formation, in terms of atomic details of the α-β conformational transition and of the structure of nascent aggregates.

For what concerns the study of more complex molecular systems, such as proteins, the development of enhanced sampling algorithms is necessary to overcome the limitations of insufficient sampling. The so-called Essential Dynamics Sampling (EDS) technique will be here described in detail (section 2.3.2) since it is applied in the present thesis to the study of the folding process of an experimentally well studied protein, the cytochrome c (**Chapter 6**), and to the accessibility of the closed and open domain conformations of an important enzyme, the citrate synthase (**Chapter 7**).

CHAPTER

TWO

Methods for Molecular Simulations

2.1 Introduction

In this chapter some basic concepts and methodologies of molecular simulations are introduced with a particular attention devoted to the methods relevant to this thesis. Several books on these subjects can be found with a deeper insight into these problems.[30–32] As the method used in the present thesis to study the properties of large molecular systems, like macromolecules in solution, is classical Molecular Dynamics (MD), a very brief description of its basic principles is presented (section 2.2), with a particular attention to the techniques devoted to the study of folding processes (section 2.3) and to the methods employed to evaluate free energy changes from MD simulations (section 2.4).

2.2 Classical Molecular Dynamics

The aim of computer simulations of molecular systems is to compute macroscopic behavior from microscopic interactions. A model of the real world is constructed, both measurable and unmeasurable properties are computed and the former are compared with experimentally determined properties. If the model used is validated by the comparison, it could even be used to predict unknown or unmeasurable quantities. A theoretical treatment of the motions and interactions of molecules should be founded, rigorously speaking, on quantum mechanics principles, due to the microscopic nature of these objects. Unfortunately, first-principle approaches are often unpractical because they require very large computational facilities and they are definitely prohibitive for systems containing thousands of atoms. Hence, a certain level of approximation becomes necessary and it should be chosen in such a way that those degrees of freedom that are essential to a proper evaluation of the quantity or property of interest can be

sufficiently sampled. When excluding chemical reactions, low temperatures or details of hydrogen atoms motion, it is relatively safe to assume that the system is governed by the laws of classical mechanics.

In classical MD, a trajectory (configurations as a function of time) of the molecular system is generated by simultaneous integration of Newton's equations of motions for all atoms in the system:

$$\frac{d^2\boldsymbol{r}_i}{dt^2} = m_i^{-1}\boldsymbol{F}_i \tag{2.1}$$

$$\boldsymbol{F}_i = -\frac{\partial V(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)}{\partial \boldsymbol{r}_i} \tag{2.2}$$

The force acting on atom i is denoted by $\boldsymbol{F}_i$, the mass by m_i and time is denoted by t. MD simulations require calculation of the gradient of the potential energy $V(\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)$, which therefore must be a differentiable function of the atomic coordinates $\boldsymbol{r}_i$. This potential energy function, or *force field*, is called an *effective interaction* function since the average effect of the omitted (electronic) degrees of freedom has been incorporated in the interaction between the (atomic) degrees of freedom explicitly present in the model.

The choice of molecular model and force field is essential to a proper prediction of the properties of a system. Therefore, it is of great importance to be aware of the fundamental assumptions, simplifications and approximations that are implicit in the various types of models used in the literature.

2.2.1 Force Field Models

A huge variety of force fields is currently used in the molecular dynamics community, sometimes differing for minor changes, e.g. CHARMM,[33] AMBER,[34] GROMOS.[35] A typical molecular force field, or effective potential, for a system of N atoms with masses m_i $(i = 1, 2, \ldots, N)$ and Cartesian position vectors $\boldsymbol{r}_i$ has the following form:

$$\begin{aligned} V(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N) &= \sum_{bonds} \frac{1}{2} K_b (b - b_{eq})^2 + \sum_{angles} \frac{1}{2} K_\theta (\theta - \theta_{eq})^2 \\ &+ \sum_{dihedrals} K_\phi [1 + cos(n\phi - \delta)] + \sum_{imp.dihedrals} \frac{1}{2} K_\xi (\xi - \xi_{eq})^2 \\ &+ \sum_{pairs} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \sum_{pairs} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \end{aligned} \tag{2.3}$$

The first term represents the covalent bond stretching interaction between two atoms linked by a harmonic potential where b_{eq} is the minimum energy bond length and K_b

is the force constant changing with the particular bond type. The second term is a three-body interaction corresponding to the valence angle, θ, deformation expressed as a harmonic potential where θ_{eq} is the equilibrium valence angle and K_θ the force constant. The third and fourth terms are used for the (four-body) dihedral angle interactions: a harmonic term for improper dihedral angles, ξ, that are not allowed to make transitions, i.e. to keep the aromatic rings planar, and a sinusoidal term for all the other dihedral angles, ϕ. The last two terms are sums over the pairs of non-bonded atoms and represent the effective non-bonded interactions expressed in terms of van der Waals and Coulombic interactions between atoms i and j at a distance r_{ij}. The parameters ϵ_{ij} and σ_{ij} are the typical constants defining the Lennard-Jones potential, q_i and q_j are the atom charges and ϵ_0 is the dielectric constant in vacuum.

The parameters used in the force field (Eq. 2.3) can be determined in different ways. Generally two main approaches are followed. The first one is to fit them with results obtained from *ab initio* calculations on small molecular clusters. The alternative way is to fit the force field parameters to experimental data, like crystal structures, energy and lattice dynamics, infrared or X-ray data on small molecules, liquid properties like density and enthalpy of vaporization, free energy of solvation, nuclear magnetic resonance data, etc. Whatever method is used, the resulting model is far to be universal. It is worth to note that every force field is usually well suited for specific general conditions, i.e. particular thermodynamic conditions (temperature, density, pressure, etc.) and also boundary conditions. Moreover, they are optimized for specific classes of molecules, such as inorganic molecules, organic molecules, biomolecules (DNA, proteins, lipids), etc.

2.2.2 The Boundary Conditions

An important characteristic of the molecular dynamics simulations is the way in which the boundaries are treated. Due to computational limits, a typical simulated system contains $10^4 - 10^5$ atoms, and then is quite small compared to macroscopic matter. This means that, if the molecules are arranged in a cubic box, a relatively great part of them will lie on the surface and will experience quite different forces from molecules in the bulk. The consequence of the finite size of the system is that the boundary conditions may affect seriously the results of the simulations, especially when the system of interest is a homogeneous liquid or a solution. Usually, periodic boundary conditions (PBC) [30] are adopted to reduce the surface effects. This technique consists on simulating the system in a central cubic box surrounded by an infinite number of copies of itself. During the simulation, the molecules in the original box and their periodic images move exactly in the same way. Hence, when a molecule leaves the central box one of its images will enter through the opposite side. As a result, there are no physical boundaries neither surface molecules. Note that other shapes of the box can be used

as the truncated octahedron or the rhombic dodecahedron.

2.2.3 Integration of Motion Equations

Newton's equations of motion, a second-order differential equation, can be written as two first-order differential equations for the particle positions $\boldsymbol{r}_i(t)$ and velocities $\boldsymbol{v}_i(t)$ respectively:

$$\frac{d\boldsymbol{v}_i(t)}{dt} = m_i^{-1}\boldsymbol{F}_i \tag{2.4}$$

$$\frac{d\boldsymbol{r}_i(t)}{dt} = \boldsymbol{v}_i(t) \tag{2.5}$$

A standard method for solution of the previous ordinary differential equations is the finite difference approach. The general idea is the following. Given the molecular positions, velocities and forces at time t, we attempt to obtain the positions, velocities and forces at a later time $t + \delta t$, to a sufficient degree of accuracy. The equations are solved on a step-by-step basis; the choice of the time interval δt will depend somewhat on the method of solution, but δt will be significantly smaller than the typical time taken for a molecule to travel its own length.

Many different algorithms fall into the general finite difference pattern, like Verlet, and its computational efficient variant *leap-frog*,[36, 37] Beeman [38] or the Gear predictor-corrector.[39]

2.2.4 Enhanced efficiency methods

Since the first published application of MD to biomolecular systems,[40] a little more than 20 years ago, people have devised methods to increase the time scales of MD simulations. When Newton's equations of motion are integrated, the limiting factor that determines the time step that can be taken is the highest frequency that occurs in the system. In solvated biological macromolecules, the vibrations of bonds involving hydrogen atoms form the highest frequency vibrations. The bond stretching frequency of an O-H bond is typically about 10^4 Hz, so the average period would be of the order of 10 fs.[41] This limits the time-step to be taken in MD simulations to about 0.5 fs (a rule of thumb exists that states that for a reasonable sampling of a periodic function, samples should be taken at least twenty times per period). The introduction of a method to constrain these bonds (or, in fact, all covalent bonds) allowed to increase the time step to a typical value of 2 fs (*SHAKE*).[42] Since these bond vibrations are practically uncoupled from all other vibrations in the system, constraining them does not notably alter the rest of the dynamics of the system. This is not true, however, for bond-angle fluctuations, which form the second-highest frequency vibrations. Constraining bond-angles has a

severe effect on many other fluctuations in the system, including even global, collective fluctuations, limiting the use of methods that use bond-angle constraints to only a few specific cases.[41]

The notion that a number of discrete classes of frequencies of fluctuations in simulations of biomolecules can be distinguished, however, can be utilized to design more efficient algorithms. Forces that fluctuate rapidly need to be recalculated at a higher frequency than those that fluctuate on a much longer time scale. Although not trivial to implement, a number of successful applications of so-called *multiple time-step* algorithms have been reported in the literature (for a review, see Schlick *et al.* [41]). Speed up factors of 4-5 have been claimed for such methods with respect to unconstrained dynamics, making them only slightly more efficient than simulations with covalent bond-length constraints.

Another approach to reach equilibrium conformational properties at an enhanced rate is by constraining the rotational and translational motions in polyatomic systems.[43] This method is generally used to study biomolecules in solution. In such a system, the internal motions of the solute are often more interesting than its rotational and translational motions. This algorithm is implemented in a *leap-frog* integration scheme coupled with SHAKE. The use of the *roto-translational constraint* presents several advantages, like a reduction of the molecular relaxation time and the possibility of reducing the amount of solvent molecules to be used.[44]

2.2.5 Long-range Interactions

One of the most challenging problems in molecular dynamics simulations is the treatment of long-range interactions, which usually correspond to the electrostatic forces between molecules. To reduce the computational cost, the size of the simulated system is generally small and, as a consequence, a correct evaluation of the intermolecular interactions is not trivial. Many different methods were developed to reproduce reasonably the thermodynamics of bulk liquids. Here we consider two of the most used techniques: the use of a cut-off radius and the Ewald sum.

The *cut-off method* is based on the truncation of the forces when the distance between the interacting particles is greater than a specified value, called the cut-off radius, r_c. In this way, the only interactions felt by the i-th particle are those due to the particles contained in a sphere of radius r_c and centered at $\boldsymbol{r}_i$. This method is doable only if the intermolecular forces decay rapidly with the distance. In fact, when the forces are negligible at a distance $\geq r_c$, the main structural and dynamical properties are correctly reproduced. Otherwise deviations from the correct bulk behaviour are expected.

Another methodology in MD simulations is the use of a *periodic lattice method* in which all the interactions between the molecular system in the central cubic cell and its virtual replica are included. The Coulomb interaction energy in a periodic system

of N charged particles is obtained by a sum over all pairs of which one atom lies in the central box and the other is its periodic image:

$$E = \frac{1}{8\pi\epsilon_0} \sum_{|\boldsymbol{n}|=0}^{\infty} \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \frac{q_i q_j}{|\boldsymbol{r}_{ij} + \boldsymbol{n}|} \right) \tag{2.6}$$

The sum over $\boldsymbol{n}$ is a summation over all simple cubic lattice points $\boldsymbol{n} = (n_x L, n_y L, n_z L)$, with L the side length of the cubic cell and n_x, n_y, n_z integers. The case $i = j$ is omitted for $\boldsymbol{n} = 0$. It was shown that the sum over $\boldsymbol{n}$ for such kind of potential (r^{-1}) is only conditionally convergent, then its limit may vary or even diverge if the order of terms in the sum is changed. A solution to this problem was developed following a physical idea:[30] each point charge is surrounded by a charge distribution of equal magnitude and opposite sign, which spreads out radially from the charge, $\rho^G(\boldsymbol{r})$. This distribution has the effect to screen the interactions between the neighbouring point charges and hence the interaction energy becomes short-ranged. Commonly, the screening charges have a Gaussian form. The total charge distribution is given by:

$$\rho_i(\boldsymbol{r}) = \rho_i^q(\boldsymbol{r}) + \rho_i^G(\boldsymbol{r}) \tag{2.7}$$

where $\rho_i^q(\boldsymbol{r})$ is the distribution of the point charge of the i-th particle and $\rho_i^G(\boldsymbol{r})$ is the corresponding Gaussian distribution.
First, the interaction energy due to the distribution 2.7 is calculated in the real space, then, in order to recover the original charge distribution, a canceling function is added in the reciprocal space, which is equal to $-\rho_i^G(\boldsymbol{r})$, realized by means of a Fourier transform. Hence the final form of the total interaction energy is given by:

$$\begin{aligned} E &= \frac{1}{8\pi\epsilon_0} \sum_{i=1}^{N} \sum_{j=1}^{N} \Bigg(\sum_{|\boldsymbol{n}|=0}^{\infty} \frac{q_i q_j erfc(\alpha|\boldsymbol{r}_{ij} + \boldsymbol{n}|)}{|\boldsymbol{r}_{ij} + \boldsymbol{n}|} \\ &+ \frac{1}{\pi L^3} \sum_{\boldsymbol{k} \neq 0} \frac{4\pi^2 q_i q_j}{k^2} \exp(-k^2/4\alpha^2) \cos(\boldsymbol{k} \cdot \boldsymbol{r}_{ij}) \Bigg) \\ &- \frac{\alpha}{4\pi^{3/2}\epsilon_0} \sum_{i=1}^{N} q_i^2 + \frac{|\sum_{i=1}^{N} q_i \boldsymbol{r}_i|^2}{2\epsilon_0 L^3 (2\epsilon' + 1)} \end{aligned} \tag{2.8}$$

Here $erfc(x)$ is the complementary error function, which falls to zero with increasing its argument. Thus, if the parameter α is large enough, the sum over $\boldsymbol{n}$ in the first term reduces to the only term $\boldsymbol{n} = 0$. The second term is a sum over the reciprocal vectors $\boldsymbol{k} = 2\pi\boldsymbol{n}/L$. Again, if α is large, a lot of terms in the k-space sum are needed to get a convergence of the energy. The last two terms are, respectively, a correction function, due to the fact that a self-interaction of the canceling distribution is included in the recipe, and the energy contribution of the depolarizing field, which is compensated by

the effect of the external dielectrics. Note that in the Ewald sum the virtual cubic cells are ordered as concentric spherical layers starting from the central box. Clearly the infinite sum is truncated at a certain point and the resulting spherical system is immersed in a continuum dielectrics with dielectric constant ϵ'. The last term in equation 2.8 is the sum of the contributions of the depolarizing field and the reaction field due to the external dielectrics. If the sphere is embedded in a medium with an infinite dielectric constant, this term vanishes because of a perfect compensation of the two effects.

Other *periodic lattice methods* are often used in computer simulations for their computational stability and efficiency. These methods, like the Particle Mesh Ewald (PME)[45] method, can be considered of the same family of the method shown here.

2.2.6 Constant Temperature/Constant Pressure Molecular Dynamics

When Newton's equations of motion are integrated the total energy is conserved (adiabatic system) and if the volume is held constant the simulation will generate a microcanonical ensemble (NVE). However, this is not always very convenient. Other statistical ensembles, such as canonical (NVT) and isothermal-isobaric (NPT) ensembles, better represent the conditions under which experiments are performed than the standard microcanonical ensemble. Moreover, with the automatic control of temperature and/or pressure, slow temperature drifts that are an unavoidable result of force truncation errors are corrected and also rapid transitions to new desired conditions of temperature and pressure are more easily accomplished.

Several methods for performing MD at constant temperature have been proposed, ranging from *ad hoc* rescaling of atomic velocities in order to adjust the temperature, to consistent formulation in terms of modified equations of motion that force the dynamics to follow the desired temperature constraint. The three most utilized methods are described next.

The *termal bath coupling* method, or Berendsen coupling,[46] has the great advantage of being simple. This algorithm simulates a coupling of the system with an external termal bath at the temperature T_0 and the interaction between this bath and the system is modulated by a time constant τ. The coupling is obtained multiplying for a constant λ the velocities. The temperature T is scaled to the reference temperature T_0 via an exponential law.

The *isothermal* , or isogaussian, method [47] allows to fix the temperature exactly constant. Using this algorithm, a variable is added to the motion equations, acting as a friction coefficient changing in time in order to keep the kinetic energy constant. This method correctly generates the configurational properties of the canonical ensemble, while the momenta distribution is not canonical.[43]

Nosé-Hoover thermostat is based on the use of an extended Lagrangian, i.e. a Lagrangian that contains additional, artificial coordinates and velocities.[48, 49] The conventional Nosé-Hoover algorithm only generates the correct distribution if there is a single constant of motion. Normally, the total energy, that includes the artificial variables, is always conserved. This implies that one should not have any other conserved quantity. If we have more than one conservation law, we have to use the Nosé-Hoover chains to obtain correct canonical distribution.[50]

The various methods for carrying out MD at constant pressure are based on the same principles as the constant temperature scheme with the role of the temperature played by the pressure and the role of the atomic velocities played by the atomic positions.

2.2.7 Essential Dynamics

The Essential Dynamics (ED) analysis is a method to seek those collective degrees of freedom that best approximate the total amount of fluctuation of a dynamical system.[51, 52] A brief description will be given here. ED is based on a principal component analysis (PCA) of (MD generated) structures. A PCA is a multidimensional linear least squares fit procedure. To understand how this is applicable to protein dynamics, the usual three-dimensional (3D) Cartesian space to represent protein coordinates (which is e.g. used to represent protein conformations in the Brookhaven Protein Data Bank or PDB) needs to be replaced by another, multidimensional space. A molecule of N particles can be represented by N points in 3D space. With 3 coordinates per point, this adds up to 3N coordinates. In a 3N-dimensional space, however, such a structure can be represented by a single point. In this space, this point is characterized by 3N coordinates. This representation is convenient since a collection or trajectory of structures can now be regarded as a cloud of points. Like in the case of a two-dimensional cloud of points, also in more dimensions, always one line exists that best fits all points. As illustrated for a two-dimensional example (Figure 2.1), if such a line fits the data well, the data can be approximated by only the position along that line, neglecting the position in the other direction. If this line is chosen as coordinate axis, then the position of a point can be represented by a single coordinate. In more dimensions the procedure works similarly, with the only difference that one is not just interested in the line that fits the data best, but also in the line that fits the data second-best, third best, and so on (the principal components). These directions together span a plane, or space, and the subspace responsible for the majority of the fluctuations has been referred to as the 'essential subspace'. Applications of such a multidimensional fit procedure on protein configurations from MD simulations of several proteins has proven that typically the ten to twenty principal components are responsible for 90% of the fluctuations of a protein.[51–53] These principal components correspond to collective coordinates, containing contributions from every atom of the (protein) molecule. Sum-

marised, a limited number of collective motions is responsible for a large percentage of a protein conformational fluctuations.

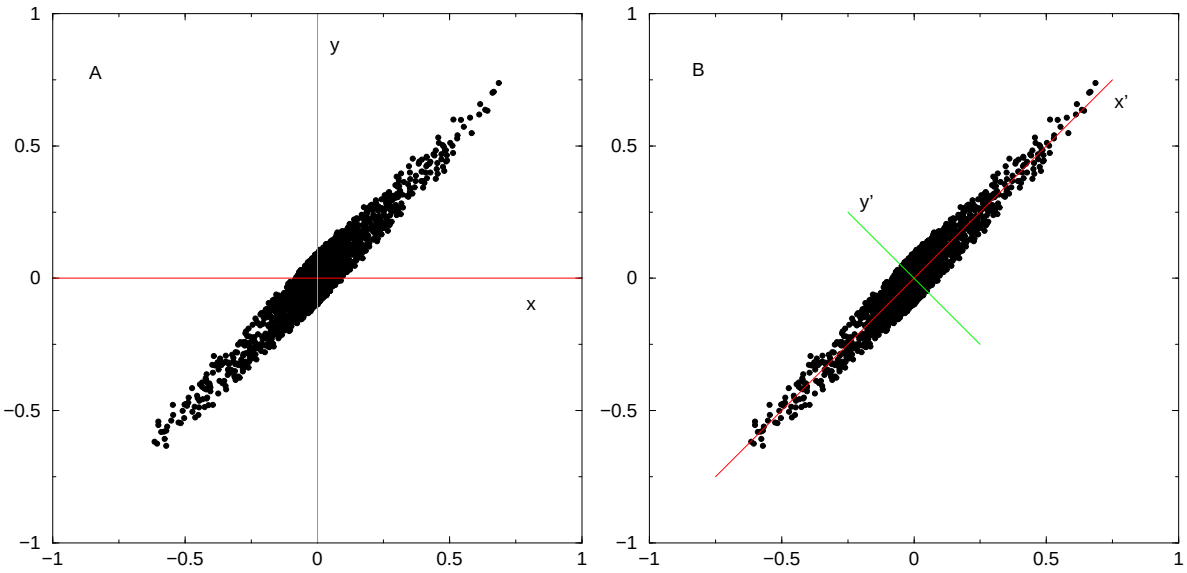


Figure 2.1: Example of Essential Dynamics in two dimensions. With a distribution of points as depicted here, two coordinates (x,y) are required to identify a point in the cluster in panel A, whereas one coordinate (x') approximately identifies a point in panel B

If all atoms in a protein were able to move uncorrelated from each other, an approximation of the total fluctuation by only a few collective coordinates would not be possible. The fact that such an approximation is successful is the result of the presence of a large number of internal constraints and restrictions ('near-constraints') defined by the interactions present in a given protein structure. Atomic interactions, ranging from covalent bonds (the tightest interactions) to weak non-bonded interactions, together with the dense packing of atoms in native-state protein structures form the basis of these restrictions.

In the study of protein dynamics, only internal fluctuations are usually of interest. Therefore, the first step in an Essential Dynamics analysis is to remove overall rotation and translation. This is done by translation of the center of mass of every configuration to the origin after which a least squares rotational fit of the atoms is performed onto to a reference structure. The actual principal component analysis is based on construction and diagonalisation of the covariance matrix of positional fluctuations. Defining the 3N dimension column vector $\boldsymbol{X}(t)$ representing the atomic coordinates of the system at time t, the covariance matrix is built up according to:

$$\boldsymbol{C} = \langle \Delta \boldsymbol{X} \Delta \boldsymbol{X}^T \rangle \tag{2.9}$$

where $\Delta \boldsymbol{X} = \boldsymbol{X}(t) - \langle \boldsymbol{X} \rangle$ and the angle brackets represent a time or ensemble average. Particles moving in a correlated fashion correspond to positive matrix elements (positive correlation) or negative elements (negative correlation) and those that move independently to small matrix elements. The orthogonal transformation $\boldsymbol{T}$ that diagonalises this (symmetric) matrix contains the eigenvectors or principal components of $\boldsymbol{C}$ as columns and the resulting diagonal matrix $\boldsymbol{\Lambda}$ contains the corresponding eigenvalues:

$$\boldsymbol{\Lambda} = \boldsymbol{T}^T \boldsymbol{C} \boldsymbol{T} \tag{2.10}$$

The eigenvalues are the positional mean square fluctuations along the corresponding eigenvectors. When the eigenvectors are sorted to decreasing eigenvalues, the first eigenvectors are those collective motions that best approximate the sum of fluctuations and the last eigenvectors correspond to the most constrained degrees of freedom. The characteristics of these collective fluctuations can be studied by projecting the ensemble of structures onto single eigenvectors and by translation of these projections to 3D space to visualize the atomic displacements connected with that eigenvector. As stated above, analyses of MD trajectories of several proteins have shown that few collective coordinates dominate the dynamics of native proteins (together often referred to as the 'essential subspace'). In a number of cases these main modes of collective fluctuation were shown to be involved in the functional dynamics of the studied proteins.[51, 54, 55]

ED analyses can be applied to any subset of atoms of the ensemble of structures and are not restricted to ensembles generated by MD simulation. Applications to collections of X-ray structures,[55, 56] NMR structures[57] and structures derived from distance constraints[58] have been reported. Since collective (backbone) fluctuations dominate the dynamics of proteins, usually only backbone or Cα coordinates are used to save computation time and to prevent problems with apparent correlation of side chain motions with backbone motions which are merely the result of poor statistics. However, even when the method is applied to only Cα atoms, the diagonalisation of the covariance matrix can still be an enormous computational task.

2.3 Protein folding simulations

A clear gap exists between time scales that can currently be obtained by computer simulation techniques applied to biological macromolecules and the times required for most biological processes. With current methods and computer state of the art, a typical protein of 1000 amino-acids (100 kD) can be simulated for time-scales of at most tens of nanoseconds, whereas most biological processes, such as protein folding, take place at times ranging between microseconds to seconds (or even minutes). Even if the present rate of increase in computer power (an order of magnitude every 5-7 years) will continue in the future, simulation of such processes at the required time scales will

be beyond those of standard MD simulation protocols in the next decade.

To overcome the limitations of insufficient sampling of the equilibrium thermodynamics and kinetics of folding processes, the development and implementation of new sampling algorithms have become necessary. Conceptually, three categories of techniques can be distinguished: (i) those that simplify the molecular models involved, thus gaining computation time by neglecting details (section 2.3.1), (ii) those that aim at mimicking biological systems as realistically as possible and focus on sophisticated methods to enhance computational efficiency (section 2.3.2) and (iii) those that use thermal unfolding simulations to infer informations on folding, relying on microscopic reversibility (section 2.3.3). This division is not exclusive; some methods cannot be assigned to either category whereas others are hybrid methods based on principles from more than one category. A number of examples from each of the categories will be discussed in the next sections, and in particular a technique from the second category, the so-called Essential Dynamics Sampling technique, will be described in detail since it will play a key role throughout the rest of this thesis.

2.3.1 Simplified models

Simulations of protein folding with a simplified protein model have been utilized extensively, especially in the presence of explicit solvent. Employed methodologies include lattice Monte Carlo (MC) models and adapted MD or Langevin Dynamics (LD) models.

Lattice models form perhaps the most simplified models with some resemblance to real proteins.[59–61] Their advantage is that exhaustive searches of the configuration space can be reached for small proteins (up to about 100 residues) by Monte Carlo methods.[62–65] However, their applicability is limited due to the lack of detail in the models and the restriction of the search space due to lattice constraints.

Continuum "minimalist" models of simplified proteins (bead models), utilizing adapted MD or LD algorithms, are more promising, because of the absence of lattice restrictions. In Langevin Dynamics, compared to classical Molecular Dynamics, forces contain an additional friction term to mimic the effect of solvent (which is not treated explicitly).[66] Although exhaustive searches can usually not be reached by these bead methods, promising results have been reported.[67–70]

Another application of simplified protein models, used in native folds prediction, are the so called *threading techniques*,[71–73] some of which make use of neural networks.[74, 75] The idea is that a discrete number of folds exists to which proteins are restricted. The sequence of a protein with unknown structure is threaded through a set of known protein folds, after which suitable scoring potentials reveal which structure is most probable for that sequence.

Another way of simplifying the complexity of the simulated system is to neglet explicit solvent degrees of freedom. Several methods of solvent treatment by *implicit*

models have been suggested over the years,[76–79] but their range of applicability is still a matter of debate.[80–82]

2.3.2 Enhanced sampling algorithms

The most widely used enhanced sampling algorithms can be divided into two classes. On one hand there are those techniques that make use of embarrassingly parallel schemes to enhance sampling, thus making efficient use of multiprocessor low-cost cluster machines, such as parallel replica dynamics and replica exchange molecular dynamics. On the other hand there are methods that make use of biasing potentials or constraint forces to enhance sampling, such as umbrella sampling, multicanonical sampling and Essential Dynamics Sampling (EDS). Brief descriptions and applications of these methods are described next, giving a particular emphasis to the latter technique, since it is extensively used in this thesis.

The simplest parallel sampling method is to run many uncoupled copies of the same system with different initial conditions.[83–87] The massive parallelism inherent in this method has been useful in project such as Folding@Home,[88] which uses the excess compute cycles of weakly coupled private computers. This simple parallel simulation method is most successful for systems with implicit solvent, the use of which increases the slowest relaxation rates by factors of 100-1000.[83] With explicit solvent, most single simulations are short compared to the system relaxation time and are strongly influenced by the initial conditions. A more sophisticated method of this class is the *parallel replica dynamics* (PRD) method.[89] In this method, independent simulations are started from the same conformational basin. When one of these simulations exits a basin, all the other simulations are restarted from the new basin. Although this method has been successfully applied on peptides,[90] it is suspect when applied to proteins because of the difficulty of identifying when a barrier-crossing event has occurred.[91]

In the *replica exchange molecular dynamics* (REMD) simulations,[92, 93] M non interacting copies (or replicas) of the original system are simulated in parallel in the canonical ensemble at M different temperatures. At fixed time intervals, replicas having neighbouring temperatures are exchanged with periodic Metropolis Monte Carlo temperature-exchange trials. The REMD method has many advantages. It is particularly easy to implement, produces information over a range of temperatures and is easily adapted for use with implicit or explicit solvent.

One of the oldest methods to enhance the calculation of static properties is *umbrella sampling*. In the umbrella sampling method, separate simulations starting from different points of the configurational space are carried out with modified potential functions. These separate simulations are then combined and the resulting phase space distribution is corrected to determine what it would have been if the sampling had been done with the original unbiased potential.[94, 81, 95, 15]

A less obvious method of umbrella sampling is to use a biasing potential that is solely a function of the potential energy. Determining this bias self-consistently, so that all potential energies are equally sampled, allows the system to do random walk in potential energy space and easily surmount large enthalpic barriers. This is the *multicanonical method* created by Berg and Neuhaus[96] and applied by others to the study of peptide folding.[97–99] Although widely used, the determination of the biasing function is difficult, especially for systems with explicit solvent, since iterative simulations are required to evaluate the biasing function. The parallel version of this algorithm was adapted for use with REMD. [92]

Another approach incorporates experimental measurements, such as NOE and ϕ values, directly into the simulations as restraints limiting the regions of conformational space that are explored in each simulation. This strategy has enabled rather detailed structures to be generated for transition, intermediate and denatured states of several proteins.[100–102]

Essential Dynamics Sampling (EDS)

The *Essential Dynamics Sampling* technique is based on the dominant modes of collective fluctuation of proteins revealed by the ED analysis. Once an approximation of the collective degrees of freedom (essential eigenvectors) has been obtained (see paragraph 2.2.7), constraint (non-deterministic) forces are used to move the system preferentially in the subspace spanned by only these coordinates (essential subspace). This method exploits the limited dimensionality of the essential subspace to achieve a more efficient sampling than can be obtained by more conventional techniques.[103, 104] The EDS technique can be used to increase (*expansion mode*) or decrease (*targeting mode*) the distance from a reference structure. To this end, the distance is calculated in the new reference system (the one obtained by the ED analysis) using only a subset of the generalized degrees of freedom of the system, i.e. a subset of the eigenvectors. At each time frame the usual MD step is performed and the distance in the subspace between the current conformation and the reference conformation is calculated. The step is accepted if this distance does not decrease, in the case of expansion, or does not increase, in the case of targeting. Otherwise the coordinates and velocities are projected radially onto the hypersphere (in the subspace) centred on the reference conformation, with a radius given by the distance from the reference in the previous step (see Figure 2.2). It has to be pointed out that with this biased MD simulation no deterministic force is added to the system.

Although proposed in 1996, this technique has been applied only in the *expansion mode*, to enhance native state protein dynamics,[103, 104] but it was never used before in the *targeting mode* to follow the folding process of a protein towards its native structure. Its first application to protein folding is the subject of chapter 6 of this thesis while an

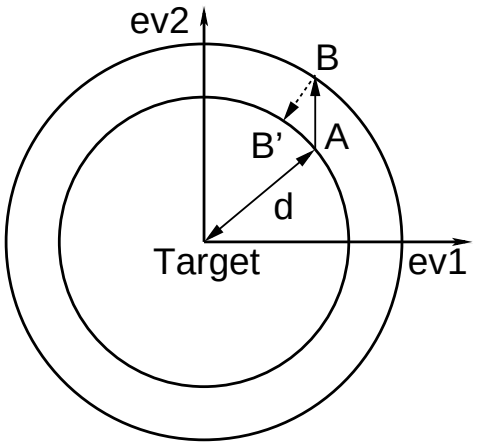


Figure 2.2: Essential dynamics sampling; example for the contraction procedure in a bidimensional case. A: structure at step 'i'; B: structure at step 'i+1'; B': new structure at step 'i+1'. ev1 and ev2 represent eigenvectors 1 and 2, respectively.

application to enzymatic functionality is the subject of chapter 7.

2.3.3 High Temperature simulations

An alternate means of studying protein folding in a fully atomic representation of the protein is through temperature (or denaturant)-induced unfolding simulations. Then to infer folding from unfolding trajectories, one has to rely on microscopic reversibility and reverse sequence of events observed in unfolding.[105–108] However, the unfolding process may not necessarily be the reverse of the folding process and therefore the issue of whether unfolding simulations are representative for the folding process is still open.[109]

2.4 Free energy calculations

In general terms, a microscopic description of a particular molecular system can be given in the form of a Hamilton operator or function. This is often simply expressed as the Hamiltonian $H(\boldsymbol{p}, \boldsymbol{q})$ of the generalized coordinates $\boldsymbol{q}$ and their conjugate momenta $\boldsymbol{p}$. For example, the Hamiltonian for a classical system of N atoms, expressed in terms of the Cartesian coordinates $\boldsymbol{r}$ and momenta $\boldsymbol{p}$ of each of the atoms, has the form $H(\boldsymbol{p}, \boldsymbol{r}) = K(\boldsymbol{p}) + U(\boldsymbol{r})$, where K is the kinetic and U the potential energy. In the canonical ensemble the fundamental formula for the Helmholtz free energy, A, is: [110]

$$A(N, V, T) = -k_B T \ln Q(N, V, T) \tag{2.11}$$

where the partition function Q is:

$$Q(N,V,T) = h^{-3N} \int\int e^{-H(\boldsymbol{p},\boldsymbol{r})/k_BT} d\boldsymbol{p}d\boldsymbol{r} \tag{2.12}$$

where V is the volume of the system, T the absolute temperature, k_B Boltzmann's constant, h Planck's constant, and it is assumed that the N atoms are distinguishable. The factor before the integral actually comes from quantum mechanics. The essential difficulty in calculating the free energy of a system is evident from Eqn. 2.12, which is dependent on a $6N$-dimensional integral to be carried out over phase space.

By means of statistical mechanics, free energy differences may also be expressed in terms of averages over ensembles of atomic configurations for the molecular system of interest. Such an ensemble can be generated by MC or MD simulation techniques. If the *ergodic hypothesis* is verified, that is the simulated trajectory will visit all possible microstates available to it, given an infinite amount of time the following equivalence holds:

$$\langle \mathcal{A}(\boldsymbol{q}(t),\boldsymbol{p}(t)) \rangle_{ensemble} = \langle \mathcal{A}(\boldsymbol{q}(t),\boldsymbol{p}(t)) \rangle_{time} \tag{2.13}$$

that is the ensemble average of a generic physical observable, $\mathcal{A}(t)$, is equivalent to its time average. In principle this equivalence offers a valid method, the time average, to obtain physical properties from our "virtual" experiment, namely computer simulations. However, despite its inherent simplicity, the computation of thermodynamic properties from molecular simulations remains far from trivial due to the limit of infinite sampling of phase space and to unavoidable numerical errors.

Within the framework of statistical mechanics, a variety of formulae for determining the difference in free energy between two states of a system, or the projection of such a difference in free energy along a spatial (reaction) coordinate, have been derived. The different formulations available are all equivalent within the limit of infinite sampling of phase space. In practice, as only a part of the total phase space accessible to a realistic system can ever be sampled by molecular simulations techniques, there are often significant differences in accuracy between the free energy estimates obtained from different formulae. Below a list of the most useful statistical mechanical formulae and computational methods to obtain the difference in free energy $\Delta A_{A\to B} = A_B - A_A$ between a state B and a state A of a molecular system in a canonical ensemble is provided.

2.4.1 Probability ratio method

In equilibrium thermodynamics, free energy changes are related to the populations (or probabilities) of states. Hence, the most straightforward way to determine the difference in free energy between two states of a system is simply to count the number of configurations in the two corresponding states. For example, in the case of folding, this

involves counting the number of folded conformations N_F and the number of unfolded conformations N_U in an ensemble generated during a MD or MC simulation, with the difference in free energy being given by

$$\Delta A_{U\rightarrow F} = -k_B T \ln \frac{Q_F}{Q_U} = -k_B T \ln \frac{p_F}{p_U} = -k_B T \ln \frac{N_F}{N_U} \tag{2.14}$$

where k_B is the Boltzmann constant, T is the temperature, Q_F and Q_U are the partition functions of the folded and unfolded states, respectively, and p_F and p_U are the probability densities of finding the system in the folded or unfolded states, respectively. This technique is only appropriate when folded and unfolded conformations occur with sufficient frequency in the ensemble to obtain reliable statistics. An example of the use of Eqn. 2.14 to determine the difference in folding free energy can be found in chapter 4 of this thesis. Direct counting has the advantage that it does not depend on the definition of a reaction coordinate and it is particularly well-suited to situations in which the end states are themselves ensembles of structures, such as in the study of protein/peptide folding.

2.4.2 Thermodynamic Integration

Integrations methods determine the change in free energy between two states of a system from the integral of the work required to go from an initial state to a final state *via* a reversible path. In *Thermodynamic Integration* (TI) method an arbitrary coupling parameter, λ, is introduced in the the Hamiltonian $H(\boldsymbol{p}, \boldsymbol{q}, \lambda)$. The coupling parameter is chosen such that when $\lambda = \lambda_A$ the Hamiltonian of the molecular system corresponds to that of state A, i.e. $H(\boldsymbol{p}, \boldsymbol{q}, \lambda_A)$=$H(\boldsymbol{p}, \boldsymbol{q})$ and when $\lambda = \lambda_B$ the Hamiltonian of the system corresponds to that of state B, i.e. $H(\boldsymbol{p}, \boldsymbol{q}, \lambda_B)$=$H(\boldsymbol{p}, \boldsymbol{q})$. If the Hamiltonian is a function of λ the free energy in Eqn. 2.11 will also be a function of λ, and the derivative of the free energy with respect to λ will be given by

$$\frac{dA(\lambda)}{d\lambda} = \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_\lambda \tag{2.15}$$

From this, it follows directly that the free energy difference between state A and state B of a molecular system is given by

$$A(\lambda_B) - A(\lambda_A) = \int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_\lambda d\lambda \tag{2.16}$$

which is the so-called thermodynamic integration formula.[111] The ensemble average $\langle \partial H/\partial \lambda \rangle$ is most commonly determined from simulations at a series of λ values between λ_A and λ_B and the integral in Eqn. 2.16 evaluated numerically. The choice of λ is arbitrary and λ may equally refer to a spatial coordinate or to a non-physical coordinate

in parameter space. In either case, the functional dependence of the system on λ effectively describes the pathway from the initial to the final state.

2.4.3 Perturbation method

An alternative to the TI method is to adopt a perturbation approach. In the *perturbation method* (PM) the free energy change is expressed by the following relation:[112]

$$A_B - A_A = -k_B T \ln \frac{Q_B}{Q_A} = -k_B T \ln \left\langle e^{\Delta H/k_B T} \right\rangle_B \tag{2.17}$$

where Q_B and Q_A are the partition functions of state B and A respectively, $\Delta H = H_B - H_A$ is the energy difference, k_B is the Boltzmann constant and T the absolute temperature. The subscript on the brackets $\langle ... \rangle$ indicates that the ensemble average is performed with respect to the probability function representative of the final state, B, of the system. Thus, the free energy change is calculated directly from one MD simulation of the state B averaging the quantity $e^{\Delta H/k_B T}$. Usually, due to the known insufficient sampling of the tails of the distribution, this method gives accurate results when the energies of the initial and final states of the system differ by a relatively small amount ($\leq 2k_B T$). Otherwise, it is possible to decompose the total free energy change by defining intermediate states along a given path between the initial and final states, hence computing as a sum of partial free energy changes.

2.4.4 Potential of Mean Force

The difference in free energy between two states of a molecular system is a single number. Often we would like to know how the free energy of a system, or the *potential of mean force* (PMF), changes as a function of a particular coordinate within the system, most commonly a spatial coordinate. Chosen this coordinate, r, and considering the partial derivative of the free energy with respect to this coordinate, we obtain:

$$\frac{\partial A}{\partial r} = -k_B T \frac{1}{Q} \frac{\partial Q}{\partial r} = -k_B T \frac{1}{Q} \int \int -\frac{\partial U(\boldsymbol{q})}{\partial r} \frac{1}{k_B T} e^{-H(\boldsymbol{p},\boldsymbol{q})/k_B T} d\boldsymbol{p} d\boldsymbol{q} \tag{2.18}$$

Considering that $-\partial U(\boldsymbol{q})/\partial r$ is the force acting along r, $\boldsymbol{F}(r)$, and that the average value of a generic function, $f(\boldsymbol{p}, \boldsymbol{q})$, is given by:

$$\langle f(\boldsymbol{p}, \boldsymbol{q}) \rangle = \frac{1}{Q} \int \int f(\boldsymbol{p}, \boldsymbol{q}) e^{-H(\boldsymbol{p},\boldsymbol{q})/k_B T} d\boldsymbol{p} d\boldsymbol{q}, \tag{2.19}$$

Eqn. 2.18 becomes

$$\frac{\partial A}{\partial r} = -\langle \boldsymbol{F}(r) \rangle \tag{2.20}$$

Hence, if we are interested in the free energy change between two positions r_A and r_B, we get

$$A_B - A_A = \int_{r_A}^{r_B} -\langle \boldsymbol{F}(r) \rangle dr \tag{2.21}$$

Usually the ensemble average $-\langle \boldsymbol{F}(r) \rangle$ is most commonly determined from simulations at a series of r values between r_A and r_B and the integral in Eqn. 2.21 evaluated numerically.

CHAPTER

THREE

β-hairpin conformation of fibrillogenic peptides: structure and α-β transition mechanism revealed by molecular dynamics simulations

Summary

Understanding the conformational transitions that trigger the aggregation and amyloidogenesis of otherwise soluble peptides at atomic resolution is of fundamental relevance for the design of effective therapeutic agents against amyloid related disorders. In the present study the transition from ideal α-helical to β-hairpin conformations is revealed by long time scale, all atom molecular dynamics simulations in explicit water solvent, for two well known amyloidogenic peptides: the H1 peptide from prion protein and the Aβ(12–28) fragment from the Aβ(1-42) peptide responsible for Alzheimer disease. The simulations highlight the unfolding of α-helices, followed by the formation of bent conformations and a final convergence to ordered in register β-hairpin conformations. The β-hairpins observed, despite different sequences, exhibit a common dynamic behaviour and the presence of a peculiar pattern of the hydrophobic side chains, in particular in the region of the turns. These observations hint at a possible common aggregation mechanism for the onset of different amyloid diseases and a common mechanism in the transition to the β-hairpin structures. Furthermore the simulations presented herein evidence the stabilization of the α-helical conformations induced by the presence of an organic fluorinated cosolvent. The results of molecular dynamics in 2,2,2-trifluoroethanol (TFE)/water mixture provide a further evidence that the peptide coating effect of TFE molecules is responsible for the stabilization of the soluble helical

conformation.

3.1 Introduction

The incorrect folding of globular proteins is the result of amino acid mutation, chemical modification, environmental changes, or other unknown factors. The misfolded proteins are often degraded. In some cases, however, they aggregate and form amyloid fibrils, which are associated with some of the most distressing neurodegenerative deseases, [113] such as prion and Alzheimer's deseases. [114, 115] Many evidences suggest that these diseases are associated with an α to β conformational transition of part of the protein [26, 27] and that a small fragment of the protein plays a key role, as misfolding and aggregation precursor. [28, 29]

Prion deseases arise through a post-translational change to the so called prion protein, PrP, whose NMR structure was first resolved by Riek *et al.*,[116] and are characterized by the accumulation of an abnormal form of the prion protein, PrP^{Sc}, in the brain.[117, 118] Residues 109-122 (H1 peptide) are considered to be important for the α to β conformational transition and amyloid formation. According to several experimental evidences on the isolated H1 peptide of the normal cellular prion protein, PrP^{C}, it adopts in water β-sheet structure from which amyloid fibrils precipitate, [119, 120] it is able to induce the α-helix to β-sheet conformational transition of the isolated 129-141 fragment (H2 peptide) [119] and, as part of the synthetic fragment PrP(90-145), it can convert PrP^{C} to a PrP^{Sc}-like form. [121]

Similarly, Alzheimer's disease is the result of deposition in brain tissues of Aβ(1–42) peptides, a product in the amyloid protein metabolism. [115] Shorter and synthetic fragments of the Aβ–peptide (1–28, 25–35, 10–35 and 12–28) have been studied and characterized, in particular the Aβ(12-28) fragment, which was shown to have behavioral effects in mice, [122, 123] formation of fibril aggregates [124] and toxic effects in vitro. [125]

In 2,2,2-trifluoroethanol (TFE) or membrane mimicking environments both H1 and Aβ–(12-28) fragments were shown to adopt an α–helical conformation. [126, 27]

Unfortunately, the insoluble and massive character of the fibrils rules out the possibility to investigate their structure at atomic resolution with conventional experimental techniques, so that the β–structures of these fragments are not available and the mechanism of the conformational transition is largely unknown. In such cases, one has little choice but to turn to the use of theoretical approaches.

Several studies using molecular dynamics simulations on model peptides and aggregates have recently appeared in the literature. Levy and coworkers [127] observed the helix-coil transition of the slightly different PrP106-126 peptide performing a set of 34 MD simulations. Klimov and Thirumalai [128] showed via MD that the oligomerization of Aβ(16–22) requires the peptide to undergo a random coil to α-helix to β-strand transition. Straub and coworkers [129] used the computation of a variationally optimized dynamical trajectory connecting fixed end points of known structures to speed up the

conformational transitions among the different secondary structure motifs . On the aggregation side, Caflish and coworkers, [130] for instance, used a simplified implicit model based on the solvent accessible surface to describe the main solvent effects, to simulate the aggregation process of the heptapeptide GNNQQNY from the yeast prion protein . Other MD simulations with explicit representations of solvent were run on oligomers of the Alzheimer's peptides. The results indicate that Aβ(16–22) has a preference to aggregate in an extended conformation, forming antiparallel β-sheet structures. Longer fragments, instead, tend to aggregate as β–hairpins.[131] These studies however do not consider the dynamic mechanism leading to the formation of the β–hairpins.

In the present study we report, for the first time at atomic resolution, the spontaneous transition to β-hairpin conformations of the Syrian hamster PrP peptide H1 and of the Aβ(12-28) fragment obtained with long timescale, all atom, MD simulations in explicit water. The analysis of the trajectories helps define the common and peculiar pattern of the hydrophobic side chains in the β-hairpin conformation and the common α to β conformational transition mechanism of these two peptides, that despite having different sequences, give rise to analogous aggregation phenomena. This unbiased MD approach to study amyloid peptides was also tested by the use of a 30% (v/v) TFE/water mixture model of the solvent to check its ability to reproduce the stabilization of the helical conformation in membrane mimicking environments, by the TFE coating effect already noticed in a previos study. [132]

The convergence of the results is particularly significant as actually the peptide H1 and the Aβ(12-28) fragment, although simulated with the same MD package and force field, were studied independently from each other in two different laboratories.

3.2 Methods

3.2.1 MD simulations protocol.

MD simulations, in the NVT ensemble, with fixed bond lengths,[133] were performed with the GROMACS software package [134] and with the GROMOS96 force field. [35] The force field uses an explicit representation of acidic hydrogens and of hydrogen atoms on aromatic rings. Water was modeled by the simple point charge (SPC) model[135] and TFE by the Fioroni *et al.* model.[136] A twin range cut-off was used for the calculation of the non-bonded interactions. The short range cutoff radius was set to 0.8 nm and the long range cut-off radius to 1.4 nm for both Coulombic and Lennard-Jones interactions. The Berendsen algorithm [46] was used for the temperature control. The peptides, in their different starting conformations, were solvated with water or the TFE/water mixture and placed in a periodic truncated octahedron large enough to contain the peptide and $\approx$1.0 nm of solvent on all sides. In all the simulations of the H1 peptide, a negative counter ion, Cl^-, was added by replacing a water molecule at the most positive electrical

potential to achieve a neutral simulation cell. The side-chains were protonated as to reproduce a pH of about 7 and the N-terminal and C-terminal were amidated and acetilated respectively to reproduce the experimental conditions. [119] For the simulations of the Aβ(12-28) fragment no counterions needed to be added, since the total charge of the peptide resulted to be zero. The protonation of the side-chains and the N-terminal and C-terminal were consistent with the experimental pH = 5 condition used by Jarvet *et al.* [137] All the simulations, starting from the crystallographic structure, were equilibrated with 100 ps of MD runs with position restraints on the protein to allow relaxation of the solvent molecules. These first equilibration runs were followed by other 50 ps runs without position restraints on the protein. The temperature was gradually increased from 50 K to the chosen temperature performing short runs, of 50 ps each, every 50 K.

3.2.2 MD simulations of the H1 peptide.

Different all atom MD simulations in explicit water of the H1 peptide (MKHMAGAAAAGAVV) were carried out:
1) 450 ns of MD simulation at 300K starting from an ideal α-helix solvated with 5133 water molecules;
2) $\approx$21 ns of MD simulation in water at two different temperatures: $\approx$11 ns at 360 K followed by $\approx$11 ns at 300K. The starting structure was obtained by a clustering procedure performed on the first 200 ns of the previous simulation. The central structure of the most populated cluster resulted to be the one at 84 ns;
3) 35 ns of MD simulation at 300K in water, starting from a model configuration provided by low resolution X-ray diffraction data solvated with 6411 water molecules; [120]
4) 50 ns of MD simulation at 300K starting from an ideal α-helix in a mixture of 30% (v/v) TFE/water. Although experimentally 1,1,1-3,3,3-hexafluoropropan-2-ol (HFIP) was used as cosolvent, we preferred to perform simulation in TFE since its smaller size allow a larger computational efficiency. Furthermore, TFE is less effective as secondary structure stabilizer than HFIP used in the experiments, so we expect that the results of our simulations are further validate by this fact.
The Protein Data Bank (PDB) coordinates file of the low resolution X-ray structure was downloaded from http:
//www.mad-cow.org/~tom/prion_QuatStruc.html.

3.2.3 MD simulations of the Aβ(12-28) peptide.

The Aβ(12–28) fragment (VHHQKLVFFAEDVGSNK) was studied with five long timescale all atom MD simulations in explicit solvent:
1) two 100 ns long MD simulation in water at 295K and 320K (in order to speed up the phase space sampling), respectively, starting from an α-helical conformation, solvated with 3463 water molecules, taken from the PDB 1IYT.pdb file corresponding to the

whole Aβ(1–42) peptide;
2) two MD simulation in water, 30 ns at 295K and 20 ns at 320K, respectively, starting from an extended, all trans, conformation. This starting conformation was solvated with 11712 water molecules. The timespan of these two simulations was shorter than the first two due to the much higher number of solvating water molecules;
3) a 100 ns long MD simulation in water at 320K starting from a conformation representative of most populated cluster obtained from the statistical clustering of the conformations explored in the four simulations described above, which was solvated by 3757 water molecules;
4) a 50 ns long MD simulation of the same peptide at 300K, in a TFE/water mixture, starting from the same α-helical conformation of item 1.

3.2.4 Clustering procedure

Cluster analysis was performed using the Jarvis-Patrick method: [138] a structure is added to a cluster when this structure and a structure in the cluster have each other as neighbours and they have at least P neighbours in common. The neighbours of a structure are the M closest structures. In our case P is 3, M is 9.

3.3 Results

3.3.1 α-helix to β-hairpin transition of the H1 peptide.

A first simulation was performed at 300K for 450 ns starting from an ideal α-helix. The α-helix structure was completely lost after ≈10 ns and after ≈408 ns a transition to an ordered β-hairpin structure, stable for the remaining 50 ns, was observed [Figure 3.1(a)]. The β-hairpin structure [Figure 3.2(a)] was of the type 2:2 with a type II' β-turn sequence of (A113-)G114-A115(-A116) and was characterized by a shift in the β-sheet register of the inter-strand hydrogen bonds (HB) pattern, with an antiparallel bulge involving G119. Interestingly, the hydrophobic residues, and in particular alanines A113, A115 and A116 in the turn region were mostly exposed to the solvent, providing a possible seed for the aggregation process. Furthermore, the hydrophobic solvent–accessible surface area monitored in the last stages of the β-hairpin folding process, clearly showed a sharp transition from a state of compact bent conformations with buried hydrophobic side-chains, to the β-hairpin state with an increased water accessibility of the hydrophobic residues (Figure 3.3). The final average solvent accessible surface area value is around 9 nm^2, slightly lower than the value calculated for the α-helical starting conformation of 9.4 nm^2.

To speed up the sampling of the conformational space, a further simulation at higher temperature was performed starting from a configuration obtained by a cluster analysis [138] of the first 200 ns of the previous simulation. The temperature was initially set to 360K and after ≈11 ns quenched to 300K, when an almost complete β-hairpin

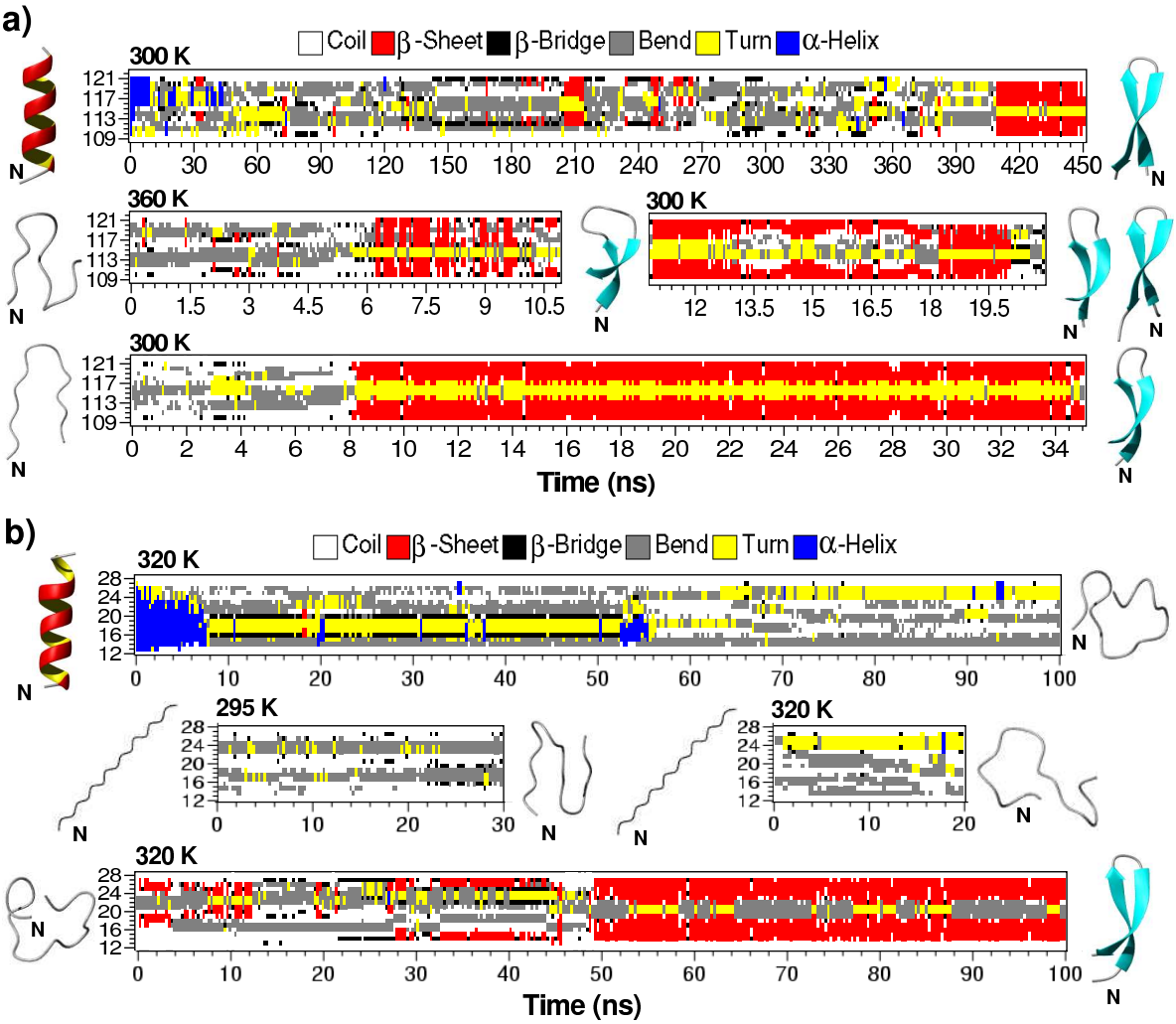


Figure 3.1: Time evolution of secondary structure. The analysis was performed with the DSSP program. [139] The starting and final structures are shown on the left and right sides, respectively. The N-terminal in each snapshot is indicated with "N". **a**) Time evolution of the H1 peptide secondary structure. Upper panel: MD simulation at 300K starting from an ideal α-helix. Note the formation of the β-hairpin at t≈408 ns. Middle panel: MD simulation at variable temperature starting from the central structure of the most populated cluster obtained from the first 200 ns of the previous simulation. Two β-hairpins are formed at t≈11 ns and t≈18 ns, respectively. Lower panel: MD simulation at 300K starting from the low resolution X-ray structure. The β-hairpin is formed at t≈8 ns. **b**) Time evolution of the Aβ(12-28) peptide secondary structure. Upper panel: MD simulation at 320K starting from an ideal α-helix. Middle panel: MD simulation at 295K and at 320K starting from an extended conformation. Lower panel: MD simulation at 320K starting from the representative structure of the previous simulations. Note the formation of the β-hairpin at t≈48.

structure was observed [Figure 3.1(a)]. In the last ≈10 ns, two different β-hairpins with an occurrence of 45% and 25%, respectively, were sampled: a 4:4 β-hairpin with a type

Figure 3.2: β-hairpin structures. **a**) Structure of the 2:2 β-hairpin, with a type II' β-turn, of the H1 peptide obtained in the 450 ns long MD simulation at 300K. Note that the alanines, in particular in the turn region (A113, A115 and A116), are exposed to the solvent. **b**) Structure of the 4:4 β-hairpin, with a type IV β-turn, of the H1 peptide obtained in the 35 ns long MD simulation at 300K, starting from the low resolution X-ray structure. Note that the alanines, in particular in the turn region (A115, A116 and A117), are exposed to the solvent. **c**) Structure of the 2:2 β-hairpin, with a type II' β-turn, of the Aβ(12-28) peptide obtained in the 100 ns long MD simulation at 320K. Note that the central hydrophobic residues (L17,V18,F19,F20,A21) are exposed to the solvent. The residues in italics belong to the turn.

VIII β-turn sequence of G114-A115-A116-A117 [Figure 3.2(b)] and a 2:2 β-hairpin with the same HB pattern and turn sequence observed in the previously reported simulation at 300K.

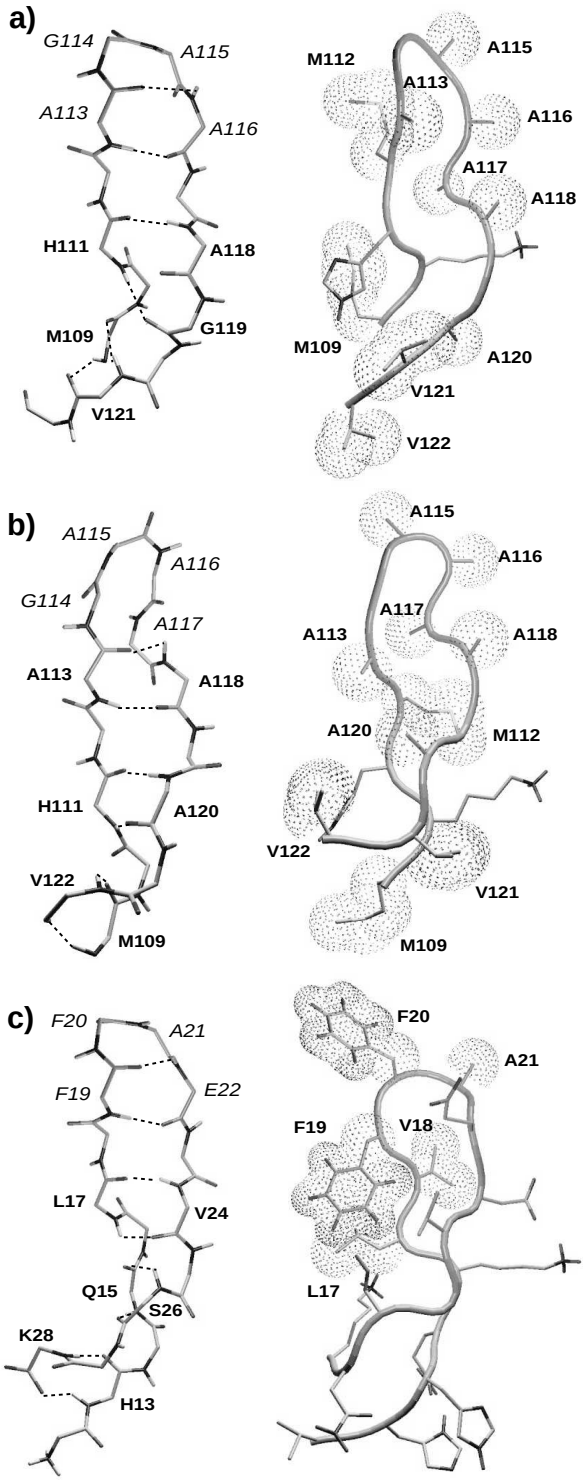
The only available experimental model, obtained by low resolution X-ray diffraction measurements on fibers, [120] suggested the presence of a β-bend with an intramolecular turn. Using this structure as starting point, we performed a 35 ns long MD simulation in water [Figure 3.1(a)]. After $\approx$ 8 ns the peptide adopted a very stable 4:4 β-hairpin conformation [Figure 3.2(b)], with the same HB pattern and β-turn sequence observed in the simulation at variable temperature. The only difference was in the β-turn type, being IV instead of VIII.

It has to be pointed out that the two types of β-hairpin observed in the different simulations (4:4 and 2:2) have the turn region surprisingly rich in alanines and a peculiar high solvent accessibility of the hydrophobic residues, hinting at a reasonable starting point for the aggregation process. A β-hairpin like conformation of the H1 peptide, in the scrapie form of the prion protein, was hypothesized by Prusiner and coworkers [140] and by Daggett and coworkers. [141]

3.3.2 α-helix to β-hairpin transition of the Aβ–(12-28) peptide.

The conformational evolution of the Aβ(12-28) peptide was investigated by long time scale simulations at two different temperatures, namely 295 and 320K, using different starting structures.

Two 100 ns long simulations at the two above referenced temperatures were started from the helical conformation. The α–helix was only marginally stable at both the temperatures and the peptide showed a high tendency to populate a compact bent conformation [Figure 3.1(b)]. This was stabilized by the formation of a salt bridge between K16 and E22 or D23, and by the packing of the side-chains of residues 17–21 (LVFFA), the central hydrophobic core (CHC). V24 also packed on this nascent hydrophobic patch.



Two additional simulations (30 ns at 295K and 20 ns at 320 K) were started from a completely extended structure to check the convergence to the same family of compact states as described above. In both simulations the peptide evolved into a compact conformational ensemble, characterized by the same features of the bent structure ob-

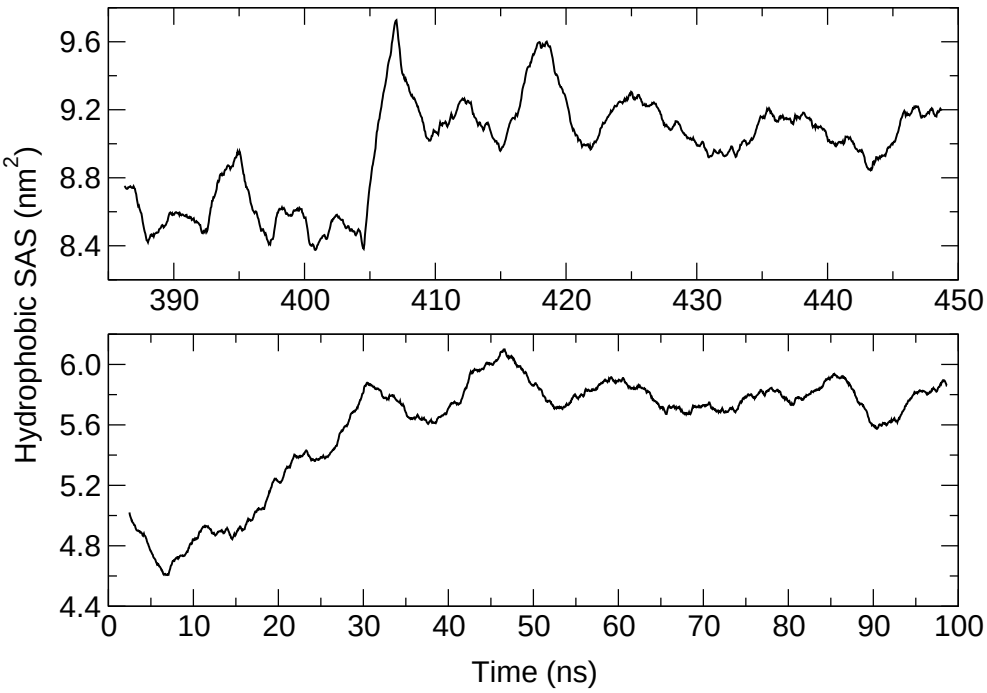


Figure 3.3: Hydrophobic Solvent Accessible Surface (SAS) area as function of time. **Upper panel**: the SAS variation corresponding to the conformational transition from bent to β-hairpin geometry at 300K for the H1 peptide. Note the increase of the hydrophobic SAS at t≈408 ns, corresponding to the α-β transition. **Lower panel**: β transition at 320K of the Aβ(12-28) peptide in the MD simulation at 320K. The starting structure in the figure corresponds to the representative structure obtained by a cluster analysis (see text). The maximum hydrophobic SAS is obtained in correspondence of the transition from disordered bent to ordered β-geometry at ≈48 ns. For the Aβ(12-28) peptide just the central hydrophobic residues are considered. The plots shown are results of box car averaging over a 10-ps window. General features are insensitive to the nature of this averaging.

tained from the two previously described simulations. A clustering procedure [138] was then applied to the four trajectories obtained, and the representative structure of the most populated cluster was isolated and used as a starting point for further MD analysis. This structure, characterized by the presence of a loop comprising residues 22–23 (E–D) and residues 12–21 and 24–28 in extended-bend conformation, was simulated for 100 ns at 320K. After ≈48 ns a sharp transition to a very ordered β-hairpin structure was observed [Figure 3.1(b)]. The β-hairpin structure [Figure 3.2(c)] was of the type 2:2 with a type II' β-turn sequence of F19-F20-A21-E22.

Very interestingly, the hydrophobic side chains of LVFFA sequence, as a consequence of both being consecutive in the sequence and of the formation of the turn, were mostly exposed to water and, consistent with the observations in the H1 peptide simulations,

an analogue increase in the hydrophobic SAS was observed [Figure 3.3], on going from the compact bent conformation to the ordered β-hairpin. The final average solvent accessible surface area value is around 5.8 nm^2 and is lower than the value calculated for the α-helical starting conformation of 8 nm^2. The final ordered structure of the β-hairpin is consistent with several experimentally based hypotheses on the conformation of the monomer in the fibrils. [29, 142, 137]

3.3.3 α-helix stabilization in TFE/water mixture.

Conformational studies of both the H1 peptide and Aβ(12-28) fragment have been performed in mixtures of fluorinated solvent and water that were shown to stabilize the helical conformations. [126, 27] To investigate this effect and test the simulations against those experimental data, two simulation in $\approx$ 30% TFE/water mixture were carried out. In Figure 3.4, the time percentage of α-helical conformation per residue for both the simulations is reported. A representative structure of each peptide is also shown. The H1 peptide retains the central core of the initial α-helix (residues 112-117) during all the simulation time. This result is in excellent agreement with the solid state NMR data obtained by Heller *et al.* [126] using the stronger helix stabilizer HFIP and by Satheeshkumar and Jayakumar [143] on the slightly different 113-127 peptide. In the case of the Aβ–(12-28) fragment, the helical conformation is retained in the region 13-17 and in region 21–24, while in the central part the structure has the tendency to bend, once more in agreement with the experimental observations. [27]

In a recent molecular dynamics study of peptide forming secondary structure in TFE/water mixture, [132] it has been shown that in a TFE/water mixture the organic cosolvent aggregates around the peptide forming a matrix that partly excludes water. This process in turn promotes the formation of local interactions and, as a consequence, stabilizes the folded structures. [144, 132] A similar coating effect is at the basis of the increased stability of helical conformations also in the cases examined herein. The average number of contacts of the TFE molecules with the peptide in the first 200 ps and in the last 5 ns are 191$\pm$30 and 349$\pm$33 for the H1 peptide and 293$\pm$33 and 525$\pm$42 for the Aβ(12-28) peptide, respectively. In both cases, an increase of $\approx$ 85% is observed, showing a clear propensity of TFE to coat the solute, in agreement with the previous observation by Roccatano *et al.* on Melittin peptide. [132]

3.4 Conclusions

Taken together, the results obtained from our totally unbiased simulations indicate an extremely high conformational flexibility for both peptides in water solution, with a general high tendency to form stable and ordered β–sheet structures. The observed β–hairpin conformations are characterized by an increased hydrophobic SAS area with respect to the compact bent conformation preceding the transition. In the ordered β-

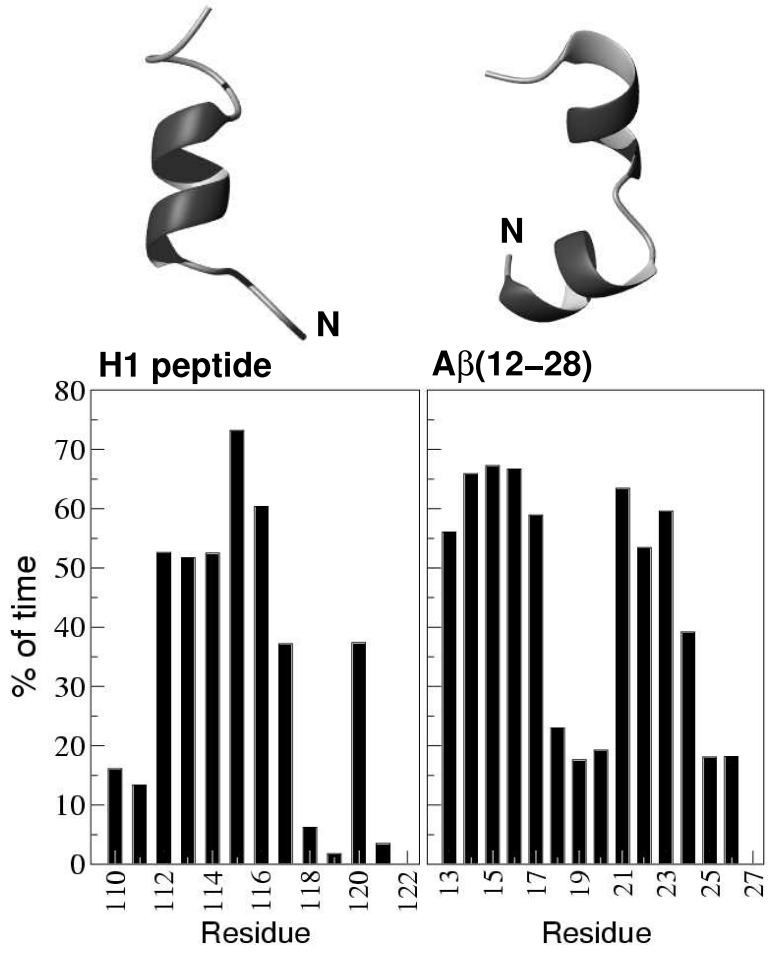


Figure 3.4: Time percentage of α-helical conformation per residue for the H1 peptide (left) and for the Aβ(12-28) peptide (right) in the TFE/water mixture simulations. A representative structure extracted at 50 ns for both peptides is also shown. The N-terminal in each snapshot is indicated with "N".

structure all of the hydrophobic side chains lie on the same plane around the turn region of the hairpin, and point into the same direction in 3D space. In contrast, in the starting α-helical conformations of both peptides, the hydrophobic side chains are "scattered" on different faces of the helix, and as a result of geometrical and sequence constraints, they point in different directions. This is a new feature in the β-hairpin conformations, as the structures so far obtained for other peptides show large intramolecular hydrophobic interactions [145] with a clear tendency to remove hydrophobic side-chains from water contact in order to avoid aggregation phenomena. Although the presence of five consecutive hydrophobic residues (LVFFA) in Aβ peptide and the unusual hydrophobic patch in H1 are not well represented in protein sequences, analogous highly hydrophobic sequences in α–helical geometries are also present in some 'nonamyloidogenic' proteins known to aggregate in amyloidogenic conditions, such as in the case of lysozime, [146]

or myoglobin. [147] Thus, it is conceivable that these very peculiar sequences and the conformational transitions into the stable β–hairpin geometry conspire to expose part of the hydrophobic cores of the two peptides. This type of structure can be considered as highly frustrated and offers a clear starting point for aggregation. Despite showing a lower or comparable global hydrophobic SAS area with respect to the α-helical conformation, the steric properties and the ordered directionality of the exposed patches in the β-hairpin conformations may define a possible ordered hydrophobic interaction area with other molecules sharing the same structural features. This sort of preorganized interaction area is absent in the helical conformation, and these observations can theoretically support the observation of high percentages of β-structures in experimental studies. [26, 28, 137, 142]

More specific interactions, such as coulombic interactions, in addition to the hydrophobic collapse, should be considered necessary for the subsequent ordering of the nascent fibrillar aggregates, as shown in the case of experimentally studied small peptide models. [148]

The transition from α–helical to β structure requires the peptides to populate intermediate β–bend geometries in which several mainly hydrophobic interactions are partially formed. This is followed by the sudden collapse to ordered β–hairpin structures and the simultaneous disruption of the hydrophobic side-chain interactions with a consequent increase in the solvent exposure. For both H1 and Aβ(12-28) peptides the atomic picture of the detailed mechanism of the evolution from α to β, provided in this work, can be very useful for the design of new constrained sequences or new drug candidates.

Finally, the simulations in the TFE/water mixture evidence the stability of α-helical conformations in the presence of the fluorinated cosolvent, resulting in excellent agreement with the available experimental data. Furthermore, the analysis of the TFE distribution around the peptide confirms the mechanism of TFE stabilization proposed by Roccatano *et al.* [132] on different secondary structure forming peptides.

Acknowledgments

This work was supported by a grant from the European Community Training and Mobility of Researchers Program "Protein (mis)folding" and by the Italian National Research Council. We thank Prof.s Maurizio Brunori, Giacomo Carrea and Martin Zacharias for carefully reading the manuscript and for stimulating discussions.

CHAPTER

FOUR

Thermodynamic and kinetic characterization of a β-hairpin peptide in solution: the complete phase space sampling by molecular dynamics simulations in explicit water

Summary

The folding of the amyloidogenic H1 peptide MKHMAGAAAAGAVV taken from the syrian hamster prion protein is explored in explicit aqueous solution at 300K using long time scale all-atom molecular dynamics simulations for a total simulation time of 1.1 μs. The system, initially modeled as an α-helix, preferentially adopts a β-hairpin structure and several unfolding/refolding events are observed, yielding a very short average β-hairpin folding time of $\approx$200 ns. The long time scale accessed by our simulations and the reversibility of the folding allow to properly explore the configurational space of the peptide in solution. The free energy profile, as a function of the principal components (essential eigenvectors) of motion, describing the main conformational transitions, shows the characteristic features of a funneled landscape, with a downhill surface toward the β-hairpin folded basin. However, the analysis of the peptide thermodynamic stability, reveals that the β-hairpin in solution is rather unstable. These results are in good agreement with several experimental evidences, according to which the isolated H1 peptide adopts very rapidly in water β-sheet structure leading to amyloid fibril precipitates [Nguyen *et al.*, Biochemistry 34:4186-4192, 1995; Inouye *et al.*, J. Struct. Biol. 122:247-255, 1998]. Moreover, in this study we also characterize the diffusion behaviour in conformational space, investigating its relations with folding/unfolding conditions.

4.1 Introduction

The most stable fold of a protein is determined by its amino acid composition, solvent environment (composition, pH, ionic strength) and physical state (temperature, pressure). Considering that interactions at atomic level play a crucial role in the equilibrium between folded and unfolded conformers, molecular dynamics simulations could in principle be used to calculate the folded/unfolded equilibrium and could yield the kinetics of the folding process. However, given the high computational cost required for the complete sampling of the protein-peptide configurational space, MD simulation in atomic detail of the folding/unfolding equilibrium was not in practice considered as a possible investigation tool. For this reason in folding studies, the molecular models used were often of a simple nature: one interaction site per residue,[149, 150] implicit solvent approximation,[76–79] motions restricted to lattice sites,[60, 61] etc. On the other hand, other methods were developed to enhance the configurational space sampling in atomistic simulations such as essential dynamics sampling,[151] highly parallel simulation algorithms (PRD,REMD), [93, 152, 153] or other generalized-ensemble methods.[154] However, the direct MD simulation of the folding/unfolding equilibrium in the canonical ensemble would be the most reliable procedure to obtain both thermodynamic and kinetic properties.

Only recently, with the aid of high power computers, all atoms MD simulations, in explicit water, provided the folding of peptides into α-helix [17] or very short β structures.[18] In the previous study[155] (chapter 3) the more complex folding of a 14 residue peptide (the prion protein H1 peptide) into a in-register β-hairpin conformation starting from an ideal α-helix has been achieved. The syrian hamster prion protein residues 109-122 (H1 peptide) is considered to be important for the α to β conformational transition that leads to amyloid formation and is responsible for prion diseases. According to several experimental evidences on the isolated H1 peptide, it adopts very rapidly in water β-sheet structure from which amyloid fibrils precipitate,[119, 120] while in 2,2,2-trifluoroethanol (TFE) or membrane mimicking environments the H1 peptide adopts an α–helical conformation.[126, 27] These properties make the study of this peptide very interesting and may provide a key for understanding protein folding or the cause of amyloid diseases.

In the present chapter further simulations of the H1 peptide at physiological conditions have been performed to obtain a complete description of its conformational free energy landscape, including the folding/unfolding equilibrium, by means of long time scale (1.1 μs) all atoms MD simulations in explicit water. At our knowledge this is one of the first attempt to simulate the thermodynamic equilibrium of a complex system, such as a β-hairpin, for more than 1 μs using realistic models for both the peptide and the solvent and with a completely unbiased sampling.

Finally, in the present study we also investigate in details the diffusion behaviour in conformational space, relating its properties with folding/unfolding conditions. Results show a characteristic dual diffusion regime, observed previously in small proteins,[156]

which can be utilized to better understand the kinetics of conformational transitions.

4.2 Methods

4.2.1 MD simulations protocol.

MD simulations, in the NVT ensemble, with fixed bond lengths [133] and a time step of 2 fs for numerical integration were performed with the GROMACS software package [134] and with the GROMOS96 force field. [35] Water was modeled by the simple point charge (SPC) model. [135] A non-bond pairlist cutoff of 9.0 Å was used and the pairlist was updated every 4 time steps. The long-range electrostatic interactions were treated with the particle mesh Ewald method [157] using a grid with a spacing of 0.12 nm combined with a fourth-order B-spline interpolation to compute the potential and forces in between grid points. The isokinetic temperature coupling [158] was used to keep the temperature constant at 300 K. The peptide, in its different starting conformations, was solvated with water and placed in a periodic truncated octahedron large enough to contain the peptide and ≈1.0 nm of solvent on all sides. In all the simulations a negative counter ion, Cl^-, was added by replacing a water molecule to achieve a neutral condition. The side-chains were protonated as to reproduce a pH of about 7 and the N-terminal and C-terminal were amidated and acetilated respectively to reproduce the experimental conditions.[119]

Two all atom MD simulations in explicit water at 300 K of the H1 peptide (MKHMAGAAAAGAVV), for a total of ≈1.1 μs of simulation time, were carried out:
1) 240 ns starting from the α-helix conformation obtained from the simulation in 30% (v/v) TFE/water mixture of the previous work[155] (chapter 3);
2) 850 ns starting from the β-hairpin conformation observed in the previous simulation, using a new set of initial velocities.

4.2.2 Essential Dynamics analysis

The principles of the ED analysis are described in detail elsewhere.[51, 104] Briefly, from all the structures of both simulations a covariance matrix of positional fluctuations (C_α only) was built and diagonalized. Sorting the eigenvectors by the size of the eigenvalues shows that the configurational space can be divided in a low dimensional (essential) subspace in which most of the positional fluctuations are confined, and a high dimensional (near-constraints) subspace in which merely small vibrations occur. In Figure 4.1 the eigenvalues obtained from Cα coordinates covariance matrix are reported as a function of eigenvectors index and are ordered in descending order of magnitude. The corresponding relative cumulative positional fluctuation (with respect to the total positional fluctuation) is given in the inset. We used the two principal components with the highest eigenvalues, defining the first "essential plane", for thermodynamic

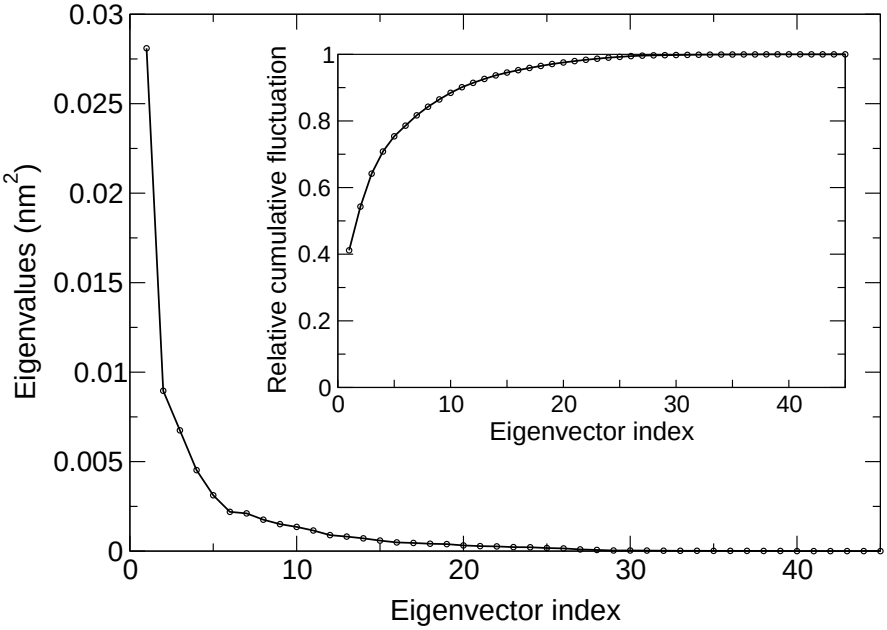


Figure 4.1: Eigenvalues, in decreasing order of magnitude, obtained from Cα coordinates covariance matrix as a function of eigenvectors index. The corresponding relative cumulative positional fluctuation is given in the inset.

and kinetic calculations. This is because such a plane accounts for almost 60% of the overall positional fluctuation (inset of Figure 4.1), hence describing the most relevant conformational degrees of freedom and the main conformational transitions of the peptide backbone. We performed similar calculations over planes defined by other eigenvectors. However, within such planes some of the relevant conformational transitions are not detectable and hence the corresponding eigenvectors are not suitable as conformational coordinates to describe the large conformatinal fluctuations as well as the folding/unfolding transitions.

4.2.3 Thermodynamic properties

Given a system in thermodynamic equilibrium, the change in free energy on going from a reference state, *ref*, of the system to a generic state, i, (e.g., from unfolded to folded) at constant temperature and constant volume can be evaluated as

$$\Delta A_{ref \to i} = -RT \ln \frac{p_i}{p_{ref}} \tag{4.1}$$

where R is the ideal gas constant, T is the temperature and p_i and p_{ref} are the probabilities of finding the system in state i and state *ref*, respectively. We will describe the free energy surface as a function of principal components (essential eigenvectors) from ED analysis. Structures sampled every 1 ps were projected onto the plane defined by

the two first principal components. A grid 20x20 has been used to divide this plane in 400 cells and for every cell the number of points were counted and the relative probability was calculated. Finally the $\Delta A_{ref \to i}$ was evaluated. We chose as the reference state the grid cell with the highest probability, i.e. the cell corresponding to the β-hairpin folded structures ensemble. Surfaces of the total (peptide + solvent) internal energy changes, $\Delta U_{ref \to i}$, and entropy changes, $\Delta S_{ref \to i}$, were calculated as well, via the average internal energy of the simulation box in each cell and

$$\Delta S_{ref \to i} = \frac{\Delta U_{ref \to i} - \Delta A_{ref \to i}}{T} \tag{4.2}$$

To evaluate the local stability of the secondary structure elements we calculated for every grid cell the ratio between the number of folded structures (β-hairpin) and the number of the unfolded ones; using equation 4.1 the ΔA of secondary structure formation, $\Delta A(formation)$, was then evaluated for every position in the essential plane.

In order to check the effect of different grid spacing on the thermodynamic properties, the same type of free energy landscapes were constructed using different grids, 10x10, 20x20 and 30x30 (data not shown). Interestingly, all the different grids provided similar free energy landscapes with the same free energy maximum variation (≈14 kJ/mol), the surface being slightly more corrugated on going from the grid with a lower cell density (10x10) to the more dense one (30x30).

4.2.4 Kinetic properties

For the study of diffusion properties, we chose the subspace defined by the first two essential coordinates. In particular different regions of the essential plane, where the coordinates do not encounter a relevant free energy gradient, were analyzed separately. To generate an ensemble of independent trajectories we used all the trajectory fragments starting within one of the selected regions and the corresponding ensemble mean square displacement, from each initial point as a function of time, was evaluated. In order to increase the statistics we averaged such a property over the first two essential degrees freedom, assuming for both a similar diffusion behaviour. All the curve fits are obtained using the graphing tool Xmgrace (http://plasma–gate.weizmann.ac.il/Grace/doc/UsersGuide.html), which makes use of the Levenberg Marquardt algorithm and provides χ^2 and correlation coefficient evaluations. Moreover, we also evaluated the noise (standard deviations, σ) for the model parameters, obtained by fitting simulation data, calculating their standard deviations over n subsets of trajectories and then extrapolating for the complete statistical sample:

$$\sigma = \left(\frac{\sum_{i=1}^{n} (a_i - \bar{a})^2}{(n-1)n} \right)^{1/2} \tag{4.3}$$

$$\bar{a} = \frac{\sum_{i=1}^{n} a_i}{n} \tag{4.4}$$

where a_i is the generic parameter evaluated in the ith subset. Note that the previous equation is based on the approximation that the parameters obtained by the whole number of trajectories are equivalent to the ones obtained averaging the corresponding values over the n subsets. In the present case we used 3 independent subsets which resulted to be a good compromise between the statistics within each subset and the sample size used in the last equation, given by the number of subsets.

4.3 Results

4.3.1 Thermodynamic characterization of the conformational transitions.

Two all atom MD simulations of the syrian hamster H1 peptide, for a total simulation time of ≈1.1 μs, were carried out. In Figure 4.2 and Figure 4.3(a) the root mean square deviation (RMSD), with respect to the β-hairpin structure, and the time evolution of the secondary structure are reported, respectively.

Within the first 0.24 μs of simulation the α-helix structure, used as the initial simulation structure, is rapidly lost and interestingly, after ≈0.20 μs, a β-hairpin structure is formed, with the same structural properties of the one observed in the previous work (chapter 3).[155] Further 0.85 μs of simulation were performed in the same conditions starting from the β-hairpin conformation. Many unfolding/refolding events of the β-hairpin are observed, with an average folding time of ≈200 ns, ensuring the reversibility of the folding of this peptide in the conditions used. The long time scale accessed by our simulations and the reversibility of the folding allow to properly explore the configurational space of the 14 residues peptide at physiological conditions.

α-helix and β-hairpin structures are populated for ≈5% and ≈30% of the total time, respectively. The rest is populated by partial folded β-hairpins, unfolded or "molten globule" like structures.

In Figure 4.3(b) the free energy surface as a function of the two first essential components (see methods) is reported. This free energy profile, obtained by the probability per grid cell as described in the methods section, shows a characteristic funneled landscape (i.e. a surface characterized by a single deep minimum) with a downhill free energy change toward the β-hairpin basin of ≈-14 kJ/mol. Such a free energy surface was obtained projecting the complete 1.1 μs simulation trajectory. To estimate its reliability we tested the convergence of the free energies within grid cells of the plane (note that free energy values are defined with respect to the grid cell corresponding to the global minimum). The results show rather stable values (within 0.1-0.2 kJ/mol) after about 0.3 μs for grid cells close to the free energy minimum. In grid cells located far from this region a worse convergence is observed, although after about 0.3 μs ΔA values are obtained within a noise of about 1.5 kJ/mol. In Figure 4.4(a) we report such a convergence plot for two given grid cells belonging to the previously mentioned subspaces,

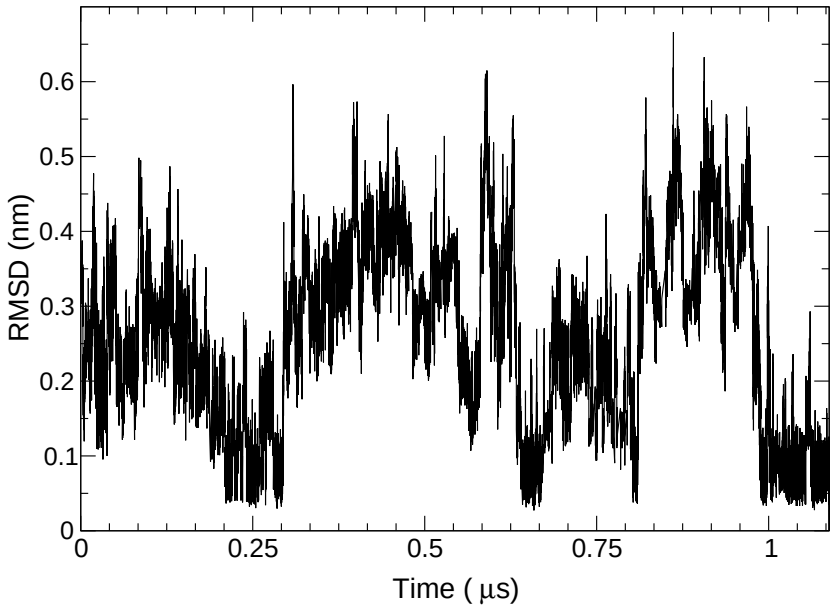
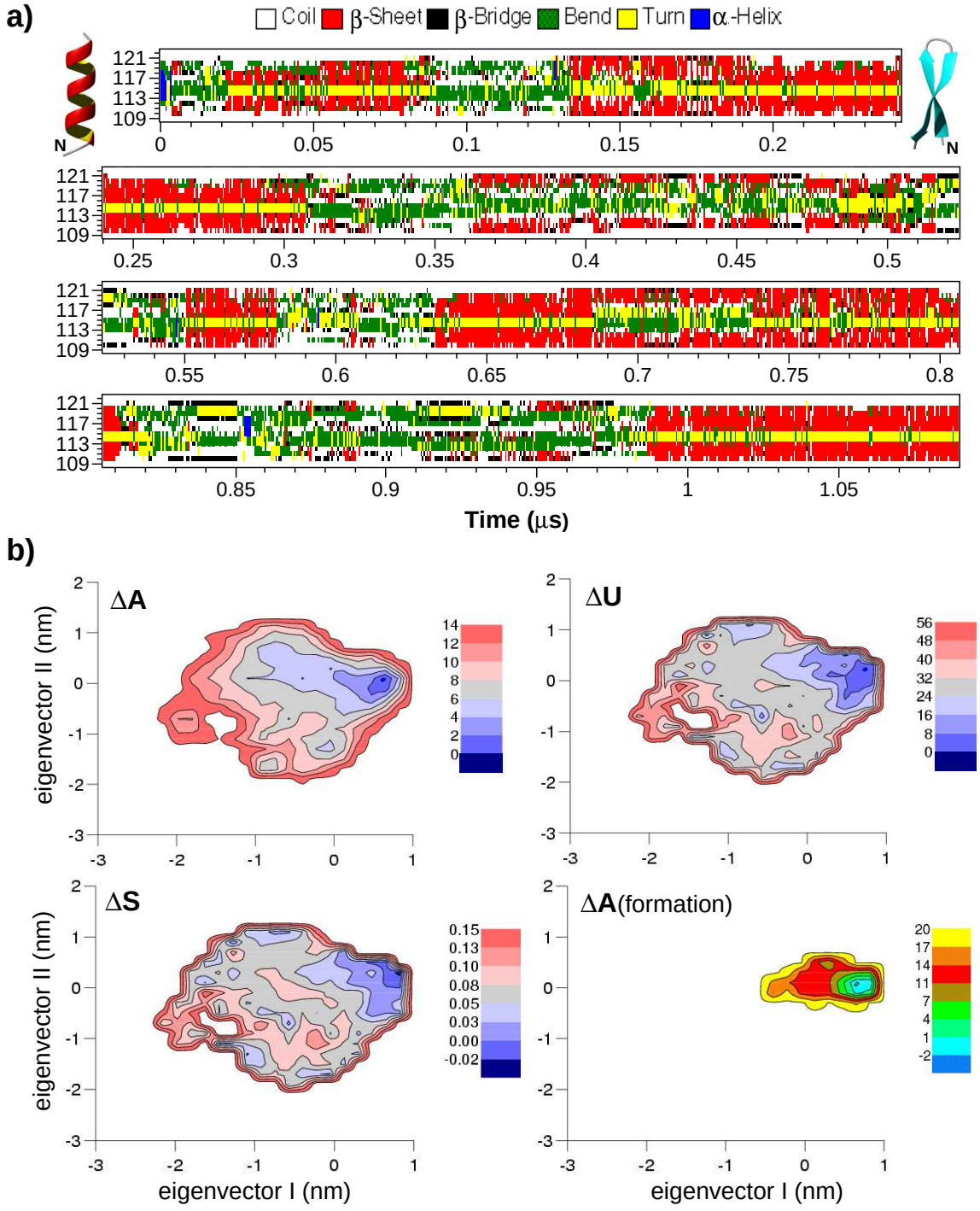


Figure 4.2: Root mean square deviation (RMSD) of the backbone atoms with respect to the β-hairpin structure vs time.

i.e. within the contour lines ΔA=0 kJ/mol and ΔA ≈3 kJ/mol of Figure 4.3(b). In Figure 4.4(b) we report the probability distribution of the free energy standard deviations, $\sigma_{\Delta A}$, over the grid cells utilized, showing rather small statistical errors affecting the free energy values. It is worth noting that the level of convergence observed for the free energy variations considered and the very limited corresponding statistical errors, do not necessarily mean that all the possible conformational transitions are sampled with the same accuracy.

The absolute free energy minimum, as well as the adjacent region within the contour line at ΔA≈2 kJ/mol in Figure 4.3(b), is mainly populated by β-hairpin structures, including the complete β-hairpin conformation also observed in the previous chapter.[155] Such a structure corresponds to a 2:2 β-hairpin with a type II' β-turn sequence of (A113-)G114-A115(-A116) and is characterized by 6 inter-strand hydrogen bonds (HB), with an antiparallel bulge involving G119 [Figure 4.5(a)]. A free energy plateau, within the contour line at ΔA ≈6 kJ/mol, is characterized by an ensemble of either completely unfolded or partial β-hairpin structures which only rarely evolve into the complete β-hairpin. Such partial β-hairpin structures mainly involve two types of structured conditions: either they share the same turn of the complete β-hairpin structure, but with flanking terminals (i.e. some HB are lost), or they have a different turn type. Note that among the latter, a 4:4 β-hairpin with a type IV β-turn sequence of G114-A115-A116-A117 [Figure 4.5(b)], was already observed in the simulations of the previous work (chapter 3).[155] Three "molten globule" like states are present with free energy local minima at ≈6, 8 and 10 kJ/mol respectively and are characterized by bent



conformations. A representative structure is given in Figure 4.5(c). The rest of the accessible essential subspace corresponds basically to completely unfolded structures.

α-helix structures [Figure 4.5(d)] are sampled 7 times throughout the simulations but each time for a very short period, about 500 ps. α-helix conformers do not populate any free energy minimum and are rather "disperse" through the gray plateau, within

Figure 4.3: **a)**: Time evolution of the H1 peptide secondary structure. The starting and final structures are shown on the left and right sides, respectively. The N-terminal in each snapshot is indicated with "N". In the first panel the first part of the simulation, starting from an ideal α-helix, is reported. The formation of the 2:2 β-hairpin at t$\approx$0.18 μs can be observed. In the following panels the second part of the simulation, starting from the β-hairpin structure with a new set of velocities, is reported. The analysis of the secondary structures was performed with the DSSP program.[139] **b)**: Contour maps of the free energy, ΔA, internal energy, ΔU, entropy, ΔS, and free energy change associated to the β-hairpin formation, $\Delta A(formation)$, as a function of the position in the essential plane. ΔA, ΔU and ΔS are calculated with respect to the state with the highest probability, i.e. the one corresponding to the β-hairpin folded structures ensemble. Energy and entropy values are given in kJ/mol and kJ mol^{-1} K^{-1}, respectively.

the contour line at ΔA$\approx$8 kJ/mol, with rather high free energies.

ΔU and ΔS profiles share the same funneled like shape of the free energy [see Figure 4.3(b)]. Note that the internal energy and the entropy values are calculated for the whole system, i.e. peptide and solvent. Interestingly, the absolute free energy minimum region (the subspace inside the contour line at ΔA$\approx$2 kJ/mol) includes the absolute internal energy and entropy minima, thus meaning that this state is the most energetically stable, with the lowest entropy.

To evaluate the local stability of the complete β-hairpin structure, we used the same essential plane to evaluate the free energy change, $\Delta A(formation)$, associated to its formation from any other possible structure [Figure 4.3(b)]. This was accomplished for every position (grid cell) of the essential plane, calculating the probability for the complete β-hairpin (p_β) and for any other possible structure (p) to occur which was then used to obtain the β-hairpin formation free energy $\Delta A(formation) = -RT \ln(p_\beta/p)$. Interestingly, except for a small region corresponding to the absolute free energy minimum, β-hairpin formation free energies are always positive, thus revealing that the H1 peptide has a rather unstable secondary structure.

In Figure 4.4(c) we also show the convergence of such a free energy change for two grid cells, clearly showing that also for this evaluation 1.1 μs is enough to obtain reliable results. A similar evaluation for the α-helix structure is not really possible because of its rare occurrence (5% of the total simulation time). However its very high approximate free energy shows, as expected, that the α-helix structure is very unstable (data not shown), although a relatively high number of α-helix unfolding/refolding transitions (7 times) was observed.

Finally we evaluated the global β-hairpin formation free energy, i.e. over the whole accessible conformational space. This was accomplished considering the essential plane as a unique cell and evaluating the corresponding probabilities, p_β and p, from which

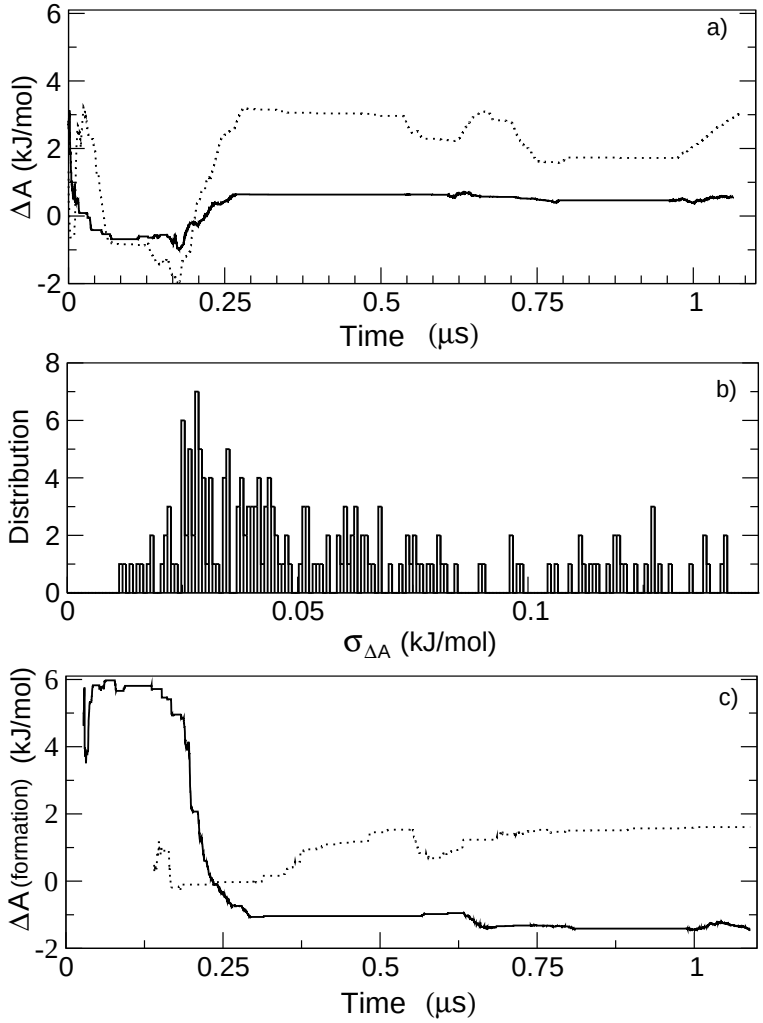


Figure 4.4: Time convergence of the ΔA [panel **a)**] and $\Delta A(formation)$ [panel **c)**] for two given essential plane positions. We chose a grid cell in the free energy minimum region, i.e. within the contour line at ΔA=0 of Fig. 3(b) (solid line), and a grid cell within the contour line at ΔA ≈3 of Figure 4.3(b) (dashed line). In panel **b)** the probability distribution of the free energy standard deviations, $\sigma_{\Delta A}$, for all the cells is reported.

the ΔA was then obtained. ΔU and ΔS were calculated as described in the method section. It is interesting to note that the global β-hairpin formation free energy obtained, ΔA≈+2.5 kJ/mol, shows that the "folded structure" is not the thermodynamic most stable condition for this peptide in water. Such a feature is due to the entropy decrease (≈-0.070 kJ mol^{-1} K^{-1}) which overcompensates the internal energy stabilization (≈-18.7 kJ/mol). Even in the absolute free energy minimum, where the β-hairpin structure is mostly stable, its formation free energy, $\Delta A(formation)$, is only about -2 kJ/mol [Figure 4.3(b)].

Our results are consistent with experimental data on a non-amyloidogenic 16 resi-

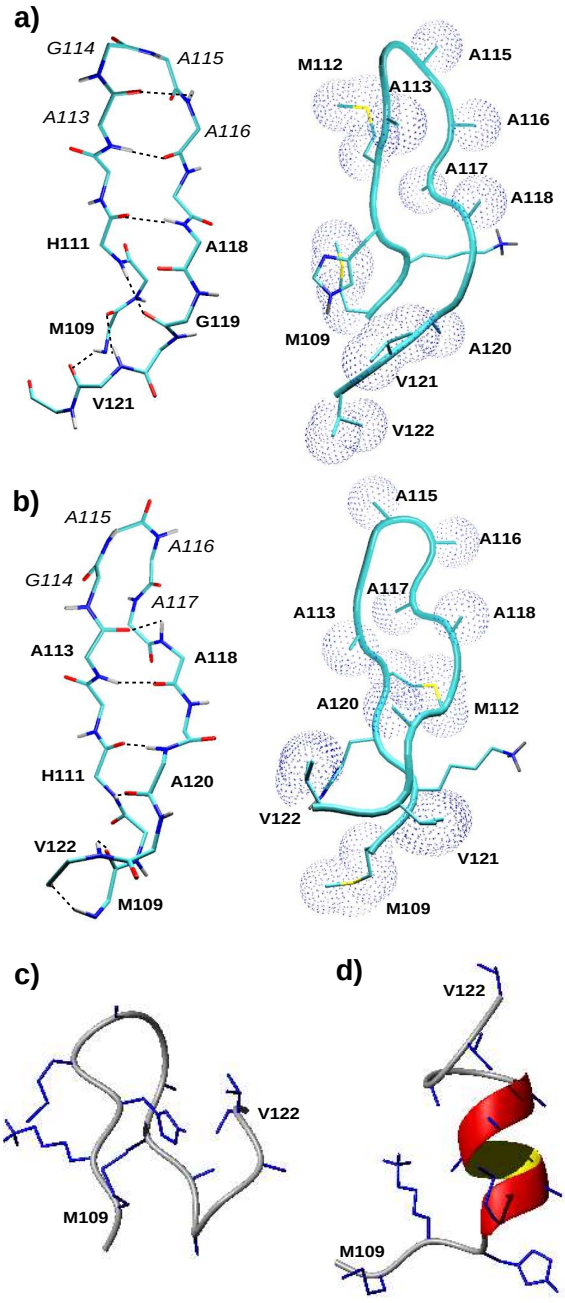


Figure 4.5: Structures of the H1 peptide observed along the simulations. **a**): a 2:2 β-hairpin with a type II' β-turn; **b**): a 4:4 β-hairpin with a type IV β-turn. Note that in the β-hairpins the alanines, in particular in the turn regions, are exposed to the solvent; **c**) a representative "molten globule" like structure; **d**) a representative α-helix structure.

dues β-hairpin peptide using a nanosecond laser temperature jump technique.[159] These experimental data provided an apparent ΔG for the β-hairpin folding transition of $\approx$-2.5 kJ/mol, a value close to our estimate for the H1 peptide, although the latter is $\approx$5 kJ/mol less stable. This relative instability could explain the amyloidogenic nature of

the H1 peptide.

4.3.2 Kinetic characterization of the conformational transitions.

In the previous subsection we characterized the thermodynamics within the conformational space of the H1 peptide. In this subsection we characterize its kinetics in the essential plane used for the previous thermodynamic analysis. Note that the essential eigenvectors defining such a plane represent the most relevant conformational degrees of freedom of the peptide backbone, hence describing the main conformational transitions. In a previous paper [156] it was shown that the kinetics of the essential degrees of freedom in proteins can be described by a diffusion behaviour characterized by a dual regime: a fast type of diffusion within a single energy local minimum, switching exponentially to a slower one, probably corresponding to "hopping" between multiple harmonic wells. Such a dual diffusion behaviour should be determined by the relaxation of the medium (defined by all the other coordinates) associated to the hopping and resulting into an increase of viscosity.

In the previous paper [156] we characterized the diffusion using relatively short time intervals (up to 20 ps). In this study, due to the huge simulation time available, we can afford a better statistical characterization of the diffusion over essential degrees of freedom, extending our investigation over longer times (up to 100 ps). However, a single-exponential relaxation of the velocity autocorrelation function, that was utilized to describe such a conformational diffusion in the previous study,[156] is not really suitable to describe this process over longer time intervals (up to 100 ps) afforded in the present study. A more accurate model can be obtained considering two relaxation modes of the velocity autocorrelation function (corresponding to a bi-exponential switching from the fast to the slow diffusion-regime), which can be considered as a generalization of the previous model, see Appendix. The equation obtained from this generalized model, for the mean square displacement, neglecting the initial fast (within a few tens of fs) relaxation, is for a given q (essential) degree of freedom

$$\begin{aligned} \left\langle \Delta q^2(t) \right\rangle &\cong 2D_\infty t + 2[D_0 - A_1]\tau_1[1 - e^{-t/\tau_1}] \\ &+ 2[D_0 - A_2]\tau_2[1 - e^{-t/\tau_2}] \end{aligned} \tag{4.5}$$

where D_∞ is the long-time diffusion constant, D_0 the short-time diffusion constant, τ_1, τ_2 the "relaxation times" of the two switching modes and A_1, A_2 two parameters defined by the integral of the velocity autocorrelation function, see Appendix.

Such a model was used to describe the diffusion in the configurational subspace defined by the first two C_α essential degrees of freedom, assuming at least for the "relaxation times" the same behaviour (this was actually checked to be a good approximation). In order to increase the statistics, we averaged the mean square displacement

over the first two essential degrees of freedom.

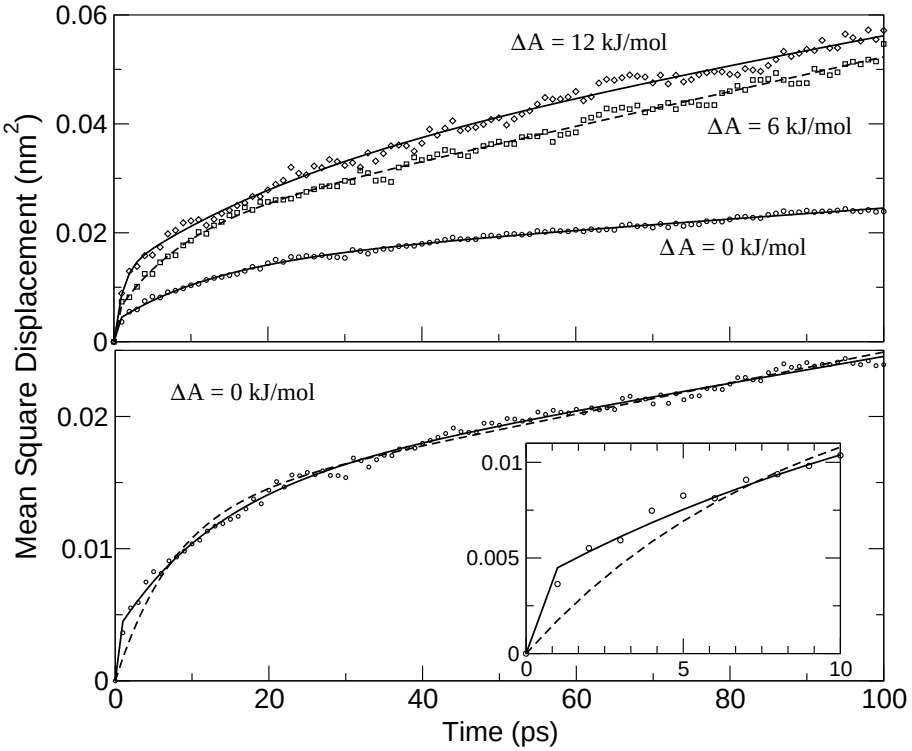


Figure 4.6: **Upper panel**: Mean Square Displacement, as a function of time, averaged over the first two principal eigenvectors. The theoretical models (solid line), derived in the Appendix, were parameterized fitting the simulation data. In particular we report results for three selected regions of the essential plane, one in the region of the free energy minimum (ΔA=0 kJ/mol) (circles), another in the region of the free energy plateau within the contour line at ΔA$\approx$6 kJ/mol (squares) and the third in the completely unfolded region (ΔA$\approx$12 kJ/mol) (diamonds). **Lower panel**: a comparison between the previous model, based on a single-exponential relaxation (dotted line), and the present one, based on a bi-exponential relaxation (solid line), is shown for one of the regions (ΔA=0 kJ/mol). In the inset the first 10 ps are shown in more details.

In Figure 4.6 we show the comparison between the theoretical models and the ensemble mean square displacements obtained by simulations. In particular we report the results obtained for three selected regions of the essential plane, one in the region of the free energy minimum (ΔA=0 kJ/mol), another in the region of the free energy plateau within the contour line at ΔA$\approx$6 kJ/mol and the third in the completely unfolded region (ΔA$\approx$12 kJ/mol). The plot reported in the upper panel clearly shows the high accuracy of the model used in the whole time range. Note that for all the three theoretical models the χ^2 values are in the range $10^{-5} - 10^{-4}$ with correlation coefficients always higher than 0.997 and full fitting convergence was achieved within 500 steps. In the lower panel a comparison between the simpler model, as used in the previous paper,[156] and the present generalized one is shown for one of the regions (ΔA=0 kJ/mol). It

is evident how the bi-exponential relaxation of the velocity autocorrelation function provides a more accurate model. In particular a dramatic improvement is observed in the first few ps of diffusion (see inset of Figure 4.6).

The diffusion constants and the "relaxation times" are reported in Table 4.1. While the short-time diffusion constants, D_0, are of the same order of magnitude for all the regions, revealing a similar diffusion behaviour within a single local energy well, the long-time diffusion constants, D_∞, are similar for the less structured regions, but it is significantly lower in the β-hairpin region. Thus, when the system enters its long-time diffusion regime, hopping between multiple energy basins, the more structured state encounters a greater viscosity of the medium defined by the other coordinates including the solvent.

Table 4.1: **Diffusion constants and "relaxation times" for three selected regions of the essential plane, one in the region of the free energy minimum (ΔA=0 kJ/mol), another in the region of the free energy plateau within the contour line at ΔA$\approx$6 kJ/mol and the third in the completely unfolded region (ΔA$\approx$12 kJ/mol).**

region	${D_0}^*$ nm^2ps^{-1}	${D_\infty}^*$ nm^2ps^{-1}	${\tau_1}^*$ ps	${\tau_2}^*$ ps
ΔA=0 kJ/mol	0.026 (0.001)	$5.3{\cdot}10^{-5}(0.4{\cdot}10^{-5})$	< 1	13.2 (2.4)
ΔA$\approx$6 kJ/mol	0.032 (0.001)	$15.9\cdot10^{-5}(1.5{\cdot}10^{-5})$	< 1	8.5 (1.1)
ΔA$\approx$12 kJ/mol	0.032 (0.001)	$12.8{\cdot}10^{-5}(1.9{\cdot}10^{-5})$	$\approx$1.1	24.8 (2.1)

*D_0 is the short-time diffusion constant, D_∞ the long-time diffusion constant and τ_1, τ_2 the "relaxation times" of the two switching modes (see Appendix). Standard deviations (see Methods) are given in parentheses.

For what concerns the "relaxation times", i.e. the time required to switch from the fast to the slower diffusion behaviour, the most striking difference can be observed for the slower mode relaxation time, τ_2, values which are similar for the two more structured regions while for the completely unfolded one its value is almost double. This could be explained considering the roughness of the internal energy surface in the unfolded region (left side of the ΔU landscape in Figure 4.3(b)). The presence of deep valleys and high mountains, with internal energy differences up to $\approx$25 kJ/mol, could be the cause of the longer time required for spreading the trajectories over such a corrugated internal energy region. Interestingly a similar trend is observed for the faster mode relaxation time, τ_1.

4.4 Conclusions

The thermodynamic and kinetic properties of the H1 peptide MKHMAGAAAAGAVV taken from the syrian hamster prion protein was explored in explicit aqueous solution at 300K using long time scale all-atom molecular dynamics simulations in the canonical ensemble for a total simulation time of 1.1 μs. At our knowledge this is one of the first attempt to simulate the thermodynamic equilibrium of a complex system, such as a β-hairpin, for more than 1 μs using realistic models for both the peptide and the solvent and with a completely unbiased sampling of the configurational space. The peptide, initially modeled as an α-helix, preferentially adopts β-hairpin structures; furthermore many unfolding/refolding events of the β-hairpin were observed, with an average folding time of $\approx$200 ns. The free energy profile, as a function of the first two essential eigenvectors, that represent the most relevant conformational degrees of freedom of the peptide backbone, has the characteristic features of a funneled landscape, with a downhill surface toward the bottom. ΔU and ΔS profiles share the same funneled like shape of the free energy and their absolute minima almost correspond to the absolute free energy minimum region. Although complete β-hairpin structures mostly populate the free energy minimum, its global free energy of formation, from any other structure, is $\approx$+2.5 kJ/mol. This positive value clearly shows that the "folded structure" is not the thermodynamic most stable condition for this peptide in water. Such a feature is due to the entropy decrease ($\approx$-0.071 kJ mol^{-1} K^{-1}) which overcompensates the internal energy stabilization ($\approx$-18.7 kJ/mol).

According to several experimental evidences, the H1 peptide adopts very rapidly in water β-sheet structure from which amyloid fibrils precipitate, [119, 120] in agreement with our results. Considering the relative instability of the β-hairpin structure in water, revealed in the present study, the interaction with other monomers could be a source of stabilization, leading to amyloid fibril formation.

Furthermore in this study we also characterize the diffusion behaviour in conformational space, investigating its relations with folding/unfolding conditions. The results show that it is possible to accurately describe the kinetics, over the same essential plane used for the thermodynamic characterization, with a dual diffusion model. A first diffusion regime, up to a few ps, probably corresponding to the diffusion of the essential coordinates in a single energy basin, is characterized by a higher diffusion constant. The second diffusion mode is probably connected with the motions from one energy well to another and is characterized by a lower diffusion constant, resulting from an increased friction due to the solvent and the other non-essential coordinates. Moreover in our model a bi-exponential switching (i.e. two relaxation times) from the faster to the slower diffusion mode, yields a very accurate description of the diffusion of the essential coordinates over time intervals up to 100 ps.

Different diffusion behaviours have been observed in relation to the degree of unfolding of the peptide. The more structured regions of the essential plane seem to be associated with a slower long-time diffusion (i.e. higher viscosity of the medium of the

other coordinates) with respect to the less structured ones, for which the D_∞ values are almost three times larger. Interestingly, the relaxation times required to switch from the faster to the slower diffusion regime are longer for the largely unfolded conformational region, being almost double with respect to the folded or partially folded regions. This could be due to the higher roughness of the internal energy surface in the unfolded region, resulting in higher energy barriers to be crossed for spreading the trajectories from an energy well to the others.

Acknowledgements

This work was supported by the European Community Training and Mobility Research Network Project "Protein (mis)folding": HPRN-CT-2002-00241.

CHAPTER

FIVE

Molecular dynamics simulation of the aggregation of the core recognition motif of the islet amyloid polypeptide in explicit water

Summary

The formation of amyloid fibrils is associated with major human diseases. Nevertheless, the molecular mechanism that directs the nucleation of these fibrils is not fully understood. Here, we used molecular dynamics simulations to study the initial self-assembly stages of the NH_2-NFGAIL-COOH peptide, the core-recognition motif of the type II diabetes associated islet amyloid polypeptide. The simulations were performed using multiple replicas of the monomers in explicit water, in a confined box starting from a random distribution of the peptides at T=300 K and T=340 K. At both temperatures the formation of unique clusters was observed after a few nanoseconds. Structural analysis of the clusters clearly suggested the formation of "flat" ellipsoid-shaped clusters through a preferred locally parallel alignment of the peptides. The unique assembly is facilitated by a preference for an extended conformation of the peptides and by intermolecular aromatic interactions. Taken together, our results may provide a description of the molecular recognition determinants involved in fibril formation, in terms of the atomic detailed structure of nascent aggregates. These observations may yield information on new ways to control this process for either materials development or drug-design.

5.1 Introduction

Amyloid fibrils are well-ordered self-assembled protein structures in the nanometric scale. These fibrillar structures are associated with a large variety of diseases of unrelated origin.[19−23,160,161] A partial list of disorders includes Alzheimer's disease, type II diabetes, prion diseases, and primary and secondary amyloidosis. All these diseases are characterized by the formation of large protein deposits (also know as "protein plaques") in various organs and tissues. Fibrils from different sources (e.g., from the pancreas of Type II diabetes patients as compared to the brain of Alzheimer's disease patients) show remarkable ultrastructural and biophysical similarity. Ultrastructural analysis of the deposits, using electron microscopy and atomic force microscopy, demonstrate the existence of fibrils with a diameter of 7-10 nm and a length of several microns.[19−23] Furthermore, X-ray diffraction patterns of several fibrils show a predominant β-sheet structure. Nevertheless, in spite of the high similarity between the fibrils that are formed by the various proteins in different diseases, there is no clear homology between the diverse amyloid-forming polypeptides.

The complexity of amyloid formation underlines the need for highly simplified systems, in which the effects of the perturbation of single properties on aggregation can be pinpointed. Both empirical and rational approaches have been used to design such systems for amyloidogenesis.[29,162−165] In this context, the core recognition motif of the islet amyloid polypeptide (IAPP) serves as an excellent model system to study the process of amyloid formation.[166] This NH_2-NFGAIL-COOH hexapeptide forms fibrils that show remarkable ultrastructural similarity to those that are formed by the full-length IAPP in the pancreas of type II diabetes patients.[166] Using an alanine-scan, the fundamental role of the phenylalanine residue in driving amyloid fibril formation by a peptide that contained the core recognition motif, has been previously demonstrated.[167] The substitution of the phenylalanine to an alanine completely abolished the ability of the fragment to form amyloid fibrils *in vitro* .[167] However, the substitution of the phenylalanine to the less-hydrophobic tryptophan residue resulted in efficient self-assembly of amyloid-related structures.[168, 23] Based on these observations, the remarkable occurrence of aromatic residues in other short amyloid related sequences, and the well-known role of aromatic interactions in processes of self-assembly in chemistry and biochemistry, it has been speculated that interactions between aromatic residues may play a role in the acceleration of the process of amyloid fibrils formation.[162, 23, 169, 170]

Previous Molecular dynamics (MD) simulations of NFGAIL and NFGAILSS peptides[171−174] focussed on the stability of β-clusters by building several different models with two and three strands and by successively performing molecular dynamics simulations to investigate their structure and stability. The results showed that the presence of the two serines resulted in a higher stability of the nascent protofibril.

In the present paper we use MD simulations of NFGAIL sequences to investigate the possible initial steps in the aggregation/nucleation of model peptides and characterize possible molecular recognition mechanisms involved in these processes. Dif-

ferently from previously reported studies, our starting point was not an already preformed aggregate of peptides,[171−174] but a solution of 26 peptides with initial positions and orientations taken at random. The initial conformation of each peptide was chosen as completely extended, consistently with information derived from Solid State NMR studies on fibrils,[160, 175] and in analogy with other simulation studies on the same peptide.[171, 173] Two more control simulations were run imposing two different sets of starting conformations on the peptides. In the first each peptide was modelled in a turn conformation, with the turn spanning residues Gly and Ala. In the second, an -helical conformation was imposed on the four central residues of each peptide. The concentration conditions were the same in every case (vide infra). These two control simulations were run to rule out the (unwarranted) assumption that the peptides have to be extended prior to clustering. The peptides were solvated with explicit water in a confined box to mimic conditions of high local concentration. The starting conformations and concentration were chosen in order to mimic local *microscopic* conditions that can favour the formation of aggregates on all-atom, explicit water MD accessible timescales.

The results clearly demonstrated a specific assembly of the peptide monomers into well-ordered ellipsoid-shaped structures that show a specific network of aromatic interactions.

5.2 Methods

MD simulations, in the NVT ensemble, with fixed bond lengths,[133] were performed with the GROMACS software package[134] and with the GROMOS96 force field.[35] Water was modeled by the simple point charge (SPC) model.[135] A twin range cut-off was used for the calculation of the non-bonded interactions. The short range cutoff radius was set to 0.8 nm and the long range cut-off radius to 1.4 nm for both Coulombic and Lennard-Jones interactions. The Berendsen algorithm[46] was used for the temperature control. 26 replicas of the same NFGAIL peptide were placed in a periodic truncated octahedron large enough to contain the peptides and ∼1.0 nm of solvent on all sides. The initial position of each peptide was chosen at random. In analogy to other simulation studies on the same peptide[171, 173] the initial conformation of every single unit was a totally extended one. Nevertheless, as reported in the results section, several torsional transitions were observed in the simulation.The number of water molecules added was 20266. Two different MD simulations were performed at T=300 K and T=340 K, respectively. Two control simulations were also run: in the first, the peptides were simulated starting from a turn conformation (10 ns) and in the second starting from an α-helix conformation.

The aggregation process in both simulations was determined by following the formation and development of peptide clusters: a peptide unit is added to a pre-existing cluster when its distance to any element of the cluster is less than 0.35 nm, i.e. when at least one atom of the unit forms a van der Waals contact with an atom belonging

to the cluster. This clustering procedure is included in release 3.2.1 of the GROMACS package. To evaluate the dimensions of a cluster, and its global geometrical properties, principal geometrical axes were calculated. This was accomplished by a principal component analysis, at each MD frame, of the spatial atomic fluctuation from the geometrical center of the cluster, as obtained diagonalizing the 3×3 atomic positional covariance matrix $C_{i,j}$:

$$C_{i,j} = \frac{1}{N}\sum_{k=1}^{N}(x_{i,k} - \bar{x}_i)(x_{j,k} - \bar{x}_j) \tag{5.1}$$

where i and j correspond to x,y and z, for the k^{th} atom, N to the total number of atoms and $\bar{x}$ refers to the geometrical center of the cluster. Its eigenvectors represent the principal geometrical axes in 3 dimensional space and the corresponding eigenvalues yield the mean square geometrical fluctuation of the atomic distribution along the three principal geometrical axes (eigenvectors) of the cluster. Note that we define such three geometrical axes according to the decreasing order of the corresponding eigenvalue, i.e. the first eigenvector corresponds to the largest eigenvalue and the third to the smallest. Using the same procedure, the principal geometrical axes were calculated also for each single peptide. In the hypothesis Gaussian statistics for the atomic positional distribution around the geometrical center of the cluster, an estimate of the cluster size along each principal geometrical axes, within 99% of confidence, is given by 6 times the square root of the corresponding eigenvalue (RMSF) (i.e. ±3 standard deviations).

The extent of aromatic packing, and the relative orientations of aromatic rings with respect to each other were evaluated by the calculation of two representative angles γ and θ. For a pair of phenylalanines, γ is the angle between the two ring surface normals; θ is the angle between the normal and the vector; $\mathbf{R}_{cen}$ is the vector connecting the two geometrical centers of the aromatic rings, according to G. B. McGaughey *et al.*[176]

After a stable cluster of peptides was formed in either simulation, quantitative characterization of the dynamical properties was performed, utilizing a principal component analysis of the covariance matrix of the positional fluctuations (essential dynamics analysis) of the C-alpha atoms of the peptides belonging to the cluster, as described elsewhere.[51, 52] This matrix was built from the equilibrated portion of the trajectories (beyond 9.5 ns at 300 K and 8.5 ns at 340 K, and the last ns for the two control simulations), and its diagonalization yielded the principal directions (essential eigenvectors) associated to the large-amplitude concerted motions that define the essential subspace of a cluster's internal dynamics.

5.3 Results

Two simulations at T=300 K and T=340 K were performed for 11.2 ns and 10 ns, respectively, starting from a random distribution of the peptides in the simulation box. In Figure 5.1 the trajectory of the cluster analysis is reported. At 300 K, after 1.5 ns,

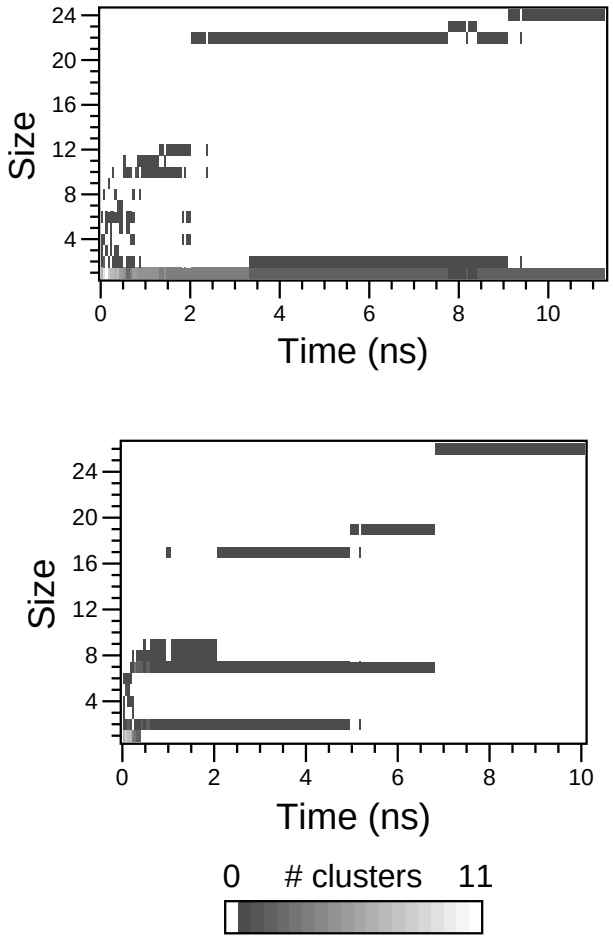


Figure 5.1: Cluster analysis for the simulations at 300 K (top panel) and 340 K (bottom panel) along time. The different gray scales correspond to the number of clusters with a given cluster size. The cluster size corresponds to the number of molecules present in the cluster.

a large cluster of 22 peptides is observed. At t= 9.5 ns two more peptides are added to the cluster in such a way that only two peptides are not included in the aggregate. At T=340 K the formation of a unique cluster can be observed at t=7.0 ns.

The distribution of the ϕ,ψ backbone dihedral angles over the whole simulation time for the two different temperatures (Figure 5.2) shows that the B region is the most populated, showing that the peptides are mostly in an extended conformation.

It has to be pointed out that also the other regions of the Ramachandran plot are populated. Bent and random coil conformations are in fact visited by several of the replicas before the rapid collapse into the main peptide aggregate.

The preferential population of the B region is confirmed by the distributions over the whole simulation time of the eigenvalues corresponding to the principal geometrical axes of each peptide, providing (within 99% of confidence) the peptide time-average sizes of 1.44, 0.36, and 0.12 nm, respectively, which correspond to a rather elongated geometry

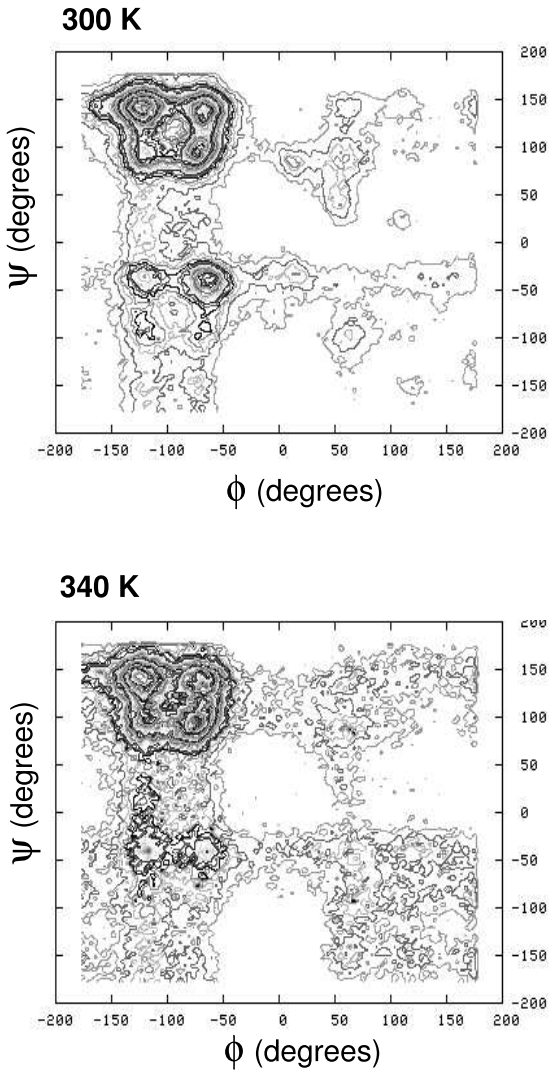


Figure 5.2: Ramachandran plot of the ϕ, ψ backbone dihedral angle distribution at 300 K (top panel) and 340 K (bottom panel).

along the first geometrical axis. Similarly, the eigenvalues of the principal geometrical axes of a cluster can account for its geometrical shape. We calculated, along the MD trajectories, the principal geometrical axes of the cluster in each simulation, after the formation of a unique stable cluster (at t=9.5 ns at 300 K and t=7.0 ns at 340 K). From these data, clusters resulted to be "flat" ellipsoids with the first two principal axes associated to larger eigenvalues. The dimensions along the three geometrical axes were ∼6.6, 6.0 and 3.0 nm (see Figure 5.3).

In Figure 5.4(a) the last snapshot for each simulation, each with two different orientations, one in the plane of the two main geometrical axes (I and II), the other perpendicular to it, are reported. The figure shows that, locally, the peptides prefer to be parallel to each other. This alignment is favored by the relative orientation of

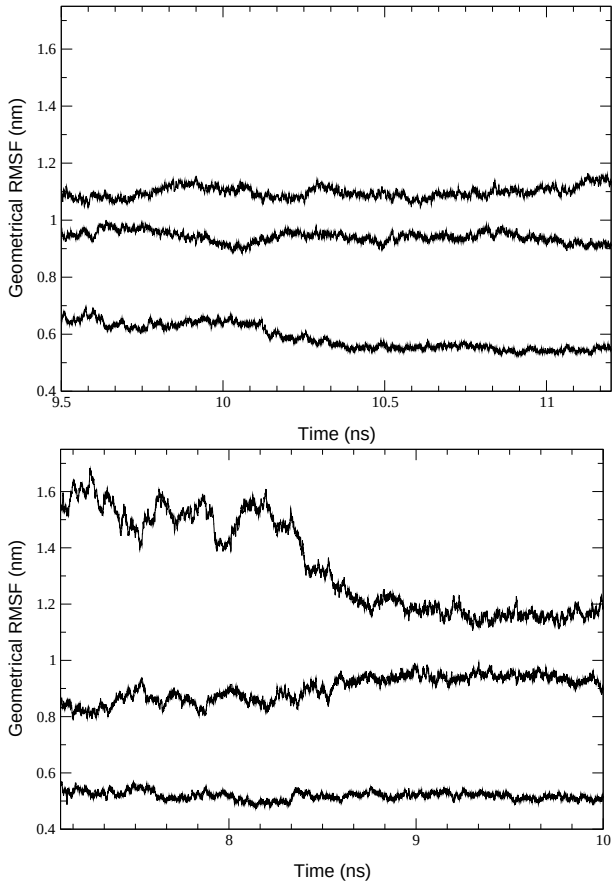


Figure 5.3: Trajectory of the square root of the eigenvalues corresponding to the principal geometrical axes of the cluster at 300 K (top panel) and 340 K (bottom panel).

the aromatic rings of adjacent phenylalanines. We described the orientation of one aromatic ring with respect to the other evaluating two angles, γ and θ, as described in the method section. In Figure 5.5, the γ and θ distributions for aromatic ring pairs with R_{cen} <0.55 nm are reported together with a snapshot of a representative configuration. It is evident that the phenylalanines have a preference to be perpendicular to each other (T-shape). However, different orientations are also present, as both angles are distributed over a relatively large range, in agreement with previous simulation results on Phe-Phe interactions in protein hydrophobic core.[174] Moreover it is interesting to note that the phenylalanine aromatic rings are preferentially oriented toward the internal side of the cluster [Figure 5.4(b)].

In the two control simulations, starting with the peptides in either α-helical or β-turn conformations, aggregate formation follows the same trend as in the simulations reported above. In particular the overall shapes of the final aggregates are similar to the ones observed above, although slightly more elongated along the first dimension. The dimensions along the three geometrical axes were ≈7.8, 5.4 and 3.4 nm for the α-helix simulation and ≈8.4, 5.4 and 3.0 nm for the β-turn simulation, respectively.

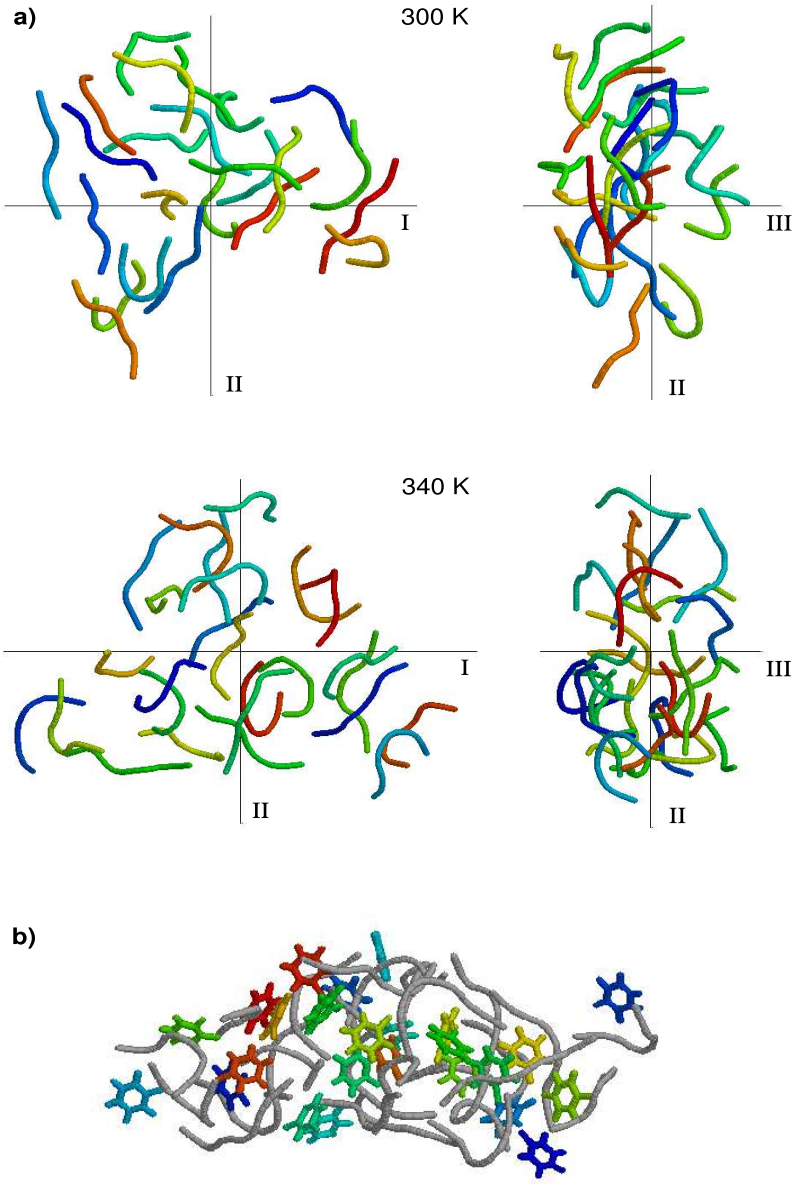


Figure 5.4: **a**) Snapshots of the last configuration for the simulations at 300 K (top) and 340 K (bottom). Left panels represent view onto the plane of the largest geometrical axes (I and II). In the right panels the side view of the snapshots shows the thickness of the "ellipsoid". **b**) Last snapshot of the simulation at 340 K showing the preferred orientation of the phenylalanine aromatic rings toward the interior of the cluster.

The distribution of the ϕ, ψ backbone dihedral angles for the two control simulations at the initial (1ns) and final (1 ns at the end) stages shows that the B region is the most populated at the end of the simulation, once an aggregate is formed (Figure 5.6, panels a and b). It is interesting to observe that also in these two cases, in the aggregate, the phenylalanines have a preference to be perpendicular to each other (T-shape) as

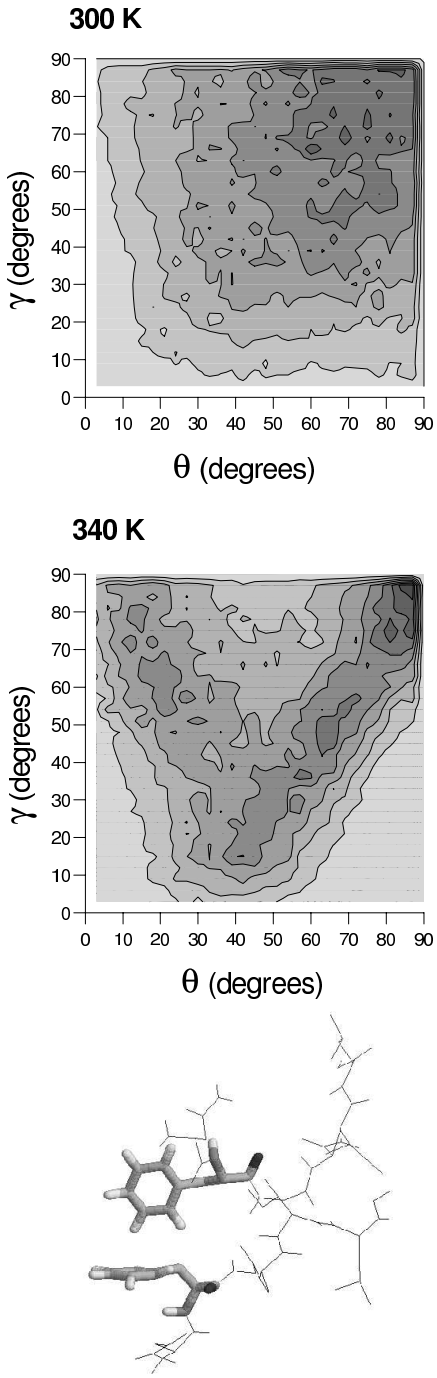


Figure 5.5: γ and θ distributions for aromatic ring pairs with R_{cen} <0.55 nm along the simulation at 300 K (top) and 340 K (middle). In the bottom panel a snapshot of a representative configuration is reported.

observed for the two simulations started from extended conformations (data not shown). Moreover, the preferential population of the B region is once again confirmed by the distributions of the eigenvalues corresponding to the principal geometrical axes of each

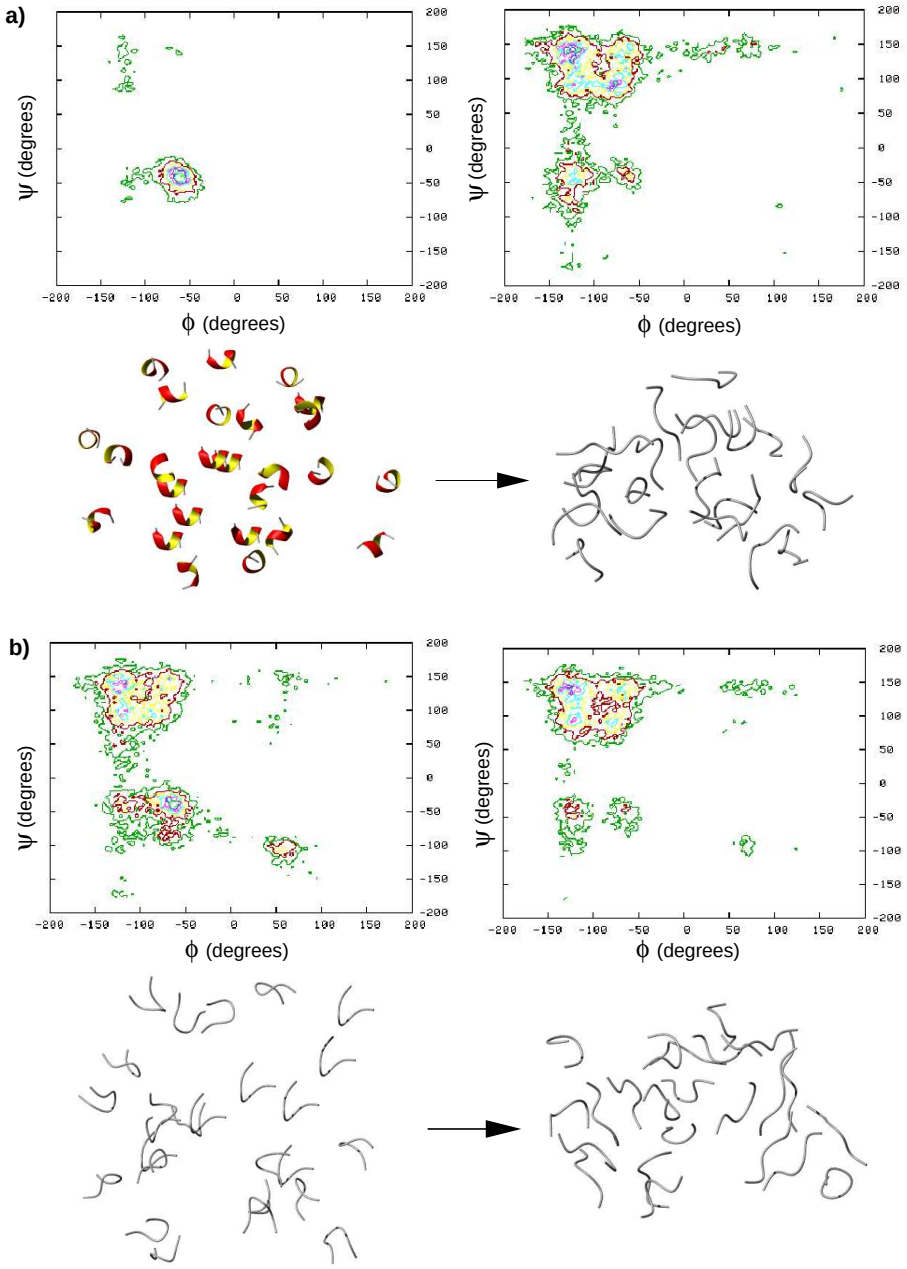


Figure 5.6: **(a)** Ramachandran plot of the ϕ, ψ backbone dihedral angle distribution and structural snapshot at the initial stage (left panel) and equilibrated stage (right panel) of the control simulation starting with α-helical conformations for all peptides. **(b)** Ramachandran plot of the ϕ, ψ backbone dihedral angle distribution and structural snapshot at the initial stage (left panel) and equilibrated stage (right panel) of the control simulation starting with β-turn conformations for all peptides

peptide in the aggregates. Within 99% of confidence the peptides time- average sizes along the geometrical peptide-axes are 1.44, 0.36, and 0.15 nm for the all α-helical control simulation and 1.42, 0.38 and 0.15 nm for the β-turn control simulation. In both

cases these values correspond to a rather elongated geometry along the first geometrical axis.

To analyze the orientation of each peptide with respect to the cluster geometry we have calculated the polar angles, α and β, of the first geometrical axes of the single peptide in the reference frame defined by the three geometrical axes of the cluster (Figure 5.7). α is the angle between the projection of the peptide first geometrical axis onto the I, II plane and the first cluster geometrical axis. β is the angle between the first peptide geometrical axis and the third cluster geometrical axis associated to the smallest eigenvalue (i.e. the direction orthogonal to the main plane of the cluster). In Figure 5.7 we show the α,β distribution, over peptides and MD frames, for the two simulations, beyond 9.5 ns at 300K and 8.5 ns at 340K respectively.

Although a very sharp and ideal distribution of peaks cannot be obtained with this sort of simulation time scales, Figure 5.7 is suggestive of the geometric trends the peptides follow at the onset of the collapse phenomenon. It has to be pointed out that in each plot the contour lines of the minima correspond to a population density that is ten times larger than the population density of the outer region. The Figure shows that at T=300 K the most probable orientation corresponds to $\alpha \approx 45^{\circ}$ and $\beta \approx 80^{\circ}$, i.e. each single peptide is essentially in the I, II plane of the cluster oriented along its bisector. At T=340 K three peaks can be observed: at $\alpha \approx 25^{\circ}$, $\beta \approx 80^{\circ}$ and $\alpha \approx 75^{\circ}$, $\beta \approx 80^{\circ}$, corresponding to orientations almost parallel to axis I and axis II, respectively, and $\alpha \approx 65^{\circ}$, $\beta \approx 60^{\circ}$, corresponding to an out of plane (cluster main plane) angle of $\approx 60^{\circ}$.

Finally, we have performed an essential dynamics analysis to determine the principal overall motions of the cluster and of each peptide within the cluster (data not shown). The principal motion of each cluster can be represented as a wave propagating on a plane surface. From the analysis it is also evident that the motion of each peptide within the cluster is limited, so that no diffusion is detected within our simulation time.

Taken together these results provide evidence that the initial steps of the aggregation between NFGAIL peptides occur through a preferred alignment of the peptides, locally parallel to each other. This is favoured by a preference for an extended conformation of each peptide and by the interaction between the aromatic rings of adjacent phenylalanines. The clusters have the shape of a flat ellipsoid and peptides are specifically oriented within its geometrical main plane.

5.4 Conclusions

Unveiling the molecular causes of the formation of amyloid fibrils is of a key medical importance. It is estimated that there are about 4 million patients that suffer from Alzheimer's disease[177] and about 18 million have type II diabetes[178] in the United States alone. Similar proportional figures are estimated for the rest of the world. As amyloid-related diseases are correlated with advanced age, the global increase in life-expectancy

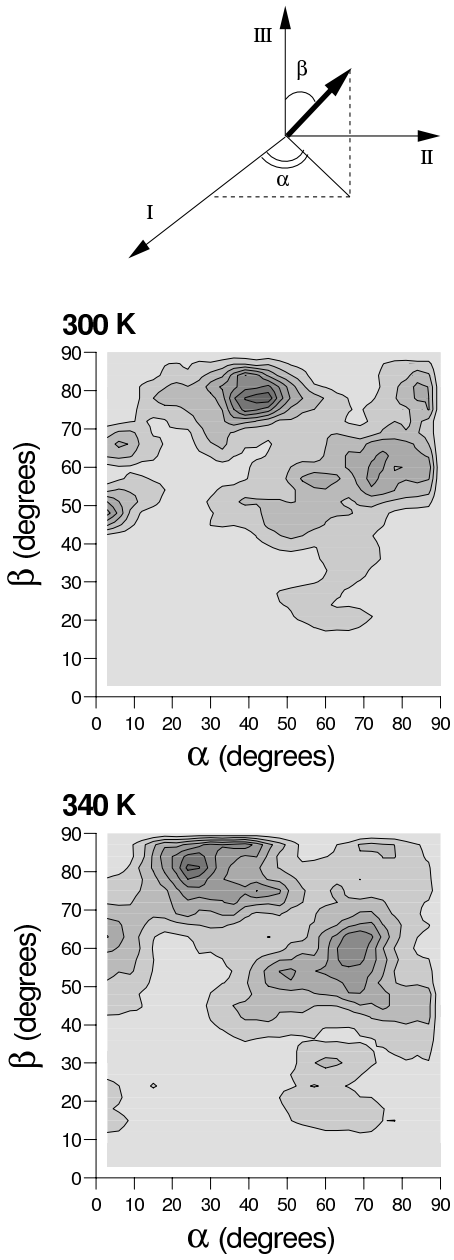


Figure 5.7: Top panel: the vector represents the first principal geometrical axis of a single peptide in the reference frame given by the three principal geometrical axes of the cluster. The density map of the α,β distribution along the trajectories at 300 K and 340 K are reported in the middle and bottom panel respectively. In each plot the contour line of the maximum of population corresponds to a population density ten times higher than the minimum population density, corresponding to the extreme outer contour line.

imply that these group of diseases will dominate the public health concerns in the 21^{st} century. Therefore, significant efforts are being directed toward the development of therapeutic agents that may inhibit the self-assembly process that might leads to the

formation of the fibrils. Atomic level understanding of the very early stage of amyloid formation is highly important for the efforts in this direction.

Here we studied the self-assembly of a short model peptide using molecular dynamics approaches to get insights into the very early steps of the molecular recognition and self-assembly processes that lead to the formation of the fibrils. While we started our simulation with a random distribution of peptide monomers mimicking a high concentration local environment, the formation of large clusters was observed within nanoseconds (Figure 5.1). The clusters appeared to be consistent with a rather extended conformation (Figure 5.2) and the analysis of the principal geometrical axes of the cluster [Figures 5.3, 5.4(a)] was consistent with a "flat" ellipsoid structure with a size of several nanometers, where single peptides present a well defined orientation within such a layer (Figure 5.7). It is worth noting that the same results in terms of dimensional and conformational features of the peptides were observed for two further control simulations started from very different initial conformations, namely α-helical and β-turn. These two simulations should rule out the hypothesis that peptides have to be necessarily extended prior to aggregation. The picture we obtain is actually consistent with the view that peptides can populate a wide set of different conformational families (i.e. non-extended conformations) before starting to aggregate, suggesting that multiple intermolecular interactions within the initial aggregates drive the peptides to mainly populate the B region of the Ramachandran plot.

Interestingly, the structures of the initial aggregates are consistent with the dimensions and organization of prefibrillar assemblies that may actually play a central role in the pathology of amyloid fibrils.[179, 175] It was in fact demonstrated that annular structures with a diameter of 5-15 nm facilitate toxic membrane permeation by the Alzheimer's β-amyloid polypeptide, the Parkinson's α-synuclein polypeptide, and the Type II diabetes islet amyloid polypeptide.

Another interesting point is the organization of the aromatic moieties within the cluster (Figures 5.4(b), 5.5). The analysis of the organization of the phenylalanine aromatic residues suggest the presence of multiple interactions among aromatic phenylalanines that are preferentially organized perpendicular to each other. The existence of the apparent aromatic interactions and their preferential organization provide further support for our hypothesis of the role of aromatic interactions in the early stages of amyloid formation.[23] According to our model stacking and T-shape (edge to face) interactions between aromatic moieties of Phe residues can provide an energetic contribution as well as directionality and orientation, due to the restricted conformational freedom of planar aromatic rings interactions. A support for this notion comes also from the observation that the very simple diphenylalanine peptide contains all the molecular information to self-assemble into well-ordered nanostructures that are structurally related to amyloid fibrils[180]. In terms of mechanism, the results suggest that conformational changes to elongated conformations and the establishment of hydrophobic patches could be taking place in parallel forming stable aggregates.

However, at this stage of calculation and given the restraints imposed by the computational costs of running longer simulations to access much longer time scales needed to observe significant equilibration in the system, we cannot completely exclude the fact that simple hydrophobic collapse might also lead to the formation of aggregates like the ones obtained here.

Hydrophobic collapse might also lead to amorphous aggregation. In the case presented here, however, the number of stereochemical constraints imposed by the sequence, the assembly of the peptides in a specific parallel arrangement, and the geometry of aromatic interactions seem to point to a rather high degree of order of the phenomenon in the microscopic scale.

We have to point out that the choice of the high local concentration conditions and extended conformations of the peptides could put a bias on the aggregation process simulated. However, we also observed several conformational transition of the non-aggregated peptides to random coil and bent conformations prior to the docking on growing cluster, which is composed of mainly extended peptides. This observation suggests the possibility of an actual high preference for the extended conformation within the fibril, consistent with both Solid State NMR and X-ray experimental data.

The same phenomenon might also be simulated with lower concentration conditions, a higher number of peptides and many more different starting geometries to increase the statistics on the conformational and molecular recognition requirements needed for fibril formation. This would also allow the characterization of possible diffusion pathways of the monomers in the cluster. This would however require such high number of particles and long simulation times, that are currently out of reach of all atom MD simulations. Therefore as different initial concentrations affect the rate of fibrillization but not the structural features of the fibrils,[181] it is practical to use high concentration of peptide at the simulation when the structural features are probed. It is worth noting that other studies based on a simplified implicit solvent model, and using a lower number of peptide replicas,[130, 182, 183] have addressed this point, showing results consistent with the ones presented here in terms of the geometry of peptide arrangement. With these caveats in mind, MD simulations can be used to create useful model to help rationalize experimental data.

Taken together, this set of data provides an atomic resolution model of the initial stages of a peptide aggregation process, yielding information on the nature of intermolecular interactions among peptides and on the preferred conformational states of the nascent protofibril that can be used in the development of anti-aggregation (amyloid breaking) sequences or lead molecules, or in the design of orderly aggregating new sequences with potential applications in material science.

Acknowledgments

This study was supported by EU community (RTN grant HPRN-CT-2002-00241 "Protein (mis)folding" to A.D.N.) and by the Israel Science Foundation (F.I.R.S.T program to E.G.).

CHAPTER

SIX

Molecular dynamics simulation of protein folding by essential dynamics sampling: folding landscape of horse heart cytochrome c

Summary

A new method for simulating the folding process of a protein is reported. The method is based on the Essential Dynamics Sampling technique. In EDS a usual molecular dynamics simulation is performed, but only those steps, not increasing the distance from a target structure, are accepted. The distance is calculated in a configurational subspace defined by a set of generalized coordinates obtained by an essential dynamics analysis of an equilibrated trajectory. The method was applied to the folding process of horse heart cytochrome c, a protein with ∼3000 degrees of freedom. Starting from structures, with a root mean square deviation of ∼20 Å from the crystal structure, the correct folding was obtained, by utilizing "only" 106 generalized degrees of freedom, chosen among those accounting for the backbone carbon atoms motions, hence not containing any information on the side chains. The folding pathways found are in agreement with experimental data on the same molecule.

6.1 Introduction

The characterization of the protein folding process represents one of the major challenges in molecular biology. Large theoretical and experimental research efforts have been devoted to this end.[14, 3, 184, 185] Computer simulations have been largely used, coupled to theoretical approaches, to address this question and molecular dynamics simulations is one of the most used computational methods. The major problem with MD simulations is due to the conformational sampling efficiency; in fact even in the 1-microsecond simulation of a 36-residue protein,[16] one of the longest simulations so far afforded , the sampled space explored represents a small fraction of the available conformational space. For this reason different techniques have been proposed to overcome this limit. Three kinds of most commonly used MD techniques can be identified: one approach is to unfold starting from the native state under denaturing conditions, mainly high temperature.[105–107] However, the unfolding process is not necessarily the reverse of the folding process and therefore the issue of whether unfolding simulations are representative for the folding process is still open.[109] Another way of addressing this problem is the so called "biased-sampling free energy" method,[94, 186, 15] in which high temperature unfolding simulations are followed by the calculation of the free-energy of a folding process at 300 K, along the previously determined path. Also this elegant, but time-consuming, method is based on the hypothesis that the unfolding process at high temperature and the folding process at 300 K follow the same path. The third method is the "targeted molecular dynamics" (TMD), in which an additional time-dependent harmonic restraint, applied on each atom, continuously decreases the all-atom root mean square deviation from the native state.[187] TMD has been previously used to calculate reaction paths between two conformations of a molecule.[188–190]

In the present study we present a different computational approach to the folding problem, based on the essential dynamics sampling.[103, 191] In the essential dynamics,[51] or principal component,[52] analysis a new Cartesian reference system is obtained; each new axis (eigenvector), obtained by the diagonalization of the covariance matrix of positional fluctuations, corresponds to a collective motion of the system and after sorting the eigenvectors, according to the displacement involved in each one (eigenvalues), the first ones correspond to the large concerted motions of the system and the last ones represent the collective quasi-constraint (usually referred as near constraint) vibrations. The EDS technique was introduced to increase (or decrease) the distance from a reference structure. To this end, the distance is calculated in the new reference system (obtained by the previously described ED analysis of an equilibrated trajectory) using only a subset of the generalized degrees of freedom of the system, i.e. a subset of the eigenvectors. As reported in the methods chapter (paragraph 2.3.2), with EDS a usual MD simulation is performed in each step; the new position is accepted if the step does not decrease (or does not increase) the distance from the reference structure in the chosen subspace. Otherwise the current structure is projected onto the closest configuration, with the same distance of the previous one in the chosen subspace. Although

proposed in 1996, this technique was never used to follow the folding process of a protein. It has to be pointed out that with this biased MD simulation no deterministic force is added to the system and the correct folding can be obtained by using a small fraction of the degrees of freedom of the protein to bias the simulation. In the present case these degrees of freedom were chosen among those accounting for backbone carbon atoms motions, hence not containing any information on the side chains.

Here we present the results obtained in the EDS folding simulation of cytochrome c (cyt c). Cyt c is a globular protein of 104 amino acids, whose folding dynamics has been subjected to extensive experimental investigations.[192–198] In particular fluorescent data [197, 199] from Trp-59 suggested an early collapse of the main chain structure within 100 μs; time-resolved circular dichroism [192] and small-angle x-ray scattering, SAXS,[193] suggested the presence of two folding intermediates having ∼0.5 ms and ∼7 ms lifetimes. The SAXS measurements also suggested, in agreement with theoretical investigations on different proteins,[200, 201, 106] that after an initial decrease of the radius of gyration, the main-chain collapse of the structure and the secondary structure formation are mostly concerted. Interestingly, recent fluorescence energy transfer studies on the iso-cytochrome c folding,[202] providing the distribution of distances between donor and acceptor labelled residues, suggested that only a small fraction of the collapsed structures correctly folds. In fact, most of those structures adopt frustrated topologies separated by large energy barriers from the folding funnel.

6.2 Methods

6.2.1 Molecular Dynamics Simulations

The starting structure for the simulation at 300 K was taken from the 1.94 Å resolution refined crystal structure of the protein cyt c (pdb-entry 1*hrc*) [203] (Figure 6.1).

The simulated system was set up as described elsewhere.[108] All MD simulations were performed using the GROMACS software package and the Gromos87 force field [204] was used with modification as suggested by van Buuren *et al.*;[205] explicit hydrogen atoms in aromatic rings were simulated.[35] The protein was solvated with water in a periodic rectangular box of dimensions 67.90x63.27x72.26 Å. The SHAKE algorithm [206] was used to constrain all bond lengths, the simple point charge [207] water model was used and the temperature was kept constant with the isokinetic temperature coupling.[158] A non-bond pairlist list cutoff of 9.0 Å was used and the pairlist was updated every 4 time steps. The long-range electrostatic interactions were treated with the particle mesh Ewald method [157] using a 56x53x60 grid combined with a fourth-order B-spline interpolation to compute the potential and forces in between grid points. A time step of 2 fs was used for numerical integration.

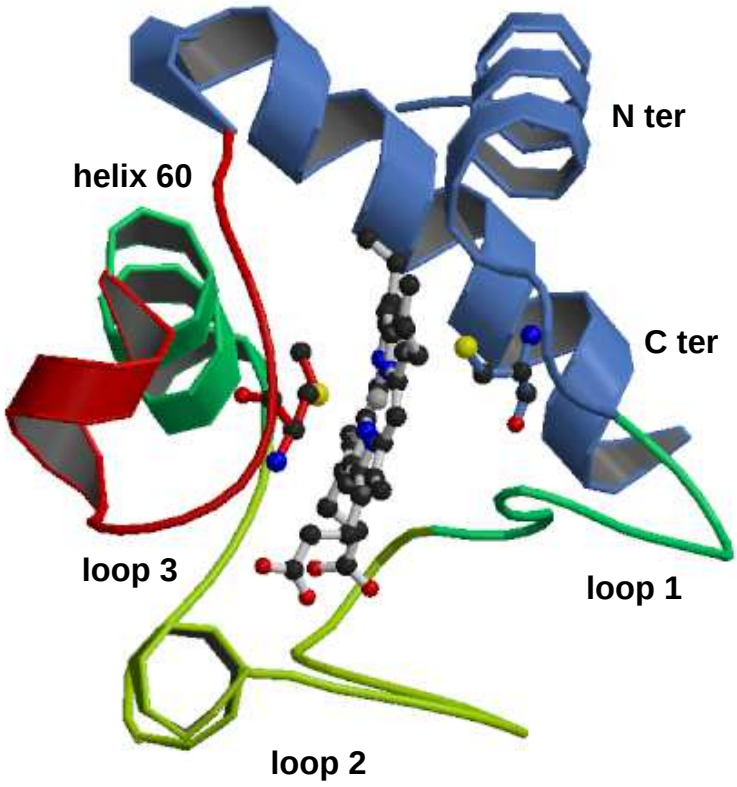


Figure 6.1: Crystal structure of cytochrome c.

6.2.2 Essential dynamics analysis

A molecular dynamics simulation at 300 K was performed for 2660 ps. From the equilibrated portion of the trajectory (beyond 160 ps) the covariance matrix, of order 312, of the positional fluctuations of the Cα carbon atoms was built up and diagonalized. The procedure yielded new axes (eigenvectors), representing the directions of the concerted motions. The corresponding eigenvalues gave the mean square positional fluctuation for each direction.[51]

6.2.3 Essential dynamics sampling

The principles of the EDS are described in section 2.3.2. It has to be pointed out that in the present case, the eigenvectors were obtained by the diagonalization of the matrix of the positional fluctuations of the backbone carbon atoms (104 carbons, i.e. 312 eigenvectors), so that they do not contain any information on the other atoms, in particular on the side chains.

6.2.4 Unfolding/refolding simulations

To produce the starting unfolded structures the EDS technique at 300 K was used in the expansion mode,[103, 191] utilizing all the 306 native eigenvectors (the last six eigenvectors represent the overall roto-translation and have zero eigenvalues). Ten unfolding simulations were performed, starting at different times of the 2660 ps simulation of the native structure, utilized in the ED analysis. Preliminary EDS folding simulations were performed with different procedures: a first simulation used all the 306 native backbone carbon atom eigenvectors to calculate the distance from the target and apply the bias (EDS procedure). Starting from the same unfolded structure three additional simulations were performed by using the EDS procedure with three lower dimensional subspaces: eigenvectors 1-100, eigenvectors 101-200 and eigenvectors 201-306, respectively. Finally 9 additional simulations, starting from the 9 previously determined unfolded structures, were performed using the last subspace (eigenvectors 201-306) for the EDS procedure.

6.2.5 Contacts

According to the GROMACS definition, a contact between residues i and j>(i+3) was considered present if the smallest distance between any two atoms, belonging to the two residues, was less than 5.5 Å. The fraction of native contacts, ρ, is calculated with respect to the crystal structure.

6.3 Results

Starting from the crystal structure, a 2660 ps simulation at 300 K, in explicit solvent, was performed. From the equilibrated portion of the trajectory (beyond 160 ps) the covariance matrix of the positional fluctuation of the Cα carbon atoms was built and diagonalized. The main structural properties of the equilibrated portion of the trajectory are reported in Table 6.1.

Starting from the structure at 2500 ps of the 300 K simulation, an unfolding simulation was performed by an EDS expansion procedure at T=300 K using all the 306 native eigenvectors and the crystal structure as reference. The final structure (RUN 1 in Figure 6.2) was characterized by radius of gyration (Rg) of 18.94 Å and (with respect to the crystal structure) by root mean square deviation (rmsd) of the Cα carbon atoms of 19.13 Å , fraction of native contacts of 0.23 and native helix content (θ) of 28%.

The refolding process was simulated by the EDS contracting procedure, using all the 306 native eigenvectors (ALL) to bias the system toward the target. The Cα rmsd, with respect to the target, reached very rapidly, i.e. within 250 ps, a value close to 1.0 Å and the average structure over the last 100 ps was close to the target one (Table 6.1).

Table 6.1: **Structural properties in the crystal (row 1), in the MD simulation of the native structure (row 2) and at the end of the refolding trajectories (rows 3-6).**

	$rmsd_{C\alpha}$[#] (Å)	$rmsd_{sc}$[#] (Å)	Rg (Å)	ρ[#]	%helix[#]	θ[#] (%)
Crystal	-	-	12.64	1.00	41	100
Native	1.44(0.22)	2.53(0.20)	12.70(0.20)	0.84(0.01)	42(4)	94(3)
ALL*	0.40(0.02)	2.44(0.04)	12.76(0.02)	0.83(0.01)	40(1)	94(2)
SET1*	2.43(0.02)	5.74(0.05)	12.98(0.02)	0.50(0.01)	10(3)	28(6)
SET2*	5.32(0.10)	7.13(0.08)	13.61(0.05)	0.54(0.01)	17(2)	35(7)
SET3*	2.33(0.07)	3.82(0.07)	13.11(0.05)	0.73(0.01)	43(2)	98(2)

*All the values are averaged over the last 100 ps of each trajectory. Standard deviations in parentheses.
#The rmsd's, $rmsd_{C\alpha}$ and $rmsd_{sc}$, the native contacts content, ρ, and the native helix content, θ, are calculated with respect to the crystal structure. %helix represents the total helix content.

To characterize the different contribution of the native eigenvectors to the refolding process, they were divided into three sets: eigenvectors 1-100, 101-200 and 201-306. Using these three sets for the EDS procedure, three new refolding simulations (SET1, SET2 and SET3) were performed. As reported in Table 6.1, only the last set gave an average final structure close to the target one. In Figure 6.3 the ribbon diagrams of sequential snapshots along the refolding trajectory using SET3 are represented. This result suggests that the most rigid quasi-constraint eigenvectors, representing in the folded protein the smallest collective vibrations, contain the proper mechanical information for the folding process. It is also worth noting that a correct folding was obtained using in the EDS procedure only 106 eigenvectors for a protein of ∼3000 degrees of freedom. These eigenvectors seem to control and constrain the internal motion of the secondary structure or loop elements, as shown in Figure 6.4, where we report the fractional decomposition of the overall Cα displacement due to each single eigenvector into internal and rototranslational (with respect to the Cα centroids) ones. The results, for the terminal helices, 60's helix and loop1, make evidence that the last set of eigenvectors mostly represents internal collective vibrations, i.e. within the secondary structure or loop element considered. In addition it is evident from the fractional mean square displacement per atom (obtained by the eigenvectors components) in the native structure simulation, calculated for the helices and the loops, along each eigenvector (Figure 6.5), that eigenvectors in the range 210-275 are mainly involved in the loops

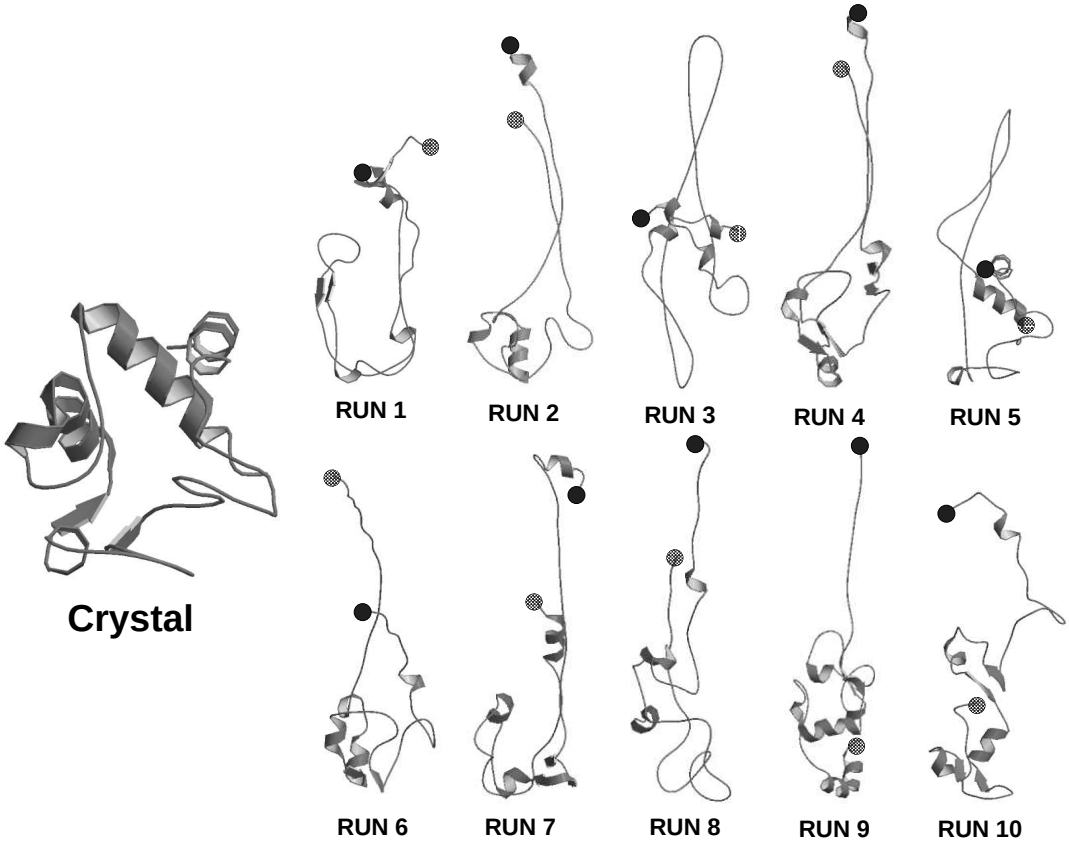


Figure 6.2: Ribbon diagrams of the crystal structure and of the starting unfolded structures of the refolding trajectories. The N- and C- terminal residues are represented by a black and a gray circle, respectively.

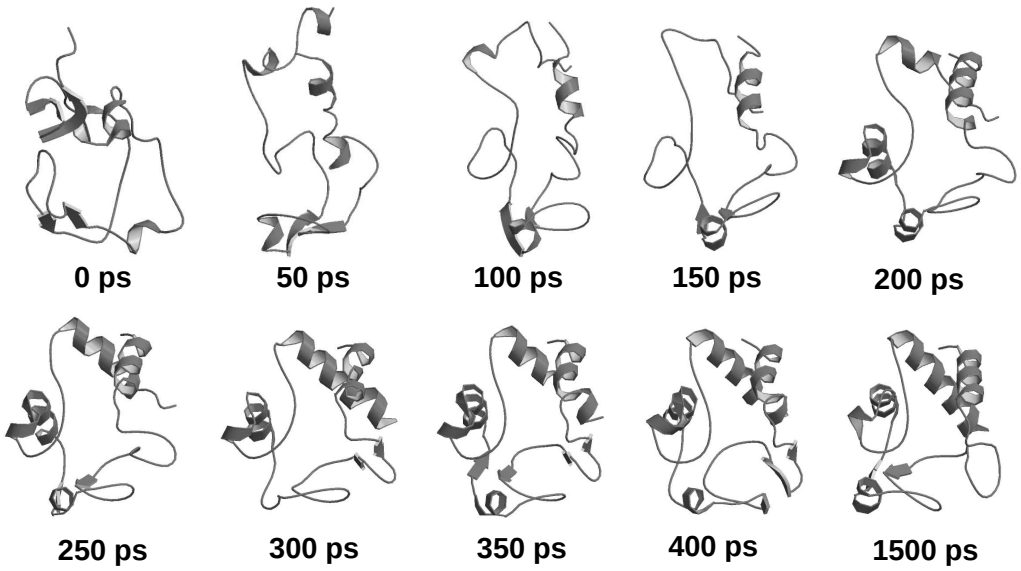


Figure 6.3: Ribbon diagrams of sequential snapshots along the refolding trajectory using SET3.

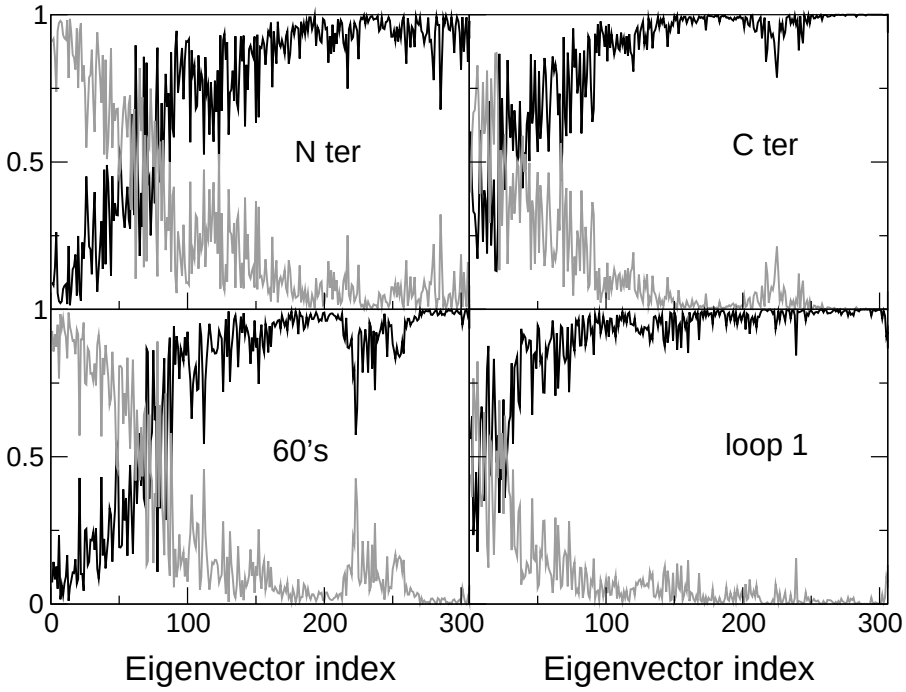


Figure 6.4: Fractional Cα internal (black) and rototranslational, with respect to the Cα centroids (gray), displacements for the Nter helix (left top), Cter helix (right top), 60's helix (left down) and loop 1 (right down) vs the eigenvector index.

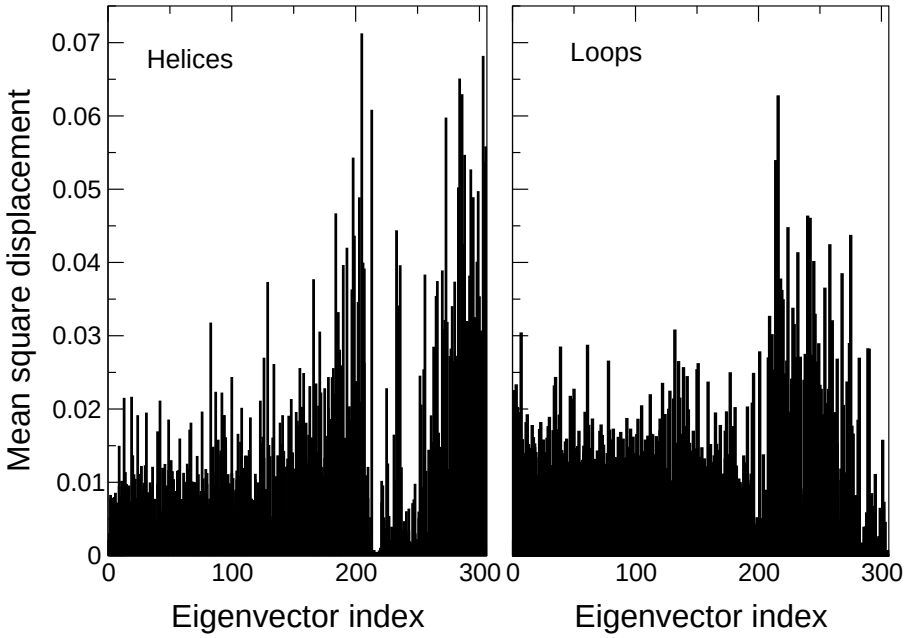


Figure 6.5: Fractional mean square displacement per atom (obtained by the eigenvector components) along each eigenvector, calculated for the helices Nter, Cter and 60's (left) and for the loops 1,2 and 3 (right).

motion, while eigenvectors in the ranges 200-210 and 275-306 are mainly involved in the helices motion. The mean square displacement per atom of a helix or a loop was calculated averaging the sum of the square components of each eigenvector of the atoms belonging to secondary structure or loop element, respectively.

Taken together these results show that, although the correct folding can be obtained using all the 306 Cα carbon eigenvectors, a folded structure of comparable quality can be obtained using only the last 106 eigenvectors. In what follows we will perform different independent folding simulations using this last set of eigenvectors. This because we want to use the least biased procedure in our folding simulations and find out the main mechanical information necessary for the folding process.

Table 6.2: **Structural properties of the starting unfolded structures of the refolding trajectories.**

	rmsd$_{C\alpha}$* (Å)	rmsd$_{sc}$* (Å)	Rg (Å)	ρ^*	%helix*	θ^* (%)
RUN 1	19.13	20.06	18.94	0.23	14	28
RUN 2	21.26	21.80	24.47	0.50	22	49
RUN 3	21.45	21.20	21.50	0.43	19	44
RUN 4	20.69	21.47	22.92	0.43	11	37
RUN 5	16.93	17.33	19.24	0.48	19	56
RUN 6	25.66	26.77	27.59	0.38	18	53
RUN 7	26.97	27.82	29.24	0.34	27	63
RUN 8	20.42	20.44	21.66	0.33	15	42
RUN 9	22.32	22.28	21.39	0.43	31	67
RUN 10	21.08	21.58	21.94	0.42	15	46

*The rmsd's, the native contacts content, ρ, and the native helix content, θ, are calculated with respect to the crystal structure. %helix represents the total helix content.

In order to have a better statistics we performed 9 additional independent unfolding simulations starting at different times of the native simulation, thus obtaining different final structures (Figure 6.2 and Table 6.2). The EDS re-folding simulations (RUN 2 to

Table 6.3: **Final structural properties of the refolding trajectories.**

RUNS*	$rmsd_{C\alpha}$# (Å)	$rmsd_{sc}$# (Å)	Rg (Å)	ρ#	%helix#	θ# (%)
RUN 1§	2.33(0.07)	3.82(0.07)	13.11(0.05)	0.73(0.01)	43(2)	98(2)
RUN 2	1.36(0.04)	2.75(0.05)	12.76(0.04)	0.82(0.01)	39(3)	86(4)
RUN 3	1.97(0.06)	3.18(0.06)	13.11(0.05)	0.77(0.01)	41(2)	96(3)
RUN 4	2.12(0.10)	3.22(0.07)	12.96(0.04)	0.77(0.01)	40(1)	93(3)
RUN 5	1.84(0.07)	2.94(0.08)	12.93(0.04)	0.78(0.01)	38(2)	89(5)
RUN 6	4.35(0.06)	5.57(0.06)	13.52(0.06)	0.61(0.01)	35(2)	80(5)
RUN 7	3.48(0.07)	4.50(0.09)	13.41(0.05)	0.62(0.01)	33(2)	83(4)
RUN 8	2.86(0.08)	4.25(0.06)	13.23(0.06)	0.65(0.01)	36(2)	83(4)
RUN 9	2.26(0.06)	3.64(0.07)	13.40(0.04)	0.75(0.01)	42(1)	98(2)
RUN 10	1.87(0.06)	3.27(0.08)	13.03(0.05)	0.77(0.01)	42(1)	96(3)

*All the values are averaged over the last 100 ps of each trajectory. Standard deviations in parentheses.
#The rmsd's, $rmsd_{C\alpha}$ and $rmsd_{sc}$, the native contacts content, ρ, and the native helix content, θ, are calculated with respect to the crystal structure. %helix represents the total helix content.
§RUN 1 coincides with SET3 of Table 6.1.

10) were performed for 1.0-1.5 ns, with the same procedure adopted for SET 3: 300 K and utilizing only eigenvectors 201-306 in the EDS procedure. The results, reported in Table 6.3 (RUN 1 of Table 6.3 coincides with SET3 of Table 6.1), show that simulations from 1 to 5 converged well to the target structure with values comparable with the native structure simulation (Table 6.1). Simulations 6, 7 and 8 did not show rmsd, native contacts or helix content in agreement with the target. Simulation 9 and 10 are doubtful because, although they show values comparable with the native structure, the terminal helices do not show a proper folding. In fact the rmsd, with respect to the crystal, of the terminal helices, averaged over the last 100 ps, is much larger than in RUNS 1-5, being ≈4.5 Å on respect to ≈2.0 Å . In addition, as discussed later, they show a small value of native contacts content between the terminal helices. Interestingly recent fluorescence energy transfer studies on the iso-cytochrome c folding,[202] measured the distribution of distances between donor and acceptor labelled residues and suggested

that only a fraction of the collapsed structures correctly fold. It has to be pointed out (Figure 6.2) that the starting structures of simulations 6-10 did not have any contact between the terminal helices, as shown by the N- and C- terminal residues represented by a black and a gray circle, respectively. Hence the contact between the terminal helices seems to be a prerequisite for a proper folding, in agreement with the hypothesized role of these contacts in the cyt c folding process.[208, 209, 196] Figure 6.6 (RUN 1-5) shows that the correct folding is obtained when the native contacts between the terminal helices precede those between helices 60's and C-ter. The process is reversed in RUN 9 and RUN 10, where the contacts between the terminal helices reached ∼50% of the native structure value.

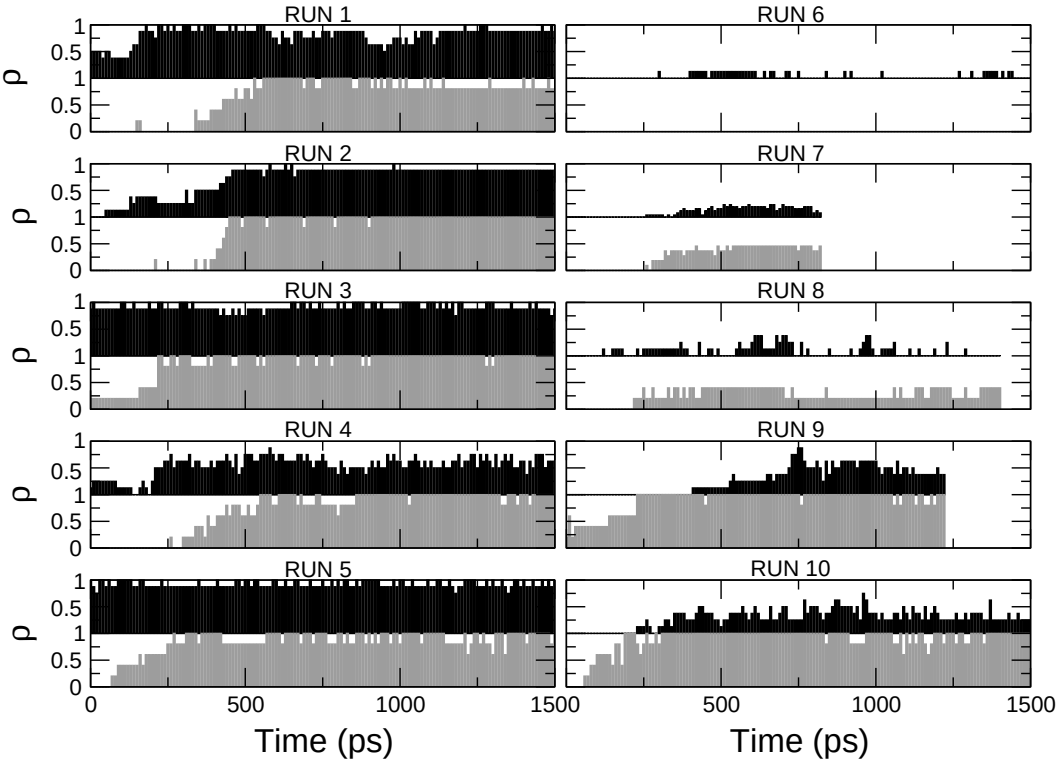


Figure 6.6: Evolution in time (RUN 1-10) of the fraction of native contacts between terminal helices (black) and between helices Cter and 60's (gray).

The correlation among the native contacts content, the radius of gyration and the helix content in the EDS folding trajectories (Figure 6.7) for RUN 1 to 10 shows that the folding process can be divided into two steps: the first one is characterized by the decrease of the radius of gyration, with no significant increase of the native contacts content and amount of secondary structure; in the last part of the simulation the radius of gyration is almost constant, while the native contacts and the secondary structure content increase in an almost concerted way. This sequence is actually in agreement with the one proposed by SAXS and CD measurements[192, 193] and MD data

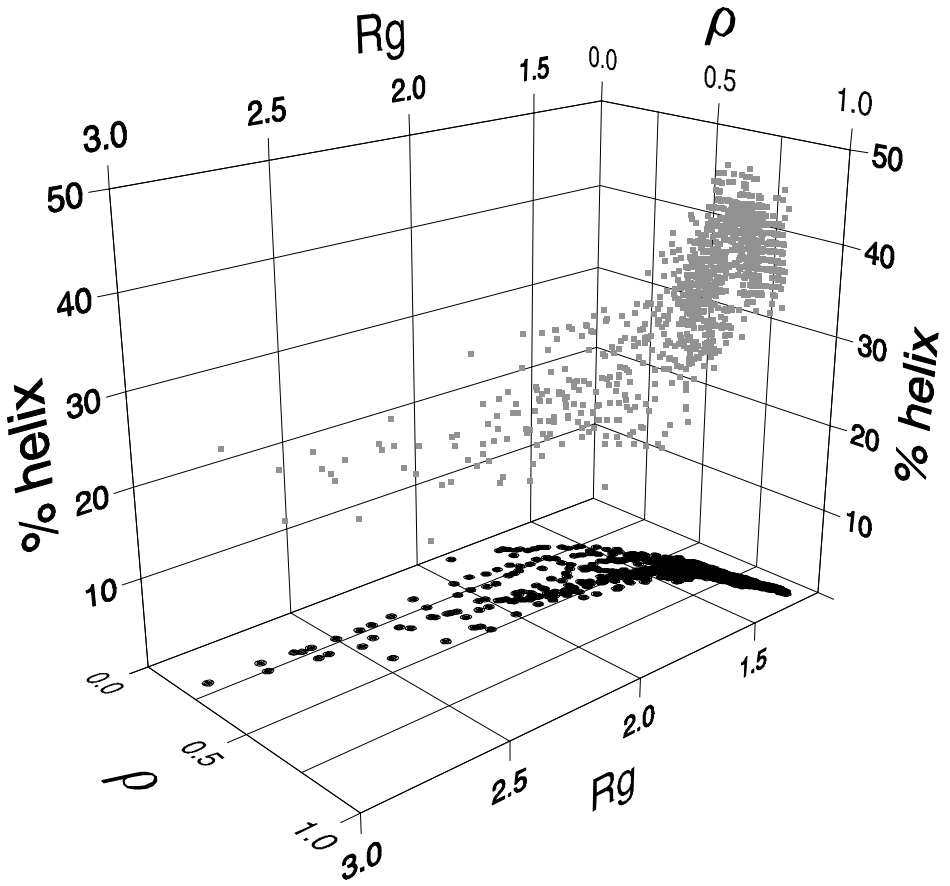


Figure 6.7: Gray squares: correlation among the native contacts content (ρ), the radius of gyration (R_g) and the total helix content (% helix). Black circles represent the projection onto the R_g-ρ plane.

on different proteins.[200, 201, 106] The SAXS and CD measurements also suggested that the folding process of cyt c is characterized by two intermediates, as evidenced by the analysis of the time dependence of the radius of gyration that was fitted by a double exponential characterized by time constants of ∼0.5 ms and ∼15 ms. In the present case the double exponential behaviour was less evident (data not shown), however the double exponential fitting gave an excellent correlation coefficient, r = 0.998, and time constants of 120 ps and 4420 ps. The difference of the time constant magnitude has to be ascribed to the EDS method that speeds up considerably the sampling toward the folded condition; however the ratios between the experimental time constants (∼30) and our time constants (∼36) are comparable.

6.4 Conclusions

In the present study a new method to simulate the folding process of a protein to its native state is reported. The method is based on the essential dynamics sampling procedure and provides a biased MD simulation, which restrains 106 over the ∼3000 degrees of freedom of the protein. These restrained degrees of freedom are obtained by the essential dynamics analysis of the positional fluctuations of the backbone carbon atoms and do not contain any information on the other backbone and side chain atoms. It has to be pointed out that in the EDS procedure no deterministic force is added to the Hamiltonian and hence the system is not systematically forced toward the target. The restraints were applied only to the last eigenvectors, representing the most rigid quasi-constraint motions, while all the other degrees of freedom were completely free to sample the configurational space, according to the usual equations of motion. The results also showed that the restrained eigenvectors are mostly involved in the internal collective motions, within helices or loops, while the essential eigenvectors (the first 10-20) provide mainly rototranslational motions of helices or loops. Such results clearly show that the last eigenvectors define the main mechanical constraints necessary in a folded protein, while the essential eigenvectors really represent the large internal motion which can occur without unfolding the protein.

The folding of cytochrome c was simulated as a test. The results evidenced that 5 essays (out of 10) were successful, 3 essays were not and 2 were doubtful. It has to be pointed out that also fluorescence energy transfer studies on the iso-cytochrome c folding[202] suggested that only a fraction of the collapsed structures correctly fold. Finally, our results showed that in the EDS simulations the folding process of cyt c is characterized by an initial decrease of the radius of gyration, with no significant increase of the native contacts and of secondary structure content; in the last part of the simulation the radius of gyration is almost constant, while the native contacts percentage and the secondary structure content increase in an almost concerted way. This folding path is in agreement with the experimental suggestions[192, 193] on cyt c and with MD data on different proteins.[200, 201, 106]

Acknowledgments

This work was supported by the European Community Training and Mobility of Researchers Program "Protein (mis)-folding", by MURST 2001 PRIN "Structural Biology and dynamics of redox proteins" and CNR ag. 2000. A. Di Nola acknowledges the "Centro di eccellenza BEMM" of the University of Roma "La Sapienza".

CHAPTER

SEVEN

Investigating the accessibility of the closed domain conformation of citrate synthase using essential dynamics sampling

Summary

A molecular dynamics study of pig heart citrate synthase is presented which aims to directly address the question whether for this enzyme the ligand-induced closed domain conformation is accessible to the open unliganded enzyme. The approach utilizes the technique of essential dynamics sampling which is used in two modes. In exploring mode the enzyme is encouraged to explore domain conformations it might not normally sample in free molecular dynamics simulation. In targeting mode the enzyme is encouraged to adopt the domain conformation of a target structure. Using both modes extensively it has been found that when the enzyme is prepared from a crystallographic open-domain structure and is in the unliganded state, it is unable to adopt the crystallographic closed-domain conformation of the liganded enzyme. Likewise, when the enzyme is prepared from the crystallographic closed liganded conformation with the ligands removed, it is unable to adopt the crystallographic open domain conformation. Structural investigations point to a common structural difference that is the source of this energy barrier, namely the shift of α-helix 328-341 along its own axis relative to the large domain. Without this shift the domains are unable to close or open fully. The charged substrate, oxaloacetate, binds near the base of this helix in the large domain and the interaction of Arg329 at the base of the helix with oxaloacetate is one that is consistent with the shift of this helix in going from the crystallographic open to closed structure. Therefore the results suggest that without the substrate the enzyme remains in a partially open conformation ready to receive the substrate. In this way the efficiency of the enzyme should be increased over one that is closed part of the time, with

its binding site inaccessible to the substrate.

7.1 Introduction

Domain movements form a large class of functional movements in enzymes, and some effort has been made to understand and characterise them. [210, 211, 212] In the simplest scenario, the substrate binds to an open conformation of the enzyme inducing closure. Once closed the reaction catalysed by the enzyme can proceed in a protected and highly specific environment. It is of interest to know whether the closed domain conformation of the enzyme is accessible or inaccessible (see footnote[1]) to the unliganded open conformation of the enzyme. Gerstein *et al.* [211] in their review of domain movements speculate that both the open and closed conformations are dynamically accessible to the unliganded enzyme at physiological temperature. Their model implies that there exists a continuous range of stable domain conformations between the most open and closed ones. They partly base their arguments on the finding that a closed unliganded form of the binding protein lactoferrin is stabilized by weak crystal packing forces. [213] Although this may be true for some domain proteins, it is not necessarily a universal truth as many domain proteins have much more complicated domain interfaces than lactoferrin. Indeed recent combined NMR and fluorescence experiments on maltose-binding protein has confirmed a barrier between the open and closed domain conformations for that protein. [214] Crystallographic work on citrate synthase does not support the idea of a continuous range of stable domain conformations but one where there are just two stable states related to enzymatic mechanism. [215]

Molecular dynamics simulations on pig heart citrate synthase [216] suggest that there is a large free energy barrier to surmount to reach the crystallographic closed domain conformation from the crystallographic open conformation in the unliganded state. On the basis of that study it was concluded that the energy to surmount this barrier comes from the interaction of the enzyme with the substrate. The concept of an energy barrier between the open and closed domain conformation would make sense for an enzyme, in that an enzyme that remained open would be more efficient than one that spent some of its time closed with its binding site inaccessible to the substrate.

Citrate synthase catalyses a step in the citric acid cycle, namely the Claisen condensation of acetyl-coenzyme A with oxaloacetate to form citrate and coenzyme. [217] It is a homodimer, where the monomer comprises a large and small domain. It is an enzyme that displays a classic domain movement as part of its function, where the binding of oxaloacetate induces domain closure, upon which the binding site for acetyl-coenzyme A is formed. [217] The MD simulation study referred to above comprised three simulations that started from the crystallographic open conformation and a further three simulations that started from the crystallographic closed conformation. In both cases any ligands were removed. The simulations starting from the open conformation appeared to show that first, there are a large number of unliganded domain conformations that are accessible from the unliganded open conformation, but second, the crystallographic

[1] We used the word "inaccessible" to indicate a free energy barrier between states making it very unlikely for there to be a transition from one to the other.

closed domain conformation cannot be reached from the crystallographic open conformation in the unliganded state. In the simulations starting from the crystallographic closed conformation the trajectories remained around the closed domain conformation, apart from one, where one of the monomers made an apparently spontaneous transition to the region explored by the open simulations. Although these results suggested closed conformation could spontaneously convert to the open in the absence of the products, this probably never occurs in reality because the enzyme needs first to open to allow citrate to escape. In order to investigate the results of the previous free MD simulations further, in this work we have used the Essential Dynamics Sampling technique. [103] This technique has been used in a number of studies on protein [104] and peptides dynamics [191] as well as folding/unfolding simulations, [108, 151] but this is its first application to the study of a functional domain movement in an enzyme. Using this technique we are able to encourage the domain conformation to explore a larger region of space than it would in free MD, and also to move towards particular target conformations.

7.2 Methods

7.2.1 Molecular dynamics simulations

The details of the protocols used in the simulations performed in this work are the same as in the previous work. [216] In short, molecular dynamics simulations were performed on fully solvated dimers (∼80,000 atoms in total) of pig heart citrate synthase using GROMACS. [218] The initial structures were the crystallographic closed structure [219] liganded both to citrate, and coenzyme A (PDB accession code: 2CTS), and the crystallographic open structure liganded to citrate, but bound differently to when it is a product (PDB accession code: 1CTS). [219] None of these ligands were included in the simulations.

7.2.2 Essential Dynamics sampling

The principles of the Essential Dynamics Sampling are described in section 2.3.2. Here the sampling is performed in two distinct modes: "targeting" and "exploring". In the former contraction is performed to a specified target conformation. In the latter initial expansion occurs from a specified reference conformation (e.g. the crystallographic open conformation), but when expansion is halted according to two parameters described below, the final conformation becomes the new reference conformation from which a new expansion is started. There are two parameters required for the EDS in the exploring mode: the maximum number of sampling cycles (ncycles) before changing the origin of expansion, and the slope which sets a minimum on the rate of expansion. These parameters were fixed to 5000 and 0.0004 nm/step, respectively. Targeting mode simulations were stopped when the radius failed to decrease any further in a number of consecutive steps. In all the simulations reported EDS was applied to one monomer only, the other being allowed to undergo free MD.

7.2.3 Sampling subspace

The sampling space was the six eigenvectors that represent the relative rigid-body motion of the two domains of a single monomer. The domains were assigned as before[216] and comprised a large domain of residues 1-55, 67-278, and 378-437, and a small domain of residues 56-66, and 278-377 which were determined by the DynDom program. [220, 221] These six eigenvectors were determined as follows. The trajectories of both monomers from the open simulations in the previous work [216] combined to give an equivalent single-monomer trajectory of 12 ns. This trajectory was used for a rigid-body essential dynamics analysis that was slightly different to that described previously. The external motion of the monomer was removed by superposing each conformation on the experimental open conformation. Superposition was done using the usual least-squares fitting procedure. Intradomain fluctuation was removed from this trajectory of intramonomer fluctuation by superposing the experimental open domain conformations on their respective domain conformations at each time frame. Superposition was done using C_α atoms only. This gave us a trajectory of the two domains as rigid bodies. Then conventional essential dynamics analysis (principal component analysis) was applied to this trajectory resulting in six non-zero eigenvalues, which collectively quantify the amount of relative motion there is between the two domains. The eigenvectors corresponding to these six eigenvalues determine the subspace for the EDS. Applying EDS in this subspace allows us to encourage one domain to move relative to the other, and to explore domain conformations accessible from a specific conformation (exploring mode), or to target a specific domain conformation (targeting mode). Note that these constraints are applied to the domain conformations only and all the intradomain degrees of freedom are allowed to undergo free MD.

7.2.4 Visualization of relative motion of the domains

The trajectories are displayed by projecting onto the two-dimensional space specified by the first two eigenvectors of the rigid-body essential dynamics analysis of the combined open trajectories from the previous paper. [216] Roughly 80% of the domain fluctuation in the combined open trajectories occurred in this two-dimensional space and the domain movement between the two crystallographic conformations could be represented to 98% by these two modes. Thus projecting the new trajectories from the EDS simulations onto this space enables one to visualize the relative motion of the domains.

7.2.5 DynDom and Dom_Select

In order to analyse the domain movements that occur, two programs have been used. The program DynDom [220] takes two conformations and determines dynamic domains, hinge axes and hinge-bending regions. It determines domains automatically based on the conformational change itself. In this work we have also used an unreleased program

Dom_Select which allows the user to specify the domains themselves by residue number ranges. Once the user specifies the domains, the hinge axis is determined in the same way as in the DynDom program.

7.2.6 Rigid-body RMSD

This quantity was used in the previous analysis. [216] Consider a part of the protein that moves from an initial to a final position. At the same time the internal conformation of this part changes. The rigid body movement of the part between the initial and final positions is calculated by superposing the initial conformation of the part on the final conformation. Thus one has the initial conformation in two positions, the initial and final. The rigid-body root mean-square deviation, RG_RMSD, is simply derived from the displacement of each atom between these two positions.

7.2.7 Helix_Shift

The unreleased program Helix_Shift calculates the shift of an α-helix along its own axis. As above, consider a helix that moves from an initial to a final position. At the same time the internal conformation of the helix changes slightly. The rigid body movement of the helix between the initial and final positions is calculated by superposing the initial conformation of the helix on the final conformation of the helix. The movement of the helix along its own axis is then determined by calculating the distance between the centres of mass of the helix in these two positions and projecting this distance onto the helical axis of the helix in its initial position. The direction of this axis is estimated by superposing an ideal α-helix of identical length with its axis along the z-axis onto the real helix. The last column of the rotation matrix from this least-squares superposition gives the direction of the axis of the real helix from which the projected distance can be calculated.

7.3 Results

7.3.1 Simulations from open conformation

Exploring mode simulations

Three exploring mode simulations were performed starting from the open conformation that were used to start production runs in the original work. [216] This open domain conformation does not coincide exactly with the crystallographic open domain conformation but are very near to it in comparison to the crystallographic closed domain conformation. Details of these simulations are given in Table 7.1. Figure 7.1 shows the domain trajectories projected onto the two main degrees of freedom for the domain movement (see Methods).

Table 7.1: **Runs originating from crystallographic open structure**

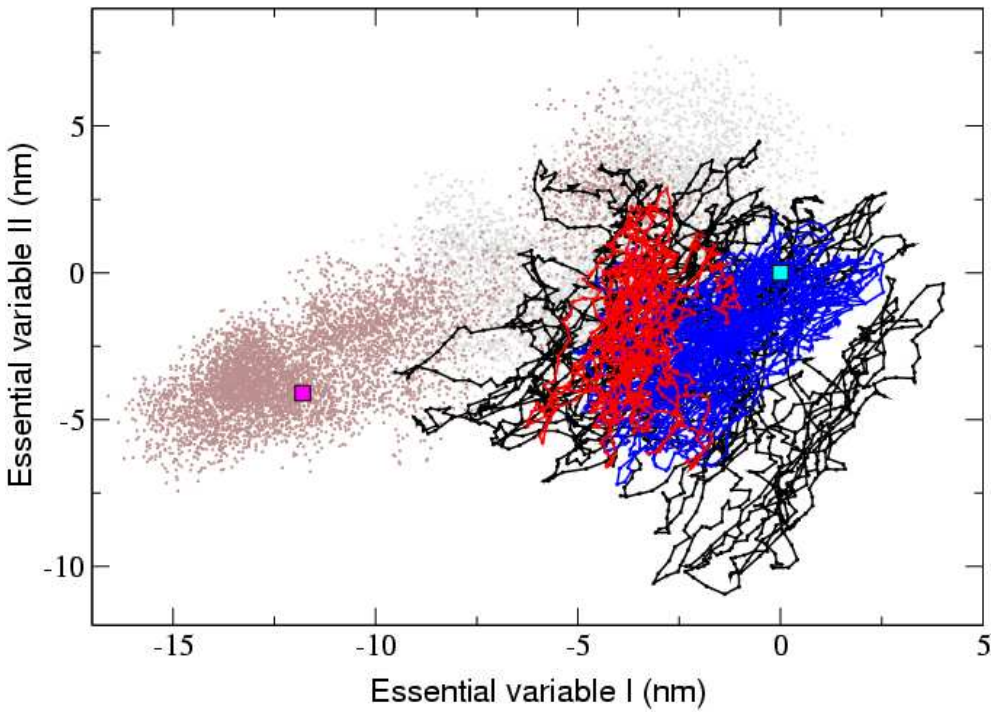
Run	Mode	Starting Structure	Simulation Length (ps)
Run 1	Exploring	Equilibrated crystal open*	1300
Run 2	Exploring	Equilibrated crystal open*	1165
Run 3	Exploring	Equilibrated crystal open*	500
Run 4	Target to closed	Equilibrated crystal open*	500
Run 5	Target to closed	Equilibrated crystal open*	500
Run 6	Target to closed	Equilibrated crystal open*	500
Run 7	Target to closed	Conformation at 800ps of Run 1	500
Run 8	Exploring	Final conformation of Run 5	500
Run 9	Target back-to-open	Conformation at 40ps of Run 8	500

*Each had a unique set of starting velocities generated from the Maxwell distribution.

Also shown in this figure are the trajectories of the original free simulations [216] starting from the open and closed. It is clear that domain conformations are explored that are not explored in the original open free simulations. In Run 1 there is some probing towards the crystallographic closed domain conformation, but generally the region around it is avoided. In order to investigate further whether trajectories that start from the crystallographic open conformation are able or unable to reach the crystallographic closed domain conformation, targeted simulations were performed.

To-closed targeting mode simulations

In these four simulations the crystallographic closed domain conformation is the target. Three simulations started from the same open conformation from which the exploring mode simulations were started, although each had a different set of velocities derived from the Maxwell velocity distribution. The fourth simulation started from the 800ps conformation of Run 1 of the exploring mode simulations. Table 7.1 gives the details of these runs and Figure 7.2 shows their domain trajectories.


Essential variable II (nm)
5
0
-5
-10
-15
-10
-5
0
5
Essential variable I (nm)

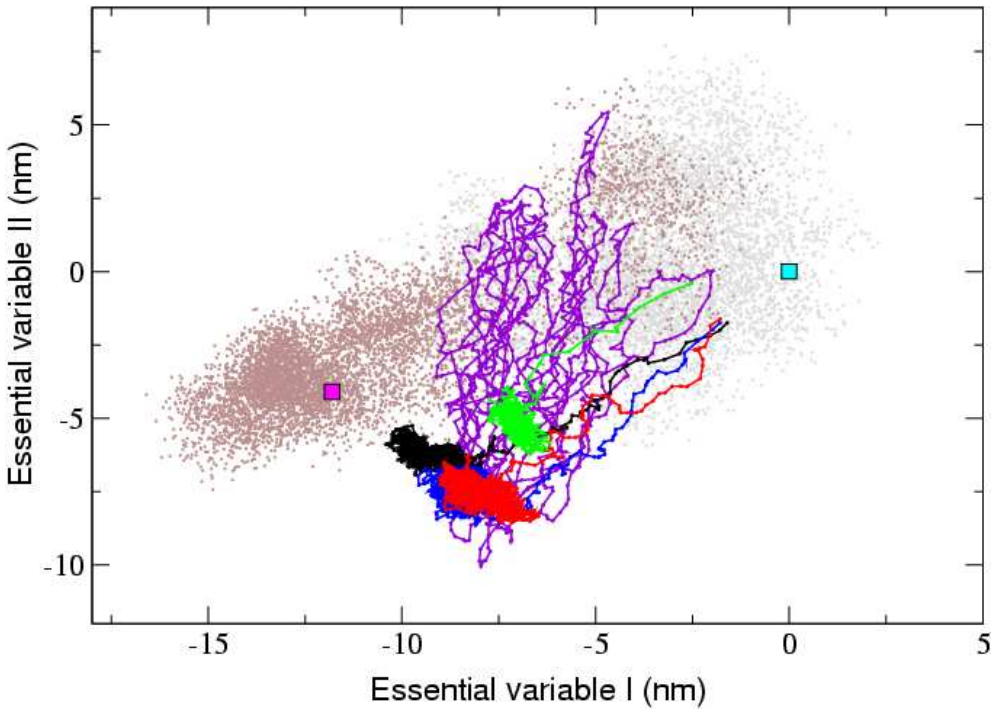

Essential variable II (nm)
5
0
-5
-10
-15
-10
-5
0
5
Essential variable I (nm)

Figure 7.1: Projections of the trajectories of the exploring mode simulations that started from the crystallographic open conformation, onto the first two eigenvectors of the rigid-body essential dynamics analysis. The three simulations shown in black, blue and red correspond to Run 1, Run 2 and Run 3 of Table 7.1, respectively. The crystallographic open and closed conformations are indicated with a cyan and magenta filled square, respectively. In the plot the trajectories of the original free simulations starting from the crystallographic open and closed conformations are also shown in grey and brown respectively.

Figure 7.2: Projections of the trajectories of the "to-closed" targeting mode simulations starting from the crystallographic open conformation, onto the first two eigenvectors of the rigid-body essential dynamics analysis. The four simulations shown in black, blue, red and green correspond to Run 4, Run 5, Run 6 and Run 7 of Table 7.1, respectively. Also shown in violet in the plot is the projection of the exploring mode simulation, corresponding to Run 8 of Table 7.1. The crystallographic open and closed conformations are indicated with a cyan and a magenta filled square respectively. In the plot the trajectories of the original free simulations [216] starting from the crystallographic open and closed conformations are also shown in grey and brown, respectively.

All trajectories rapidly move towards the target but eventually are unable to move any closer and remain stuck around a region of closest approach. The simulation with the different starting conformation gets stuck in a slightly different region from the others. These simulations confirm that the crystallographic closed domain conformation is inaccessible to conformations that start from the crystallographic open conformation. This finding supports the conclusion from the earlier work [216] where a free energy barrier between the open and closed was proposed.

Further simulations

Figure 7.2 also shows an exploring simulation that was started from the final conformation of a targeting simulation. In addition a simulation that targeted back to the crystallographic open domain conformation was also performed. The trajectory did not reach the crystallographic open domain conformation. Details of these simulations are given in Table 7.1.

7.3.2 Simulations from closed conformation

Exploring mode simulations

Three exploring mode simulations were performed from the closed starting conformation, which was the same starting point as for the free MD simulations and was close to, but not coincident with, the crystallographic closed domain conformation. Details of the simulations are reported in Table 7.2. The projected domain trajectories are displayed in Figure 7.3.

They move in a region around the main distribution of closed domain conformations determined in the free simulations. However, as one would expect in an exploring mode simulation they went beyond some of the outermost regions explored in the original closed simulations.

To-open targeting mode simulations

In these four simulations the crystallographic open domain conformation is the target. Three simulations started from the same closed conformation, although each had a different set of velocities derived from the Maxwell velocity distribution. The fourth simulation started from the 500 ps conformation of Run 2, an exploring mode simulation. Table 7.2 gives the details of these simulations and Figure 7.4 shows their domain trajectories.

All the trajectories rapidly move towards the target but eventually are unable to move any closer and remain stuck around a region of closest approach. These simulations appear to show that the crystallographic open domain conformation is inaccessible to conformations that start from the closed. These trajectories follow the path taken by the "transitional trajectory" of the free MD simulation study [216] indicating the

Table 7.2: **Runs originating from crystallographic closed structure**

Run	Mode	Starting Structure	Simulation Length (ps)
Run 1	Exploring	Equilibrated crystal closed*	500
Run 2	Exploring	Equilibrated crystal closed*	500
Run 3	Exploring	Equilibrated crystal closed*	500
Run 4	Target to open	Equilibrated crystal closed*	500
Run 5	Target to open	Equilibrated crystal closed*	500
Run 6	Target to open	Equilibrated crystal closed*	500
Run 7	Target to open	Conformation at 500ps of Run 2	500
Run 8	Exploring	Final conformation of Run 4	850
Run 9	Target back-to-closed	Conformation at 500ps of Run 8	500

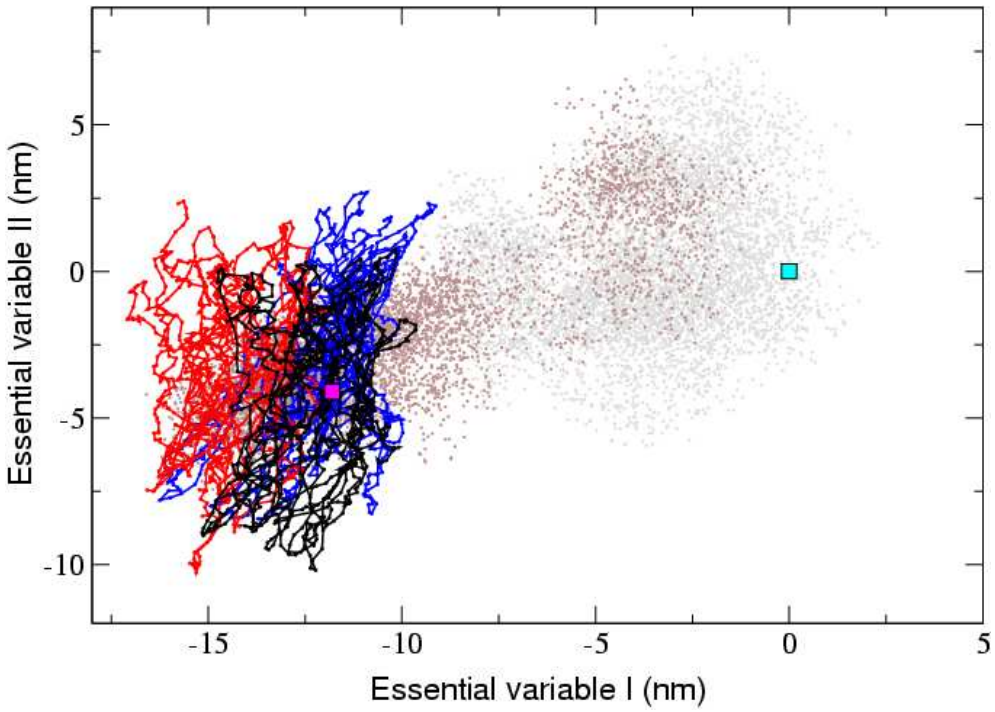
*Each had a unique set of starting velocities generated from the Maxwell distribution.

consistency of these results with that study. However, from the free MD study it was concluded that the "transitional trajectory" was able to reach the open conformation. This conclusion is not supported by the results here. The domain conformations at the end of these targeted trajectories are indeed closer to the crystallographic open domain conformation than the closed, but some internal differences prevent these conformations from reaching the crystallographic open domain conformation.

Further simulations

Figure 7.4 also shows an exploring mode simulation (Run 8 in Table 7.2) which was started from the end of the targeted simulation, Run 4. Again this trajectory explores the regions accessed by the transitional trajectory going back to the crystallographic closed domain conformation 6 times along the same path as the transitional trajectory.

This supports the finding that the crystallographic open domain conformation is inaccessible and that there is an internal difference that allows these rather open domain conformations to reach the crystallographic closed domain, but not the open one.


Essential variable II (nm)
5
0
-5
-10
-15
-10
-5
0
5
Essential variable I (nm)

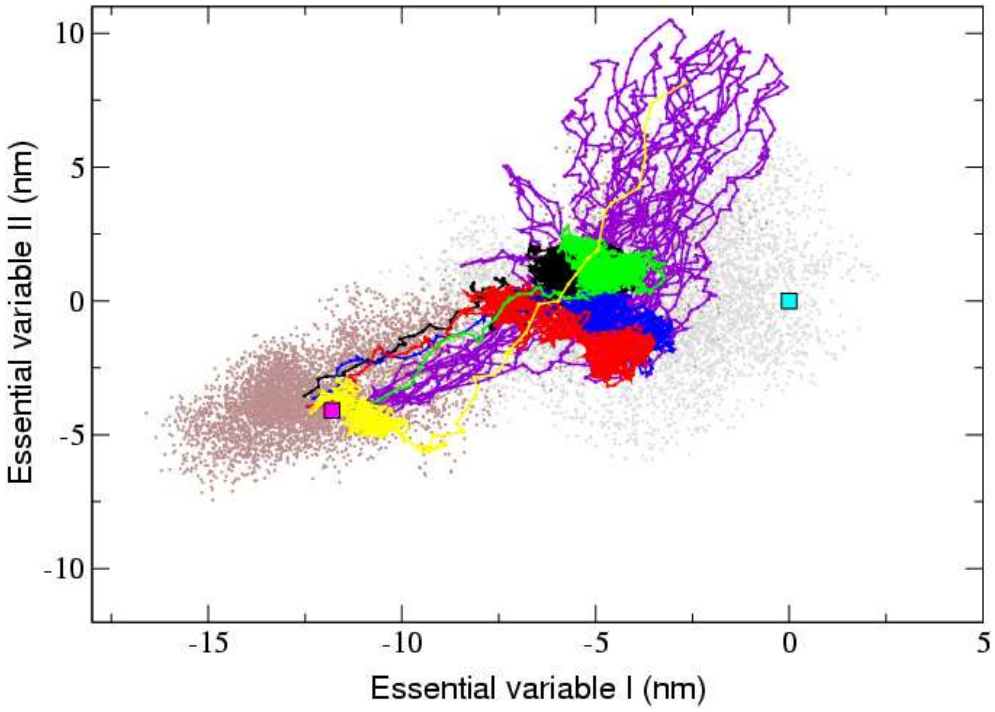

Essential variable II (nm)
10
5
0
-5
-10
-15
-10
-5
0
5
Essential variable I (nm)

Figure 7.3: Projections of the trajectories of the exploring mode simulations starting from the crystallographic closed conformation, onto the first two eigenvectors of the rigid-body essential dynamics analysis. The three simulations shown in black, blue and red correspond to Run 1, Run 2 and Run 3 of Table 7.2, respectively. The crystallographic open and closed conformations are indicated with a cyan and a magenta filled square, respectively. In the plot the trajectories of the original free simulations [216] starting from the crystallographic open and closed conformations are also shown in grey and brown, respectively.

Figure 7.4: Projections of the trajectories of the "to-open" targeting mode simulations starting from the crystallographic open conformation onto the first two eigenvectors, of the rigid-body essential dynamics analysis. The four simulations shown in black, blue, red and green correspond to Run 4, Run 5, Run 6 and Run 7 of Table 7.2, respectively. Also shown in violet and in yellow are the projections of Run 8 (exploring mode simulation) and Run 9 (targeting mode simulation) of Table 7.2, respectively. The crystallographic open and closed conformations are indicated with a cyan and a magenta filled square, respectively. In the plot the trajectories of the original free simulations [216] starting from the crystallographic open and closed conformations are also shown in grey and brown respectively.

At the 500ps conformation of this exploring mode trajectory, a targeted simulation (Run 9, Table 7.2) was started with the crystallographic closed conformation as the target. The trajectory shown in Figure 7.4 confirms that the crystallographic closed domain conformation is indeed accessible from conformations originating from the crystallographic closed conformation. This was not found in the back-to-open simulations. The path taken in this back-to-closed trajectory is again the same as the transitional trajectory and the exploring mode trajectory of Run 8 indicating a low energy pathway.

Identifying the source of the energy barrier between open and closed conformations

These results confirm one finding of the previous work: [216] that the crystallographic closed domain conformation is inaccessible from the open unliganded conformation. However, on the basis of the "transitional trajectory" it was also speculated that the crystallographic open domain conformation *is* accessible from the crystallographic closed conformation in the unliganded state. However, this work indicates that the crystallographic open domain conformation is not accessible from the crystallographic closed conformation. This suggests that an energy barrier exists between both experimental domain conformations in their unliganded states. A barrier to domain rotation is surely located at an interface region situated between the two domains. This would mean that it is likely to be assigned as a bending region in the DynDom analysis. Given that this barrier is obviously overcome in the presence of the substrate or product, structural changes in the vicinity of the binding sites for these ligands are of particular interest. Recently, Hayward has developed a method to identify residues that are involved in inducing closure in enzymes when an open unliganded structure and a closed liganded structure is available. [222] Using this method three potential "closure-inducing residues" have been identified in citrate synthase: His274, His320 and Arg329. These residues interact with the substrate oxaloacetate to induce closure and it is reasonable to expect therefore, that the barrier to closure will be located in the vicinity of these residues. Previously, it was speculated that the barrier between the open and closed conformations is located at the ψ-dihedral of His274, [216] which in the closed conformation has an angle of -134.7° , which together with a ϕ-dihedral angle of -114.7° puts it in a "disallowed" region of the Ramachandran plot. [219] However, a structure at 140ps of the back-to-closed targeting simulation (Run 9 in Table 7.2), has a domain conformation that is almost identical to the crystallographic closed domain conformation, but the ψ-dihedral angle of His274 has a value of -77.6°, which with a ϕ-dihedral angle of -61.3° puts it in a low energy region of the Ramachandran plot. It seems therefore that the extreme value of the dihedral angle in the crystallographic closed conformation is due to the interaction of this residue with the substrate oxaloacetate and that it need not have this value in order for the enzyme to reach the closed domain conformation. This means that there is no particular hindrance to the domain rotation from this region. Unlike His274, His320 is not assigned as a bending region and is located in a region

that moves as a rather rigid body in going from the crystallographic open to closed conformation. As it is not located in an interdomain region it is difficult to see how any barrier to domain closure reside in the vicinity of His320. Arg329 is assigned as a bending residue and is situated at the N-terminal of an α-helix which undergoes a significant shift upon domain rotation. Its role in domain closure will be elucidated below.

Structural analysis of targeting simulations

The domain conformations in the targeting simulations are straining to achieve their target conformations but are unable to do so. It would appear logical therefore to analyse the movements between the starting conformations, the endpoint conformations of these targeting simulations, and the target conformations themselves. Table 7.3 gives the actual domain rotation angle corresponding to these movements and the RG_RMSDs.

The data in Table 7.3 verifies that there is a significant difference between the fully open and closed and conformations as determined by crystallography and the most open and closed conformations achieved in the targeting simulations. The difference between the domain movements from starting to endpoint conformations and endpoint to target conformations can easily appreciated from Figures 7.2 and 7.4. The starting to endpoint domain movements take routes that are rather parallel to the line that directly joins the crystallographic conformations. The routes taken from the open are on opposite sides of this line to those that start from the closed (see Figure 7.5 for a schematic illustration).

The endpoint to target domain movements would be more perpendicular to this line. The symmetry implied by Figure 7.5 indicates a common structural mechanism that prevents the conformations of the unliganded enzyme originating from the open crystallographic structure reaching the closed, and vice-versa. In order to investigate this two structures were selected that form a line on our 2D projections that is parallel to the direction of the trajectories of the targeting simulations before they get stuck, and a further two that form a line that is perpendicular to this line. Then the program Dom_Select was used to characterise the rigid-body movement of the small domain relative to the large for both of these pairs. In Figure 7.6 these hinge axes are displayed with the enzyme structure.

The hinge axis depicting the movement in the parallel direction lies directly between, and is almost perfectly parallel to, a pair of parallel α-helices: helix 222-235 situated in the large domain, and helix 328-341 situated in the small. The hinge axis depicting the perpendicular movement is not parallel to these helices, but also is situated between them. The fact that in both cases the axes are located directly between these two helices indicates that they play a crucial role in the interdomain movement.

Table 7.3: **Domain rotation angles and RG_RMSD's**

Targeting from open to closed	Small domain rotation angle between start and endpoint (deg.)	RG_RMSD between small domain at start and endpoint (Å)	Small domain rotation angle between endpoint and target (deg.)	RG_RMSD between small domain at endpoint and target (Å)
Run 1	20.3	5.6	2.5	1.8
Run 2	15.5	5.4	5.2	2.6
Run 3	14.4	5.2	7.7	3.1
Run 4	14.0	4.2	8.4	3.1
Targeting to open from closed				
Run 1	15.1	4.1	7.1	2.9
Run 2	13.6	4.5	5.8	1.8
Run 3	12.1	4.3	8.5	2.1
Run 4	16.7	4.4	7.1	3.0

Shift of α-helix 328-341

Although helix 328-341 is assigned to the small domain here, DynDom often assigns a portion of this helix to belong to the large domain in terms of its rotational properties (please see the DynDom database of protein domain motions for more details on the domain movement of citrate synthase between crystallographic open and closed conformations). [226] Helix 222-235 belongs unambiguously to the large domain. The movement between the crystallographic open and closed conformations is described by a hinge axis that makes an angle of approximately 30^o with these helices. Consequently, the movement from crystallographic open to closed results in a distinct shift of the helix 328-341 "downwards" relative to the large domain (for convenience the direction "up" will be used to refer to the direction of the helical axis of this helix pointing along the direction

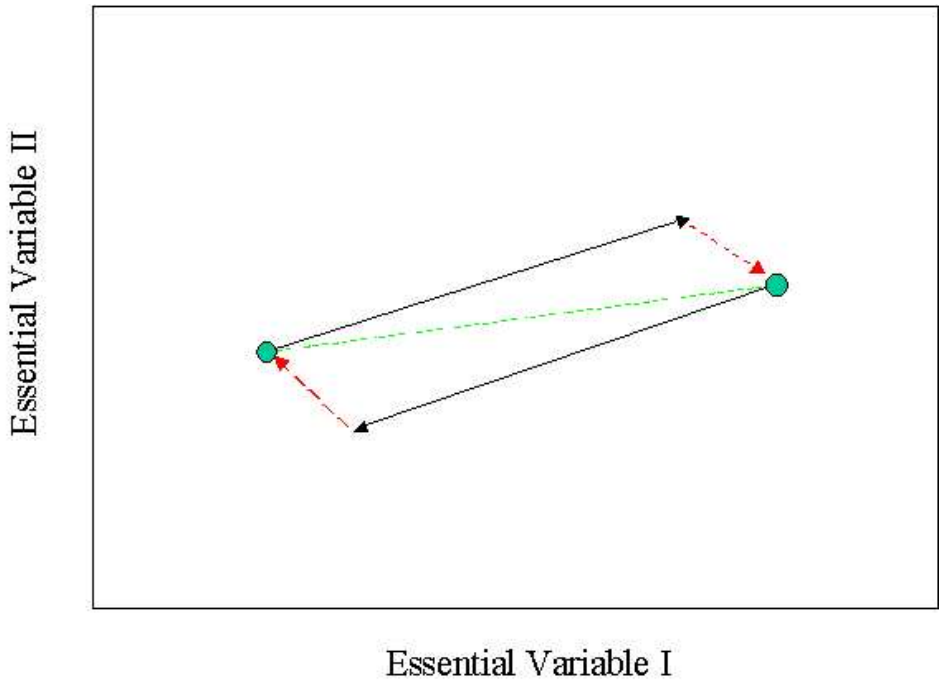
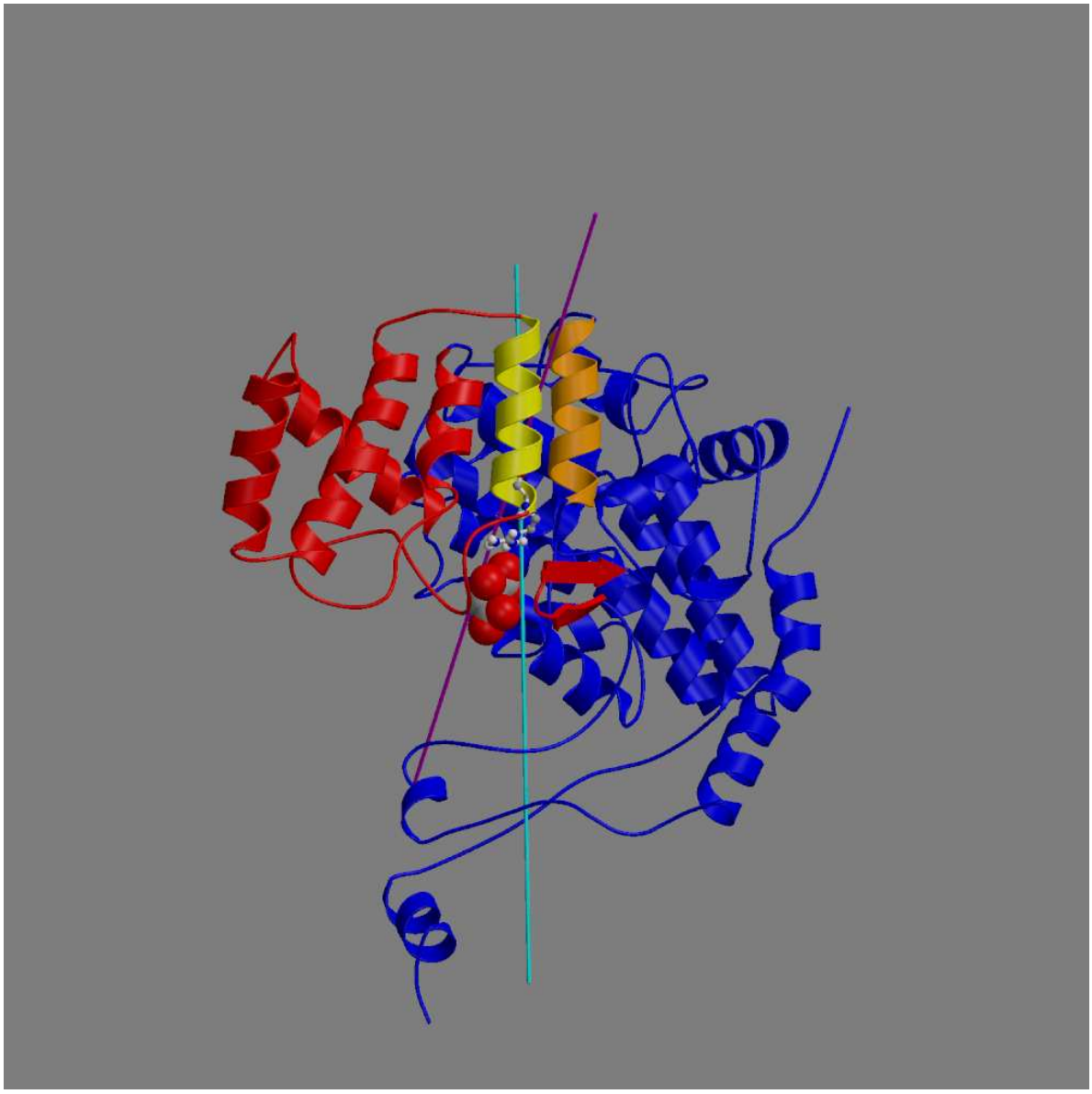


Figure 7.5: Schematic illustration of the paths taken by the trajectories of the targeted simulations in relation to the locations of the crystallographic domain conformations (see Figures 7.2 and 7.4). The filled circles indicate the crystallographic domain conformations. The unbroken arrows indicate the general direction taken by the targeted trajectories, which start from the crystallographic domain conformation indicated at the arrow's origin. The broken arrows point to the targeted crystallographic domain conformation from the final conformation located at the head of the unbroken arrow. In both cases the direction taken by the targeted simulations is rather parallel to, but does not follow the direct path between the two crystallographic domain conformations. In order to achieve these conformations a movement indicated by the broken arrows is required. However, in both cases it appears that this movement is unable to occur.

given by the right-hand rule). As the hinge axis describing the movement from starting to endpoint conformations in the targeting simulations is parallel to the helical axes, rotation about this hinge axis should not result in a shift of helix 328-341 relative to the large domain. The hinge axis describing the endpoint to target conformations, however, is not parallel to these helical axes and rotation about this hinge axis would result in a shift of helix 328-341 relative to the large domain. Thus the movement described by the crystallographic hinge axis is one the comprises two movements (see Figure 7.5), one a rotation about an axis parallel to the two helices, which does not result in a relative shift of these helices, the other a rotation that about an axis not parallel to the two helices which does result in a relative shift of the helices. The former does occur in our simulations, but the latter cannot. In order to quantify the shift in the helix



328-341 relative to the large domain, the program Helix_Shift was used (see Methods). In going from the crystallographic open to closed conformations this helix shifts -1.58A along its own axis relative to the large domain. The shift of this helix in each of the targeting simulations was calculated as the shift in the helix between the starting and endpoint conformations. Figure 7.7 schematically shows these distances. It confirms that the helix is unable to move up or down sufficiently as the domains rotate to reach the target conformation. Thus the shift of helix 328-341 relative to the large domain that occurs between the crystallographic open and closed conformations is unable to proceed sufficiently in our simulations. This result implies that the barrier to opening and closing is a barrier to the shift of this α-helix. Given that the open domain conformation does reach the closed domain conformation in the presence of the substrate

Figure 7.6: Backbone trace of the citrate synthase monomer. The large domain is coloured blue, the small domain red. α-helix 328-341 is coloured yellow, α-helix 222-235, orange. The cyan and magenta rods indicate the hinge axes for the domain movements indicated by the unbroken and broken arrows in Figure 7.5, respectively. These axes were calculated by selecting two pairs of conformations, one pair projecting parallel to the path taken in the targeted trajectories, and the other perpendicular to this path. These pairs of conformations were then passed to the program Dom_Select. The substrate, oxaloacetate is fitted into the large domain and is depicted in space-filling model. Arg329 situated at the base of α-helix 328-341, is shown in ball and stick model. In the crystallographic closed structure, Arg329 and oxaloacetate form a strong salt-bridge. It is thought that this interaction helps citrate synthase overcome the energy barrier in moving from the open to the fully closed state. The figure was created using RasMol, [223] Molscript [224] and Raster3D. [225]

oxaloacetate it is probably the interaction with oxaloacetate that is able to shift helix 328-341 relative to the large domain. Arg329 is situated at the base of this helix and is often assigned as a bending residue. In a sequential model of ligand binding and domain closure, oxaloacetate binds first to the large domain before closure occurs. [226] This would then put oxaloacetate in a position to interact with Arg329. This interaction creates a torque about the hinge axis helping to induce the closed conformation. In the closed conformation the salt-bridge between Arg329 and oxaloacetate is fully formed. The suggested movement that this interaction would induce is one that would shift the helix 328-341 downwards relative to the large domain (see Figure 7.6). Helix 328-341 is parallel to helix 222-235 in the large domain and they have many packing interactions. It would appear that the movement of helix 328-341 is quite constrained by these packing interactions and therefore the interaction between Arg329 and oxaloacetate is one that is consistent with these inter-helix contacts. Therefore our hypothesis is that it is primarily the interaction between oxaloacetate and Arg329 that is able to shift this helix downwards over the energy barrier. However, there is no easy explanation as to how the presence of the product might shift this helix back to its fully up position in the open conformation as citrate forms a strong salt bridge with Arg329 in the closed conformation. [219] If this hypothesis is correct then the energy barrier resides in the interactions between the parallel helices 222-235 and 328-341.

Importance of Arg329

Arg329 is conserved over all available citrate synthase sequences despite some of the sequences having diverged considerably. This was ascertained by using the Sequence Retrieval System, SRS, at the European Bioinformatics' Institute (http://srs/ebi.ac.uk) by selecting EC number 2.3.3.1 from the Enzyme database [227] and then linking

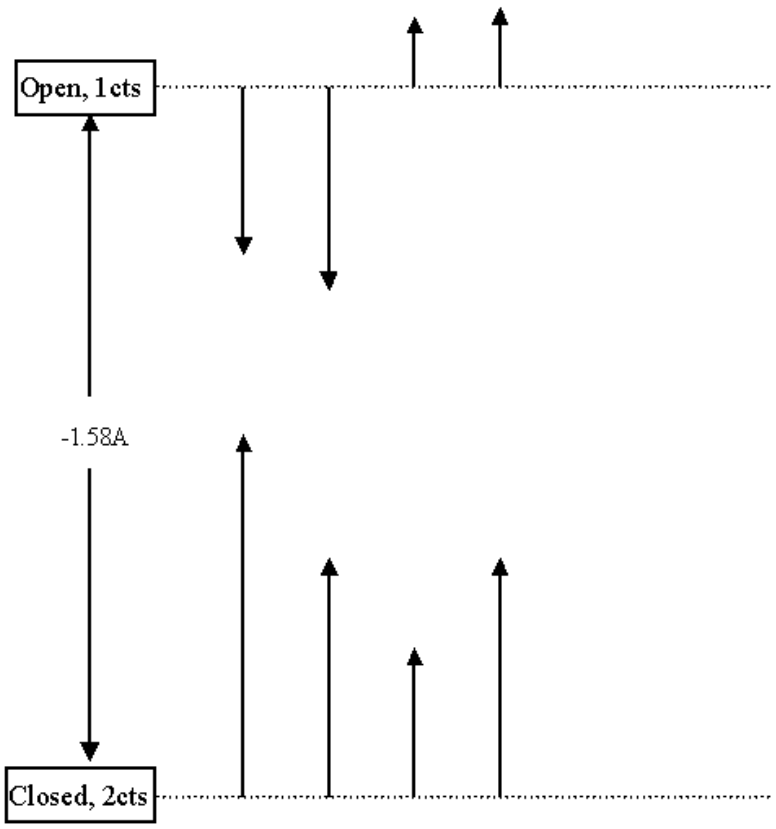


Figure 7.7: In going from the crystallographic open to closed structure the α-helix 328-341 shifts -1.58A along its own axis relative to the large domain. The arrows show the shifts of this helix relative to the large domain from the starting conformation in the targeted simulations. When starting from the open conformation and targeting to the crystallographic closed conformation this helix is unable to shift down sufficiently. Likewise when starting from the closed conformation and targeting to the crystallographic open conformation this helix is unable to shift up sufficiently. The distances were calculated by the program Helix_Shift and the lengths of the arrows in the figure correspond to the calculated distances.

to the Swiss-Prot protein sequence database. [228] All the sequences found were then aligned using Clustal W. [229] Arg329 is also the last residue in the citrate synthase PROSITE [230] motif: G-[FYA]-[GA]-H-X-[IV]-X(1,2)-[RKT]-X(2)-D-[PS]-R. In all available PDB structures of citrate synthase, this arginine is situated at the same structural position, namely at the N-terminus of one of two parallel α-helices. In all closed structures this arginine makes a salt bridge with either citrate or oxaloacetate. Unfortunately only two mutational studies of this residue are reported in the literature. Both these are on citrate synthase from *Escherichia coli* where this arginine is at position 314 in the sequence. A R314L mutant showed a complete lack of activity. [231] In another kinetics study, results on a R314Q and a D312N mutant were reported. [232] It was suggested that

either mutant may affect the step between citryl-coenzyme A formation and hydrolysis, and that it must involve both Arg314 and D312 acting in unison as both mutations produce a very similar effect. This latter point makes sense because Arg314 and Asp312 form a salt bridge and can therefore be understood as a conformational change involving both. This fits with our analysis because like Arg314, Asp312 is situated in the small domain suggesting that the forces acting on Arg314 are transmitted through to the rest of the small domain partly through their salt-bridge interaction so helping to induce closure. Thus without this interaction closure of the small domain upon the large would be impaired.

7.4 Conclusions

Targeting and exploring simulations have been performed to assess whether the crystallographic closed domain conformation is accessible to the unliganded enzyme from the crystallographic open conformation and vice-versa. The results here suggest that even with a strong bias introduced, the crystallographic closed domain conformation cannot be reached from the crystallographic open conformation. Likewise it would appear that the crystallographic open domain conformation is inaccessible from the crystallographic closed conformation. The closed conformation is unlikely to occur in the unliganded state as the enzyme needs to open to release citrate. This is supported by all available crystal structures which are all liganded in the active site when closed. It is likely therefore that the interaction with the products helps to recycle the enzyme back to the open conformation where they are finally able to escape. This means that the presence of a barrier between the open and closed conformations for the unliganded state is of particular relevance for the open conformation because it is only the open conformation that is likely to be unliganded. Our results indicate that the source of this energy barrier is related to the shift of α-helix 328-341 along its own axis relative to the large domain. In a sequential model of ligand binding and domain closure, [222] oxaloacetate binds first to the large domain. This would put it in a position to interact with Arg329 at the base of the helix such that it would pull this helix downwards in a direction that is roughly parallel with its own axis. Thus it is proposed that the source of the energy barrier lies in the interactions of this helix with its parallel partner in the large domain and that in going from the open to closed domain conformation it is the interaction primarily with Arg329 that provides the energy to overcome this barrier. In the unliganded state the domains are able to partially close or open but without the shift of this helix the full domain movement is unable to occur. Our results indicate that in citrate synthase the open unliganded enzyme will remain in a relatively open conformation ready to receive the substrate, thus increasing the efficiency of the enzyme over one that is closed part of the time, with its binding site inaccessible to the substrate. The substrate therefore is the key that is able to unlock the mechanism that prevents the domains to close fully. In this sense the substrate catalyses the domain closure.

Acknowledgements

I.D. thanks Prof. Alfredo Di Nola for his support.

CHAPTER

EIGHT

Concluding remarks

As summarized in the methods section of the present thesis, recent developments in computer simulations of biological macromolecules have enhanced the range of applicability of these techniques in the study of folding and misfolding processes. The methods used in this thesis form another contribution to this field and applications to several model systems have yielded interesting results.

The major problem with molecular dynamics (MD) simulations of the folding process of macromolecular systems, such as proteins, is due to the conformational sampling efficiency. This difficulty is also present in systems with a lower complexity, such as peptides, but is more tractable than for proteins. Experimentally, peptides fold at very fast rates, requiring probing on the nanosecond-microsecond time resolution, hence offering a unique opportunity to bridge the gap between theoretical and experimental understanding of protein folding. However, apart from some exceptions, the development and implementation of new sampling algorithms is commonly necessary to overcome the limitations of insufficient sampling for the study of more complex molecular systems.

For what concerns protein misfolding, understanding the conformational transitions featuring in the aggregation and amyloidogenesis of otherwise soluble proteins and peptides at atomic resolution would be of fundamental relevance for the development of effective therapies against amyloid related disorders. However, the insoluble and massive character of fibrils rules out the possibility to investigate their formation and their structure at atomic detail with conventional experimental techniques and the use of computational approaches becomes extremely useful, if not even necessary.

In the present thesis work, the determinants involved in fibril formation, in terms of atomic details of both the α-β conformational transition that is thought to trigger aggregation and the structure of nascent aggregates, are investigated. In particular, the transition from an ideal α-helix to a β-hairpin conformation of two well studied amyloidogenic peptides, the H1 peptide from prion protein and the Aβ(12–28) fragment from the Aβ(1-42) peptide responsible for Alzheimer disease, was here revealed for the first time by long time scale, all atom MD simulations in explicit water solvent. The

simulations highlight the unfolding of α-helices, followed by the formation of bent conformations and a final convergence to ordered in register β-hairpin conformations. The β-hairpins observed, despite different sequences, exhibit a common dynamic behaviour and the presence of a peculiar pattern of the hydrophobic side chains, in particular in the region of the turns. These observations hint at a possible common aggregation mechanism for the onset of different amyloid diseases and a common mechanism in the transition to the β-hairpin structures.

Further simulations of the H1 peptide at physiological conditions, for a total simulation time of $\approx$1.1 μs, provided an almost complete phase space sampling and a thermodynamic and kinetic characterization of its folding process was achieved. Several unfolding/refolding events of the β-hairpin structure are observed, yielding a very fast average β-hairpin folding time of $\approx$200 ns. The analysis of the peptide thermodynamic stability, reveals that the β-hairpin in solution is rather unstable. These results are in good agreement with several experimental evidences, according to which the isolated H1 peptide adopts very rapidly in water β-sheet structure leading to amyloid fibril precipitates.[119, 120] At our knowledge this is one of the first attempts to simulate the thermodynamic equilibrium of a complex system, such as a β-hairpin, for more then 1 μs using realistic models for both the peptide and the solvent and with a completely unbiased sampling of the configurational space.

The initial self-assembly stages of another fibrillogenic peptide, the core recognition motif of the type II diabetes associated islet amyloid polypeptide, was also studied by MD simulations. The simulations were performed using multiple replicas of the monomers in explicit water, in a confined box starting from a random distribution of the peptides. The formation of unique clusters is observed after a few nanoseconds. Structural analyses of the clusters clearly reveal the formation of "flat" ellipsoid-shaped clusters, showing a preferred locally parallel alignment of the peptides. The unique assembly is facilitated by a preference for an extended conformation of the peptides and by inter-molecular aromatic interactions.

For the study of more complex macromolecular systems, in the present thesis work a new enhanced sampling protocol, the Essential Dynamics Sampling (EDS), was utilized to study the folding process of an experimentally well studied protein, the cytochrome c, and to study an important enzyme, the citrate synthase, to directly address the question whether for this enzyme the ligand-induced closed domain conformation is accessible to the open unliganded enzyme.

In the EDS (see paragraph 2.3.2) the simulated protein is encouraged to adopt the conformation of a target structure, the X-ray folded structure in the case of cytochrome c and the closed domain conformation in the case of citrate synthase. In the case of cytochrome c, starting from structures with a root mean square deviation of $\sim$20 Å from the crystal structure, its correct folding was obtained using "only" 106 out of the total $\sim$3000 degrees of freedom of the protein in the EDS procedure. The folding pathways found in our simulations show that the early formations of contacts between the ter-

minal helices seems to be a prerequisite for a proper folding, in agreement with the hypothesized role of these contacts in the cytochrome c folding process revealed by experiments.

In the study of citrate synthase, we found that, when the enzyme is prepared from a crystallographic open-domain structure and is in the unliganded state, it is unable to adopt the crystallographic closed-domain conformation of the liganded enzyme. This result suggests that without the substrate the enzyme remains in a partially open conformation ready to receive the substrate, providing an increased efficiency over one that could be closed part of the time, with its binding site inaccessible to the substrate.

APPENDIX

A

Appendix

In this section the theory used to model the kinetics in the essential plane will be described in details.

Given a coordinate q, the ensemble mean square displacement from an initial point, as a function of time, can be expressed as:

$$\left\langle \Delta q^2(t) \right\rangle = 2 \int_0^t I(t')dt' \tag{A.1}$$

with $\Delta q(t) = q(t) - q(0)$ and

$$I(t') = \int_0^{t'} \gamma(t'')dt'' \tag{A.2}$$

where $\gamma(t'') = \langle \dot{q}(0)\dot{q}(t'') \rangle$ is the velocity autocorrelation function of q. As in the previous paper, [156] the function $I(t)$ is considered rapidly converging to a positive value within t_0, corresponding to a first fast relaxation of the order of 30-40 fs, while for $t > t_0$ a second slower, first order, relaxation is used to model the slowly converging tail in the velocity autocorrelation function. However, differently from the previous model where the diffusion was studied using relatively short time intervals (up to 20 ps) and a single-exponential mode was utilized, extension over longer time intervals (up to 100 ps) afforded in this study shows that a bi-exponential relaxation of the velocity autocorrelation function is necessary to model accurately the diffusion. Hence, considering $\gamma(t) = \gamma_1(t) + \gamma_2(t)$, equation A.1 becomes

$$\left\langle \Delta q^2(t) \right\rangle = 2 \left(\int_0^{t_0} I_1(t')dt' + \int_{t_0}^t I_1(t')dt' + \int_0^{t_0} I_2(t')dt' + \int_{t_0}^t I_2(t')dt' \right) \tag{A.3}$$

where

$$I_1(t') = \int_0^{t'} \gamma_1(t'')dt''$$
$$I_2(t') = \int_0^{t'} \gamma_1(t'')dt'' \tag{A.4}$$

If we assume, for $t > t_0$, a simple first order kinetics affecting the two components, then

$$I_1(t') = [I_1(t_0) - I_1(\infty)]e^{-(t'-t_0)/\tau_1} + I_1(\infty)$$
$$I_2(t') = [I_2(t_0) - I_2(\infty)]e^{-(t'-t_0)/\tau_2} + I_2(\infty) \tag{A.5}$$

with relaxation time constants τ_1 and τ_2. Therefore

$$\begin{aligned} \left\langle \Delta q^2(t) \right\rangle &= 2\Delta + 2I_1(\infty)(t-t_0) + 2[I_1(t_0) - I_1(\infty)]\tau_1[1 - e^{-(t-t_0)/\tau_1}] \\ &+ 2I_2(\infty)(t-t_0) + 2[I_2(t_0) - I_2(\infty)]\tau_2[1 - e^{-(t-t_0)/\tau_2}] \end{aligned} \tag{A.6}$$

with $\Delta = \int_0^{t_0} I(t')dt'$.

Finally, considering that for a time range up to 100 ps we can neglect the initial fast convergence, $\Delta, t_0 \approx 0$, equation A.6 becomes

$$\begin{aligned} \left\langle \Delta q^2(t) \right\rangle &\cong 2D_\infty t + 2[D_0 - A_1]\tau_1[1 - e^{-t/\tau_1}] \\ &+ 2[D_0 - A_2]\tau_2[1 - e^{-t/\tau_2}] \end{aligned} \tag{A.7}$$

where

$$\begin{aligned} D_0 &= I_1(t_0) + I_2(t_0) \\ D_\infty &= I_1(\infty) + I_2(\infty) \\ A_1 &= I_1(\infty) + I_2(t_0) \\ A_2 &= I_2(\infty) + I_1(t_0) \end{aligned}$$

Equation A.7 was used to evaluate the time behaviour of $\langle \Delta q^2(t) \rangle$ in the time range 1-100 ps. In particular three structurally different regions of the essential plane, where the coordinates do not encounter a relevant free energy gradient, were analyzed.

Bibliography

[1] C. Levinthal. *Mossbauer Spectroscopy in Biological Systems*. University of Illinois Press, Urbana, IL, P. Degennes edition, 1969.

[2] P.G. Wolynes, J.N. Onuchic, and D. Thirumalai. *Navigating the folding routes*. Science 267, 1619–1620 (1995).

[3] K. Dill and H. Chan. *From Levinthal to pathways to funnels*. Nat. Struct. Biol. 4, 10–19 (1997).

[4] M. Karplus. *The Levinthal paradox, yesterday and today*. Fold. Des. 2, 569–576 (1997).

[5] C. M. Dobson, A. Sali, and M. Karplus. *Protein Folding: A Perspective from Theory and Experiment*. Angew. Chem. Int. Ed. 37, 869–893 (1998).

[6] C.P. Schultz. *Illuminating folding intermediates*. Nat. Struct. Biol. 7, 7–10 (2000).

[7] B. Schuler, E. A. Lipman, and W. A. Eaton. *Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy*. Nature 419, 743–747 (2002).

[8] C.K. Chan, Y. Hu, S. Takahashi, D.L. Rousseau, W.A. Eaton, and J. Hofrichter. *Submillisecond protein folding kinetics studied by ultrarapid mixing*. Proc. Natl. Acad. Sci. USA 94, 1779–84 (1997).

[9] S. Takahashi, S.R. Yeh, T.K. Das, C.K. Chan, D.S. Gottfried, and D.L. Rousseau. *Folding of cytochrome c initiated by submillisecond mixing*. Nat. Struct. Biol. 1, 44–50 (1997).

[10] R. H. Callender, R. B. Dyer, R. Gilmanshin, and W. H. Woodruff. *Fast events in protein folding: the time evolution of primary processes*. Annu. Rev. Phys. Chem. 49, 173–202 (1998).

[11] T. Pascher. *Temperature and driving force dependence of the folding rate of reduced horse heart cytochrome c*. Biochemistry 40, 5812–5820 (2001).

[12] J. K. Myers and T. G. Oas. *Preorganized secondary structure as an important determinant of fast protein folding.* Nat. Struct. Biol. 8, 552–558 (2001).

[13] S. R. Yeh and D. L. Rousseau. *Hierarchical folding of cytochrome c.* Nat. Struct. Biol. 7, 443–445 (2000).

[14] J.N. Onuchic, Z. Luthey-Schulten, and P.G. Wolynes. *Theory of protein folding: the energy landscape perspective.* Ann. Rev. Phys. Chem. 48, 545–600 (1997).

[15] J. E. Shea and C. L. Brooks III. *From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding.* Annu. Rev. Phys. Chem. 52, 499–535 (2001).

[16] Y. Duan and P.A. Kollman. *Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution.* Science 282, 740–744 (1998).

[17] W.F. van Gunsteren, R. Bürgi, C. Peter, and X. Daura. *The Key to Solving the Protein-Folding Problem Lies in an Accurate Description of the Denatured State.* Angew. Chemie Intl. Ed. 40, 351–355 (2001).

[18] X. Daura, W.F. van Gunsteren, and A.E. Mark. *Folding-unfolding thermodynamics of a β-heptapeptide from equilibrium simulations.* PROTEINS: Struct. Funct. Gen. 34, 269–280 (1999).

[19] J.D. Harper. *Models of amyloid seeding in Alzheimer disease and scrapie: mechanistic truths and physiological consequences of the time-dependent solubility of amyloid proteins.* Annu. Rev. Biochem. 66, 385–407 (1997).

[20] M. Sunde and C.C.F. Blake. *From the globular to the fibrous state: Protein structure and structural conversion in amyloid formation.* Quart. Rev. Biophys. 31, 1–39 (1998).

[21] C. M. Dobson. *Protein misfolding, evolution and disease.* Trends Biochem Sci. 24, 329–332 (1999).

[22] J.D. Sipe and A.S. Cohen. *Review: History of the amyloid fibril.* J. Struct. Biol. 130, 88–98 (2000).

[23] E. Gazit. *Mechanistic studies of the process of amyloid fibrils formation by the use of peptide fragments and analogues: Implications for the design of fibrillization inhibitors.* Curr. Med. Chem. 9, 1725–1735 (2002).

[24] M. Sunde and C. Blake. *The structure of amyloid fibrils by electron microscopy and X-ray diffraction.* Adv. Protein Chem. 50, 123–159 (1997).

[25] M. Barteri and B. Pispisa. *Influence of isopropanol-water solvent mixtures on the conformation of poly-L-lysine.* Biopolymers 12, 2309–2327 (1973).

[26] K. M. Pan, M. Baldwin, J. Nguyen, M. Gasset, A. Serban, D. Groth, I. Mehlhorn, Z. Huang, R. J. Fletterick, F. E. Cohen, and S. B. Prusiner. *Conversion of α-helices into β-sheets features in the formation of the scrapie prion proteins.* Proc. Natl. Acad. Sci. USA 90, 10962–10966 (1993).

[27] D. Jayawickrama, S. Zink, D. Vander Velde, R.I. Effiong, and C.K. Larive. *Conformational analysis of the β-amyloid peptide fragment, β(12-28).* J. Biomol. Struct. Dyn. 13, 229–244 (1995).

[28] D. Peretz, R.A. Williamson, Y. Matsunaga, H. Serban, C. Pinilla, R.B. Bastidas, R. Rozenshteyn, T.L. James, R.A. Houghten, F.E. Cohen, S.B. Prusiner, and D.R. Burton. *A conformational transition at the N terminus of the prion protein features in formation of the scrapie isoform.* J. Mol. Biol. 273, 614–622 (1997).

[29] H. Mihara, Y. Takahashi, and A. Ueno. *Design of peptides undergoing self-catalyitic α to β transition and amyloidogenesis.* Biopolymers 47, 83–92 (1998).

[30] M. P. Allen and D. J. Tildesly. *Computer simulation of liquids.* Oxford University Press, Oxford, 1989.

[31] D. Frenkel and B. Smit. *Understanding Molecular Simulation. From Algorithms to Applications.* Academic Press, Boston, 1996.

[32] W. F. van Gunsteren and P. K. Weiner. *Computer simulation of biomolecular systems.* Escom Science, Leiden (NL), 1989.

[33] B. R. Brooks, R. Bruccoleri, B. Olafson, D. States, S. Swaminathan, and M. Karplus. *CHARMM: a program for macromolecular energy, minimization, and dynamics calculations.* J. Comput. Chem. 4, 187 (1983).

[34] S. Weiner, P. Kollman, D. Nguyen, and D. Case. *An all atom force field for simulations of proteins and nucleic acids.* J. Comput. Chem. 7, 230 (1986).

[35] 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.

[36] L. Verlet. *Computer "experiments" on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules.* Phys. Rev. 159, 98 (1967).

[37] L. Verlet. *Computer "experiments" on classical fluids. II. Equilibrium correlation functions.* Phys. Rev. 165, 201–23 (1968).

[38] D. Beeman. *Some multistep methods for use in molecular dynamics calculations.* J. Comput. Phys. 20, 130 (1976).

[39] C. Gear. *Numerical Initial Value Problems in Ordinary Differential Equations*. Englewood Cliffs, NJ, USA, 1971.

[40] J. McCammon, B. Gelin, and M. Karplus. *Dynamics of folded proteins*. Nature 267, 585 (1977).

[41] T. Schlick, E. Barth, and M. Mandziuk. *BIOMOLECULAR DYNAMICS AT LONG TIMESTEPS:Bridging the Timescale Gap Between Simulation and Experimentation*. Annu. Rev. Biomol. Struct. 26, 181–222 (1997).

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

[43] A. Amadei, G. Chillemi, M. A. Ceruso, A. Grottesi, and A. Di Nola. *Molecular dynamics simulations with 0 roto-traslational motions: Theoretical basis and statistical mechanical consistency*. J. Chem. Phys. 112, 9–23 (2000).

[44] H. Bekker, J. P. van den Berg, and T. A. Wassenaar. *A method to obtain a near-minimal-volume molecular simulation of a macromolecule, using periodic boundary conditions and rotational constraints*. J. Comput. Chem. 25, 1037–1046 (2004).

[45] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. *A smooth particle mesh Ewald method*. J. Chem. Phys. 103, 8577–8593 (1995).

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

[47] D. Brown and J. Clarke. *Molecular dynamics simulations of polymer fiber microstructure*. J. Chem. Phys 84, 2858 (1986).

[48] S. Nosè. *Constant temperature molecular dynamics methods*. Prog. Theoret. Phys. Supplement 103, 1–46 (1991).

[49] W. Hoover. *Canonical dynamics: Equilibrium phase-space distributions*. Phys. Rev. A31, 1695–1697 (1985).

[50] G. J. Martyna, M. L. Klein, and M. Tuckerman. *Nosé Hoover chains: The canonical ensemble via continuous dynamics*. J. Chem. Phys. 97, 2635–2645 (1992).

[51] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen. *Essential Dynamics of Proteins*. Proteins: Struct. Funct. Genet. 17, 412–425 (1993).

[52] A.E. García. *Large-amplitude nonlinear motions in proteins*. Phys. Rev. Lett. 66, 2696–2699 (1992).

[53] A. Kitao, F. Hirata, and N. Gō. *The effects of solvent on the conformation and the collective motions of protein: normal mode analysis and molecular dynamics simulations of melittin in water and in vacuum.* J. Chem. Phys. 158, 447–472 (1991).

[54] D. M. van Aalten, A. Amadei, A. B. Linssen, V. G. Eijsink, G. Vriend, and H. J. C. Berendsen. *The essential dynamics of thermolysin: confirmation of the hinge-bending motion and comparison of simulations in vacuum and water.* PROTEINS: Struct. Funct. Gen. 22, 45–54 (1995).

[55] B. L. de Groot, D. M. F. van Aalten, A. Amadei, and H. J. C. Berendsen. *Domain motions in Bacteriophage T4 lysozyme: a comparison between molecular dynamics and crystallographic data.* Proteins: Struct. Funct. Genet. 31, 116–127 (1998).

[56] D. M. van Aalten, D. A. Conn, B. L. de Groot, H. J. C. Berendsen, J. B. Findlay, and A. Amadei. *Protein dynamics derived from clusters of crystal structures.* Biophys. J. 73, 2891–2896 (1997).

[57] R. Abseher, L. Horstink, C. W. Hilbers, and M. Nilges. *Essential spaces defined by NMR structure ensembles and molecular dynamics simulation show significant overlap.* PROTEINS: Struct. Funct. Gen. 31, 370–382 (1998).

[58] B. L. de Groot, D. M. van Aalten, R. M. Scheek, A. Amadei, G. Vriend, and H. J. C. Berendsen. *Prediction of protein conformational freedom from distance constraints.* PROTEINS: Struct. Funct. Gen. 29, 240–251 (1997).

[59] Y. Ueda, H. Taketomi, and N. Gō. *Studies on protein folding, unfolding and fluctuations by computer simulation. II. A three-dimensional lattice model of lysozyme.* Biopolymers 17, 1531–1548 (1978).

[60] A.M. Gutin, V.I. Abkevich, and E.I. Shakhnovich. *A Protein Engineering Analysis Of The Transition State For Protein Folding: Simulation In The Lattice Model.* Fold. Des. 3, 183–194 (1998).

[61] D.K. Klimov and D. Thirumalai. *Mechanisms and kinetics of β-hairpin formation.* Proc. Natl. Acad. Sci. USA 97, 2544–2549 (2000).

[62] D. G. Covell and R. L. Jernigan. *Conformations of folded proteins in restricted spaces.* Biochemistry 29, 3287–3294 (1990).

[63] A. Kolinski and J. Skolnick. *Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme.* PROTEINS: Struct. Funct. Gen. 18, 338–352 (1994).

[64] A. Sali, E. I. Shakhnovich, and M. Karplus. *Kinetics of protein folding. A lattice model study of the requirements for folding to the native state.* J. Mol. Biol. 235, 1614–1636 (1994).

[65] K. A. Dill, S. Bromberg, K. Yue, K. M. Fiebig, D. P. Yee, P. D. Thomas, and H. S. Chan. *Principles of protein folding: a perspective from simple exact models.* Protein Sci. 4, 561–602 (1995).

[66] S. Chandrasekhar. *Stochastic Problems in Physics and Astronomy.* Rev. Mod. Phys. 15, 1–89 (1943).

[67] H. Grübmuller and P. Tavan. *Molecular dynamics of conformational substates for a simplified protein model.* J. Chem. Phys. 101, 5047–5057 (1994).

[68] J. D. Honeycutt and D. Thirumalai. *Metastability of the Folded States of Globular Proteins.* Proc. Natl. Acad. Sci. USA 87, 3526–3529 (1990).

[69] R. Srinivasan and G. D. Rose. *LINUS - A hierarchic procedure to predict the fold of a protein.* PROTEINS: Struct. Funct. Gen. 22, 81–99 (1995).

[70] G. F. Berriz, A. M. Gutin, and E. I. Shakhnovich. *Cooperativity and Stability In A Langevin Model Of Protein-Like Folding.* J. Chem. Phys. 106, 9276–9285 (1997).

[71] D. T. Jones, W. R. Taylor, and J. M. Thornton. *A new approach to protein fold recognition.* Nature 358, 86–89 (1992).

[72] A. E. Torda. *Perspectives in protein-fold recognition.* Curr. Opin. Struct. Biol. 7, 200–205 (1997).

[73] D. T. Jones and J. M. Thornton. *Potential energy functions for threading.* Curr. Opin. Struct. Biol. 6, 210–206 (1996).

[74] P. Fariselli, M. Compiani, and R. Casadio. *Predicting secondary structures of membrane proteins with neural networks.* Eur. Biophys. J. 22, 41–51 (1993).

[75] M. Compiani, P. Fariselli, P. L. Martelli, and R. Casadio. *An entropy criterion to detect minimally frustrated intermediates in native proteins.* Proc. Natl. Acad. Sci. USA 95, 9290–9294 (1998).

[76] M. Schaefer, C. Bartels, and M. Karplus. *Solution conformations and thermodynamics of structured peptides: molecular dynamics simulation with an implicit solvation model.* J. Mol. Biol. 284, 835–848 (1998).

[77] H. Wang, J. Varady, L. Ng, and S.S. Sung. *Molecular dynamics simulations of β-hairpin folding.* PROTEINS: Struct. Funct. Gen. 37, 325–333 (1999).

[78] J.P. Ulmschneider and W.L. Jorgensen. *Polypeptide Folding Using Monte Carlo Sampling, Concerted Rotation, and Continuum Solvation.* J. Am. Chem. Soc. 126, 1849–1857 (2004).

[79] C.D. Snow, L. Qiu, D. Du, F. Gai, S.J. Hagen, and V.S. Pande. *Trp zipper folding kinetics by molecular dynamics and temperature-jump spectroscopy.* Proc. Natl. Acad. Sci. USA 101, 4077–4082 (2004).

[80] H. Nymeyer and A.E. García. *Simulation of folding equilibrium of α-helical peptides: a comparison of the generalized Born approximation with explicit solvent.* Proc. Natl. Acad. Sci. USA 100, 13934–13939 (2003).

[81] B. D. Bursulaya and C. L. Brooks III. *Comparative study of the folding free energy landscape of a three-stranded β-sheet protein with explicit and implicit solvent models.* J. Phys. Chem. B 104, 12378–12383 (2000).

[82] R. Zhou and B.J. Berne. *Can a continuum solvent model reproduce the free energy landscape of a β-hairpin folding in water?* Proc. Natl. Acad. Sci. USA 99, 12777–12782 (2002).

[83] P. Ferrara, J. Apostolakis, and A. Caflisch. *Thermodynamics and kinetics of folding of two model peptides investigated by molecular dynamics simulations.* J. Phys. Chem. B 104, 5000–5010 (2000).

[84] P. Ferrara and A. Caflisch. *Folding simulations of a three-stranded antiparallel beta-sheet peptide.* Proc. Natl. Acad. Sci. USA 20, 10780–10785 (2000).

[85] P. Ferrara and A. Caflisch. *Native topology or specific interactions: what is more important for protein folding?* J. Mol. Biol. 306, 837–850 (2001).

[86] A. Cavalli, P. Ferrara, and A. Caflisch. *Weak temperature dependence of the free energy surface and folding pathways of structured peptides.* PROTEINS: Struct. Funct. Gen. 47, 305–314 (2002).

[87] A. Hiltpold, P. Ferrara, J. Gsponer, and A. Caflisch. *Free energy surface of the helical peptide Y(MEARA)$_6$.* J. Phys. Chem. B 104, 10080–10086 (2000).

[88] M. Shirts and V. S. Pande. *Computing: screen savers of the world unite!* Science 290, 1903–1904 (2000).

[89] A. F. Voter. *Parallel replica method for dynamics of infrequent events.* Phys. Rev. B-Cond. Matt. 57, 13985–13988 (1998).

[90] B. Zagrovic, E. J. Sorin, and V. S. Pande. *β-hairpin folding simulations in atomistic detail using an implicit solvent model.* J. Mol. Biol. 313, 151–169 (2002).

[91] A. R. Fersht. *On the simulation of protein folding by short time scale molecular dynamics and distributed computing*. Proc. Natl. Acad. Sci. USA 99, 14122–14125 (2002).

[92] Y. Sugita and Y. Okamoto. *Replica-exchange multicanonical algorithm and multicanonical replica-exchange method for simulating systems with rough energy landscape*. Chem. Phys. Lett. 329, 261–270 (2000).

[93] A.E. García and K.Y. Sanbonmatsu. *Exploring the energy landscape of β hairpin in explicit solvent*. PROTEINS: Struct. Funct. Gen. 42, 345–354 (2001).

[94] E.M. Boczko and C. L. Brooks III. *First-principle calculation of the folding free energy of a three-helix bundle protein*. Science 269, 393–396 (1995).

[95] C.L. Brooks. *Protein and peptide folding explored with molecular simulations*. Acc. Chem. Res. 35, 447–454 (2002).

[96] B.A. Berg and T. Neuhaus. *Multicanonical algorithms for 1st order phase-transitions*. Phys. Lett. B 267, 249–253 (1991).

[97] U. H. E. Hansmann and Y. Okamoto. *Prediction of peptide conformation by multicanonical algorithm: new approach to the multiple-minima problem*. J. Comput. Chem. 14, 1333–1338 (1993).

[98] N. A. Alves and U. H.E . Hansmann. *Helix formation and folding in an artificial peptide*. J. Chem. Phys. 117, 2337–2343 (2002).

[99] N. Kamiya, J. Higo, and H. Nakamura. *Conformational transition states of a beta-hairpin peptide between the ordered and disordered conformations in explicit water*. Protein Sci. 11, 2297–2307 (2002).

[100] D.M. Korzhnev, X. Salvatella, M. Vendruscolo, A. A. Di Nardo, A. R. Davidson, C. M. Dobson, and L. E. Kay. *Low-populated folding intermediates of Fyn SH3 characterized by relaxation dispersion NMR*. Nature 29, 586–990 (2004).

[101] K. Lindorff-Larsen, M. Vendruscolo, E. Paci, and C. M. Dobson. *Transition states for protein folding have native topologies despite high structural variability*. Nat. Struct. Mol. Biol. 11, 443–449. (2004).

[102] K. Lindorff-Larsen, S. Kristjansdottir, K. Teilum, W. Fieber, C. M. Dobson, F. M. Poulsen, and M. Vendruscolo. *Determination of an ensemble of structures representing the denatured state of the bovine acyl-coenzyme a binding protein*. J. Am. Chem. Soc. 126, 3291–3299 (2004).

[103] A. Amadei, A. B. M. Linssen, B. L. de Groot, D. M. van Aalten, and H. J. C. Berendsen. *An efficient method for sampling the essential subspace of proteins*. J. Biomol. Struct. Dyn. 13, 615–625 (1996).

[104] B. L. de Groot, A. Amadei, R. M. Scheek, N. A. van Nuland, and H. J. C. Berendsen. *An extended sampling of the configurational space of HPr from E. coli*. Proteins: Struct. Funct. Genet. 26, 314–322 (1996).

[105] U. Mayor, C. M. Johnson, V. Daggett, and A. R. Fersht. *Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation*. Proc. Natl. Acad. Sci. USA 97, 13518–13522 (2000).

[106] D.O.V. Alonso and V. Daggett. *Staphylococcal protein A: unfolding pathways, unfolded states, and differences between the B and E domains*. Proc. Natl. Acad. Sci USA 97, 133–138 (2000).

[107] Y. Pan and V. Daggett. *Direct comparison of experimental and calculated folding free energies for hydrophobic deletion mutants of chymotrypsin inhibitor 2: free energy perturbation calculations using transition and denaturated states from molecular dynamics of unfolding*. Biochemistry. 40, 2723–2731 (2001).

[108] D. Roccatano, I. Daidone, M.-A. Ceruso, C. Bossa, and A. Di Nola. *Selective excitation of native fluctuations during thermal unfolding simulations: horse heart cytochrome c as a case study*. Bioph. J. 84, 1876–1883 (2003).

[109] A. V. Finkelstein. *Can protein unfolding simulate protein folding?* Protein Eng. 10, 843–845 (1997).

[110] D.A. McQuarrie. *Statistical mechanics*. Harper & Row, New York, 1976.

[111] J. K. Kirkwood. *Statistical mechanics of fluid mixtures*. J. Chem. Phys. 3, 300–313 (1935).

[112] M. Mezei and D. L. Beveridge. *Computer Simulations and Biomolecular Systems*. Ann. New York Academy of Sciences, New York, d.L. Beveridge and W.L. Jorgensen edition, 1986.

[113] B. S. Shastry. *Neurodegenerative disorders of protein aggregation*. Neurochem Int. 43, 1–7 (2003).

[114] S.B Prusiner. *Prions*. Proc. Natl. Acad. Sci. USA 95, 13363–13383 (1998).

[115] D.J. Selkoe. *Alzheimer's disease: genes, proteins, and therapy*. Physiol. Rev. 81, 741–766 (2001).

[116] R. Riek, S. Hornemann, G. Wider, M. Billeter, R. Glockshuber, and K. Wuthrich. *NMR structure of the mouse prion protein domain PrP(121-321)*. Nature 382, 180–182 (1996).

[117] D. R. Borchelt, M. Scott, A. Taraboulos, N. Stahl, and S. B. Prusiner. *Scrapie and cellular prion proteins differ in the kinetics of synthesis and topology in cultured cells*. J. Cell Biol. 110, 743–752 (1990).

[118] S.B Prusiner, M. R. Scott, S. J. DeArmond, and F. E. Cohen. *Prion protein biology*. Cell 93, 337–348 (1998).

[119] J. Nguyen, M. A. Baldwin, F. E. Cohen, and S. B. Prusiner. *Prion protein peptides induce α-helix to β-sheet conformational transitions*. Biochemistry 34, 4186–4192 (1995).

[120] H. Inouye and D. A. Kirschner. *Polypeptide chain folding in the hydrophobic core of hamster scrapie prion: analysis by X-ray diffraction*. J. Struct. Biol. 122, 247–255 (1998).

[121] K. Kaneko, D. Peretz, K. M. Pan, T.C. Blochberger, H. Wille, R. Gabizon, O.H. Griffith, F.E. Cohen, M.A. Baldwin, and S.B. Prusiner. *Prion Protein (PrP) Synthetic Peptides Induce Cellular PrP to Acquire Properties of the Scrapie Isoform*. Proc. Natl. Acad. Sci. USA 92, 11160–11164 (1995).

[122] J. F. Flood, J. E. Morely, and E. Roberts. *Amnestetic effects in mice of four synthetic peptide homologous to amyloid β–protein in patients with Alzheimer's disease*. Proc. Natl. Acad. Sci. USA 88, 3363–3366 (1991).

[123] J. F. Flood, J. E. Morely, and E. Roberts. *An amyloid β–protein fragment, Aβ–(12–28), equipotently impairs post-training memory processing when injected into different limbic system structures*. Brain Res. 663, 271–276 (1994).

[124] P. E. Fraser, J. T. Nguyen, W. K. Surewiez, and D. A. Kirschner. *pH dependent structural transitions of Alzheimer's amyloid peptides*. Biophys. J. 60, 1190–1201 (1991).

[125] P. E. Fraser, L. Levesque, and D. R. McLachlan. *Alzheimer's Aβ–amyloid forms an inhibitory neuronal substrate*. J. Neurochem. 62, 1227–1230 (1994).

[126] J. Heller, A. C. Kolbert, R. Larsen, M. Ernst, T. Bekker, M. Baldwin, S. B. Prusiner, A. Pines, and D. E. Wemmer. *Solid-state NMR studies of the prion protein H1 fragment*. Protein Sci. 5, 1655–1661 (1996).

[127] Y. Levy, E. Hanan, B. Solomon, and O. M. Becker. *Helix-coil transition of PrP106-126: molecular dynamic study*. PROTEINS: Struct. Funct. Gen. 45, 382–396 (2001).

[128] D. K. Klimov and D. Thirumalai. *Dissecting the Assembly of Aβ(16–22) Amyloid peptides into antiparallel β-sheets*. Structure 11, 295–307 (2003).

[129] J. E. Straub, J. Guevara, S. Huo, and J. P. Lee. *Long time dynamic simulations: exploring the folding pathways of an Alzheimer's Amyloid Aβ peptide*. Acc. Chem. Res. 35, 473–481 (2002).

[130] J. Gsponer, U. Haberthuer, and A. Caflisch. *The role of side chain interactions in the early steps of aggregation: Molecular dynamics simulations of an amyloid–forming peptide from yeast prion Sup35.* Proc. Natl. Acad. Sci. USA 100, 5154–5159 (2003).

[131] B. Y. Ma and R. Nussinov. *Stabilities and conformations of Alzheimer's β-amyloid peptide oligomers (A β(16-22), A β(16-35) and Aβ(10-35)): Sequence effects.* Proc. Natl. Acad. Sci. USA 99, 14126–14131 (2002).

[132] D. Roccatano, G. Colombo, M. Fioroni, and A.E. Mark. *Mechanism by which 2,2,2-trifluoroethanol/water mixtures stabilize secondary-structure formation in peptides: a molecular dynamics study.* Proc. Natl. Acad. Sci. USA 99, 12179–12184 (2002).

[133] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraaije. *LINCS: A linear constraint solver for molecular simulations.* J. Comp. Chem. 18, 1463–1472 (1997).

[134] D. van der Spoel, R. van Drunen, and H. J. C. Berendsen. *GROningen MAchine for Chemical Simulation.* Department of Biophysical Chemistry, BIOSON Research Institute, Nijenborgh 4 NL-9717 AG Groningen, 1994.

[135] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma. *The missing term in effective pair potentials.* J. Phys. Chem. 91, 6269–6271 (1987).

[136] M. Fioroni, K. Burger, A.E. Mark, and D. Roccatano. *A new 2,2,2-trifluoroethanol model for Molecular dynamics simulations.* J. Phys. Chem. B 104, 12347–12354 (2000).

[137] J. Jarvet, P. Damberg, K. Bodell, L. E. Goran Eriksson, and A. Graslund. *Reversible random coil to β–sheet transition and the early stage of aggregation of the Aβ(12–28) fragment from the Alzheimer peptide.* J. Am. Chem. Soc. 122, 4261–4268 (2000).

[138] R. A. Jarvis and E. A. Patrick. *Clustering using a similarity measure based on shared near neighbors.* IEEE Trans. Comp. 22, 1025–1034 (1973).

[139] W. Kabsch and C. Sander. *Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features.* Biopolymers 22, 2577–2637 (1983).

[140] Z. Huang, S. B. Prusiner, and F. E. Cohen. *Scrapie prions: a three-dimensional model of an infectious fragment.* Fold. Des. 1, 13–19 (1996).

[141] D.O.V. Alonso, S.J. DeArmond, F.E. Cohen, and V. Daggett. *Mapping the early steps in the pH-induced conformational conversion of the prion protein.* Proc. Natl. Acad. Sci. USA 98, 2985–2989 (2001).

[142] L. C. Serpell, C. C. F. Blake, and P. E. Fraser. *Molecular Structur of a fibrillar Alzheimer's Aβ–fragment.* Biochemistry 39, 13269–13275 (2000).

[143] K.S. Satheeshkumar and R. Jayakumar. *Conformational polymorphism of the amyloidogenic peptide homologous to residues 113-127 of the prion protein.* Biophys. J. 85, 473–483 (2003).

[144] A. Jasanoff and A. Fersht. *Quantitative determination of helical propensities from trifluoroethanol titration curves.* Biochemistry 33, 2129–2135 (1994).

[145] E. Lacroix, T. Kortemme, M. Lopez de la Paz, and L. Serrano. *The design of linear peptides that fold as monomeric β-sheet structures.* Cur.r Opin. Struct. Biol. 9, 487–493 (1999).

[146] D. R. Booth, M. Sundet, V. Bellotti, C. V. Robinson, W. L. Hutchinson, P. E. Fraser, P. N. Hawkins, C. M. Dobson, S. E. Radford, C. C. F Blake, and M. B. Pepys. *Instability, unfolding and aggregation of human lysozime variants underlying amyloid fibrillogenesis.* Nature. 385, 787–793 (1997).

[147] M. Fandrich, M. A. Fletcher, and C. M. Dobson. *Amyloid fibrils from muscle myoglobin.* Nature. 410, 165–166 (2001).

[148] M. Lopez De La Paz, K. Goldie, J. Zurdo, E. Lacroix, C.M. Dobson, A. Hoenger, and L. Serrano. *De novo designed peptide-based amyloid fibrils.* Proc. Natl. Acad. Sci. USA 99, 16052–16057 (2002).

[149] B. Erman and K. A. Dill. *Gaussian Theory of Protein Folding.* J. Chem. Phys. 112, 1050–1056 (2000).

[150] Y. Zhou and M. Karplus. *Interpreting the folding kinetics of helical proteins.* Nature 401, 400–403 (1999).

[151] I. Daidone, A. Amadei, D. Roccatano, and A. Di Nola. *Molecular dynamics simulation of protein folding by essential dynamics sampling: folding landscape of horse heart cytochrome C.* Biophys. J. 85, 2865–2871 (2003).

[152] S. Gnanakaran, H. Nymeyer, J. Portman, K.Y. Sanbonmatsu, and A.E. García. *Peptide folding simulations.* Curr. Opin. Struct. Biol. 13, 168–174 (2003).

[153] W.Y. Yang, J.W. Pitera, W.C. Swope, and M. Gruebele. *Heterogeneous folding of the trpzip hairpin: full atom simulation and experiment.* J. Mol. Biol. 336, 241–251 (2004).

[154] A. Mitsutake, Y. Sugita, and Y. Okamoto. *Generalized-ensemble algorithms for molecular simulations of biopolymers.* Biopolymers 60, 96–123 (2001).

[155] I. Daidone, F. Simona, D. Roccatano, R. A. Broglia, G. Tiana, G. Colombo, and A. Di Nola. *β hairpin conformation of fibrillogenic peptides: structure and α-β transition mechanism revealed by molecular dynamics simulations.* PROTEINS: Struct. Funct. Bioinf. 57, 198–204 (2004).

[156] A. Amadei, M. A. Ceruso, and A. Di Nola. *On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins molecular dynamics simulations.* Proteins: Struct. Funct. Genet. 36, 419–424 (1999).

[157] T. Darden, D. York, and L. Pedersen. *Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems.* J. Chem. Phys. 98, 10089–10092 (1993).

[158] D. Brown and J. H. R. Clarke. *A comparison of constant energy, constant temperature, and constant pressure ensembles in molecular dynamics simulations of atomic liquids.* Mol. Phys. 51, 1243–1252 (1984).

[159] V. Muñoz, P.A. Thompson, J. Hofrichter, and W.A. Eaton. *Folding dynamics and mechanism of β-hairpin formation.* Nature 390, 196–199 (1997).

[160] R. Tycko. *Progress towards a molecular-level structural understanding of amyloid fibrils.* Curr. Opin. Struct. Biol. 14, 96–103 (2004).

[161] M.I. Ivanova, M. Gingery, L.J. Whitson, and D. Eisenberg. *Role of the C-terminal 28 residues of beta2-microglobulin in amyloid fibril formation.* Biochemistry 42, 13536–13540 (2003).

[162] Y. Mazor, S. Gilead, I. Benhar, and E. Gazit. *Identification and characterization of a novel molecular-recognition and self-assembly domain in the islet amyloid polypeptide.* J. Mol. Biol. 322, 1013–1024 (2002).

[163] F. Chiti, P. Webster, N. Taddei, A. Clark, M. Stefani, G. Ramponi, and C.M. Dobson. *Designing conditions for in vitro formation of amyloid protofilaments and fibrils.* Proc. Natl. Acad. Sci. USA 96, 3590–3594 (1999).

[164] J.W. Kelly. *Amyloid fibril formation and protein miniassembly: a structural quest for insight into amyloid and prion diseases.* Structure 5, 595–600 (1997).

[165] R.A. Kammerer, D. Kostrewa, J. Zurdo, A. Detken, C. Garcia-Echeverria, J.D. Green, S.A. Muller, B.H. Meier, F.K. Winkler, C.M. Dobson, et al. *Exploring amyloid formation by a de novo design.* Proc. Natl. Acad. Sci. USA 101, 4435–4440 (2004).

[166] K. Tenidis, M. Waldner, J. Bernhagen, W. Fischle, M. Bergmann, M. Weber, M.L. Merkle, W. Voelter, H. Brunner, and A. Kapurniotu. *Identification of a penta- and hexapeptide of islet amyloid polypeptide (IAPP) with amyloidogenic and cytotoxic properties.* J. Mol. Biol. 295, 1055–1071 (2000).

[167] R. Azriel and E. Gazit. *Analysis of the structural and functional elements of the minimal active fragment of islet amyloid polypeptide (IAPP): An experimental support for the key role of the phenylanaline residue in amyloid formation.* J. Biol. Chem. 276, 34156–34161 (2001).

[168] Y. Porat, A. Stepensky, F.X. Ding, F. Naider, and E. Gazit. *Completely different amyloidogenic potential of nearly identical peptide fragments.* Biopolymers 69, 161–164 (2003).

[169] M. Reches, Y. Porat, and E. Gazit. *Amyloid fibrils formation by pentapeptide and tetrapeptide fragments of human calcitonin.* J. Biol. Chem. 277, 35475–35480 (2002).

[170] M. Reches and E. Gazit. *Amyloidogenic hexapeptide fragment of medin: Implications for the stacking model of fibrillization.* Amyloid in press (2004).

[171] D. Zanuy, B. Ma, and R. Nussinov. *Short peptide amyloid organization: stabilities and conformations of the islet amyloid peptide NFGAIL.* Biophys. J. 84, 1884–1894 (2003).

[172] D. Zanuy and R. Nussinov. *The Sequence dependence of fiber organization. A comparative molecular dynamics study of the islet amyloid polypeptide segments 22-27 and 22-29.* J. Mol. Biol. 329, 565–584 (2003).

[173] D. Zanuy, Y. Porat, E. Gazit, and R. Nussinov. *Peptide sequence and amyloid formation; molecular simulations and experimental study of a human islet amyloid polypeptide fragment and its analogs.* Structure 12, 439–455 (2004).

[174] R. Chelli, F.L. Gervasio, O. Procacci, and V. Schettino. *Stacking and T-shape competition in aromatic-aromatic amino acid interactions.* J. Am. Chem. Soc. 124, 6133–6143 (2002).

[175] R. Tycko and Y. Ishii. *Constraints on Supramolecular Structure in Amyloid Fibrils from Two-Dimensional Solid-State NMR Spectroscopy with Uniform Isotopic Labeling.* J. Am. Chem. Soc. 125, 6606–6607 (2003).

[176] G. B. McGaughey, M Gagné, and A. K. Rappé. *π- stacking interactions.* J. Biol. Chem. 273, 15458–1546 (1998).

[177] B.P. Leifer. *Early Diagnosis of Alzheimer's disease: clinical and economic benefits.* J. Am. Geriatr. Soc. 51, 281–288 (2003).

[178] D.F. Williamson, F. Vinicor, and B.A. Bowman. *Primary prevention of type 2 diabetes mellitus by lifestyle intervention: implications for health policy.* Ann. Intern. Med. 140, 951–957 (2004).

[179] E. Gazit. *The role of prefibrillar assemblies in amyloid diseases.* Drugs Fut. 29, 613–619 (2004).

[180] M. Reches and E. Gazit. *Casting metal nanowires within discrete self-assembled peptide nanotubes.* Science 300, 625–627 (2003).

[181] J.T. Jarret and P.T.J. Lansbury. *Seeding "one-dimensional crystallization" of amyloid: a pathogenic mechanism in Alzheimer's disease and scrapie?* Cell 73, 1055–1058 (1993).

[182] E. Paci, J. Gsponer, X. Salvatella, and M. Vendruscolo. *Molecular dynamics studies of the process of amyloid aggregation of peptide fragments of transthyretin.* J. Mol. Biol. 340, 555–569 (2004).

[183] W. Hwang, S. Zhang, R.D. Kamm, and M. Karplus. *Kinetic control of dimer structure formation in amyloid fibrillogenesis.* Proc. Natl. Acad. Sci. USA 101, 12916–12921 (2004).

[184] C.M. Dobson and M. Karplus. *The fundamental of protein folding: bringing together theory and experiment.* Curr. Opin. Struct. Biol. 9, 92–101 (1999).

[185] E. Alm and D. Baker. *Matching theory and experiment in protein folding.* Curr. Opin. Struct. Biol. 9, 189–196 (1999).

[186] F.B. Sheinermann and C.L. Brooks III. *Calculation on folding of segmentvB1 of streptococcal protein G.* J. Mol. Biol. 278, 439–456 (1998).

[187] P. Ferrara, J. Apostolakis, and A. Caflisch. *Computer simulations of protein folding by targeted molecular dynamics.* Proteins 39, 252–260 (2000).

[188] J. Schlitter, M. Engels, P. Kruger, E. Jacoby, and A. Wollmer. *Targeted molecular dynamics simulation of conformational changes: Application to the T< − >R transition in insulin.* Mol. Simul. 10, 291–309 (1993).

[189] J. Diaz, B. Wroblowski, J. Schlitter, and Y. Engelborghs. *Calculation of pathways for the confomational transition between the GTP- and GDP-bound states of the Ha-ras-p21 protein: calculations with explicit solvent simulations and comparison with calculations in vacuum.* Proteins 28, 434–451 (1997).

[190] J. Ma and M. Karplus. *Molecular switch in signal transduction: reaction paths of the conformational changes in ras p21.* Proc. Natl. Acad. Sci USA 94, 11905–11910 (1997).

[191] B. L. de Groot, A. Amadei, D. M. F. van Aalten, and H.J.C. Berendsen. *Towards an Exhaustive sampling of the configurational space of the two forms of peptide hormone guanylin.* J. Biom. Str. Dyn. 13, 741–751 (1996).

[192] S. Akiyama, S. Takahashi, K. Ishimori, and I. Morishima. *Stepwise formation of α-helices during cytochrome c folding.* Nat. Struct. Biol. 7, 514–520 (2000).

[193] S. Akiyama, S. Takahashi, T. Kimura, K. Oishimori, I. Morishima, Y. Nishikawa, and T. Fujisawa. *Conformational landscape of cytochrome c folding studied by microsecond-resolved small-angle x-ray scattering.* Proc. Natl. Acad. Sci. USA 99, 1329–1334 (2002).

[194] D.J. Segel, D. Eliezer, V. Uversky, A.L. Fink, K.O. Hodgson, and S. Doniach. *Protein denaturation: a small-angle X-ray scattering study of the ensemble of unfolded states of cytochrome c.* Biochemistry 38, 15352–15359 (1999).

[195] M. Ohgushi and A. Wada. *'Molten-globule state': a compact form of globular proteins with mobile side-chains.* FEBS Lett. 164, 21–24 (1983).

[196] Y. Xu, L. Mayne, and S.W. Englander. *Evidence for an unfolding and refolding pathway in cytochrome c.* Nat. Struct. Biol. 5, 774–778 (1998).

[197] M.C. Shastry, S.D. Luck, and H. Roder. *A continuous-flow capillary mixing method to monitor reactions on the microsecond time scale.* Biophys. J. 74, 2714–2721 (1998).

[198] S.J. Hagen and W.A. Eaton. *Two-state expansion and collapse of a polypeptide.* J. Mol. Biol. 301, 1019–1027 (2000).

[199] L. Pollack, M. W. Tate, N. C. Darnton, J. B. Knight, S. M. Gruner, W. A. Eaton, and R. H. Austin. *Compactness of the denatured state of a fast-folding protein measured by submillisecond Small-Angle X-ray Scattering.* Proc. Natl. Acad. Sci. USA 96, 10115–10117 (1999).

[200] C. L. Brooks III. *Viewing protein folding from many perspectives.* Proc. Natl. Acad. Sci. USA 99, 1099–1100 (2002).

[201] Z. Guo, C. L. Brooks III, and E.M. Boczko. *Exploring the folding free energy surface of a three-helix bundle protein.* Proc. Natl. Acad. Sci. USA 94, 10161–10166 (1997).

[202] J.G. Lyubovitski, H.B. Gray, and J.R. Winkler. *Mapping the cytochrome c folding landscape.* J. Am. Chem. Soc. 124, 5481–5485 (2002).

[203] G. W. Bushnell, G. V. Louie, and G. D. Brayer. *High-resolution three-dimensional structure of horse heart cytochrome c.* J. Mol. Biol. 214, 585–595 (1990).

[204] W. F. van Gunsteren and H. J. C. Berendsen. *Gromos manual.* BIOMOS, Biomolecular Software, Laboratory of Physical Chemistry, University of Groningenm, The Netherlands, 1987.

[205] A. R. van Buuren, S. J. Marrink, and H. J. C. Berendsen. *A molecular dynamics study of decane/water interface.* J. Phys. Chem. 97, 9206–9212 (1993).

[206] J.P. Ryckaert and A. Bellemans. *Molecular dynamics of liquid n-butane near its boiling point.* Chem. Phys. Lett. 30, 123–125 (1975).

[207] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. Interaction models for water in relation to protein hydration. In B. Pullman, editor, *Intermolecular Forces*, pages 331–342. D. Reidel Publishing Company, Dordrecht, The Netherlands, 1981.

[208] W. Colon, G. A. Elove, L. P. Wakem, F. Sherman, and H. Roder. *Side chain packing of the N- and C-terminal helices plays a critical role in the kinetics of cytochrome c folding.* Biochemistry. 35, 5538–5549 (1996).

[209] J. L. Marmorino, M. Lehti, and G. J. Pielak. *Native tertiary structure in an A-state.* J. Mol. Biol. 275, 379–388 (1998).

[210] G.E. Schulz. *Domain motions in proteins.* Curr. Opin. Struct. Biol. 1, 883–888 (1991).

[211] M. Gerstein, A. M. Lesk, and C. Chothia. *Structural mechanisms for 0 movements in proteins.* Biochemistry 33(22) (1994).

[212] S. Hayward. *Structural principles governing domain motions in proteins.* Proteins: Struct. Funct. Genet. 36, 425–435 (1999).

[213] E. N. Baker, S. V. Rumball, and B. F. Anderson. *Transferrins - Insights into Structure and Function from Studies on Lactoferrin.* Trends Biochem. Sci. 12, 350–353 (1987).

[214] O. Millet and R.P. Hudson nad L.E. Kay. *The energetic cost of domain reorientation in maltose-binding protein as studied by NMR and fluorescence spectroscopy.* Proc. Natl. Acad. Sci. USA 100, 12700–12705 (2003).

[215] D.I. Liao, M. Karpusas, and S.J. Remington. *Crystal structure of an open conformation of citrate synthase from chicken heart at 2.8Å resolution.* Biochemistry 30, 6031–6036 (1991).

[216] D. Roccatano, A. E. Mark, and S. Hayward. *Investigation of the mechanism of domain closure in citrate synthase by molecular dynamics simulation.* J. Mol. Biol. 310, 1039–1053 (2001).

[217] G. Wiegand. *Citrate synthase, structure, control, and mechanism.* Annu.Rev. Biophys.Biophys.Chem. 15, 97–117 (1986).

[218] H. J. C. Berendsen, D. van der Spoel, and R. van Drunen. *GROMACS: A message-passing parallel molecular dynamics implementation.* Comp. Phys. Comm. 95, 43–56 (1995).

[219] S. Remington, G. Wiegand, and R. Huber. *Crystallographic refinement and atomic models of two different forms of citrate synthase at 2.7 and 1.7 Angstroms resolution.* J. Mol.Biol. 158, 111–152 (1982).

[220] S. Hayward and H. J. C. Berendsen. *Systematic analysis of domain motions in proteins from conformational change: new results on citrate synthase and T4 lysozyme.* Proteins: Struct. Funct. Genet. 30, 144–154 (1998).

[221] S. Hayward and R.A. Lee. *Improvements in the analysis of domain motions in proteins from conformational change: DynDom version 1.50.* J. Mol.Graph.Model. 21, 181–183 (2003).

[222] S. Hayward. *Identification of specific interactions that drive ligand-induced closure in five enzymes with classic domain movements.* J. Mol. Biol. in press (2004).

[223] R. Sayle and E.J. Milner-White. *Rasmol: Biomolecular graphics for all.* Trends in Biochem. Sci. 20, 374–375 (1995).

[224] P. J. Kraulis. *MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures.* J. Appl. Cryst. 24, 946–950 (1991).

[225] E. A. Merritt and D. J. Bacon. *Raster3D: Photorealistic Molecular Graphics.* Meth. Enzymol. 277, 505–524 (1997).

[226] R.A. Lee, M. Razaz, and S. Hayward. *The DynDom database of protein domain motions.* Bioinformatics 19, 1290–1291 (2003).

[227] A. Bairoch. *The ENZYME database in 2000.* Nucl. Acids Res. 28, 304–305 (2000).

[228] B. Boeckmann, A. Bairoch, R. Apweiler, M.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider. *The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003.* Nucl. Acids Res. 31, 365–370 (2003).

[229] J. D. Thompson, D.G. Higgins, and T.J. Gibson. *Clustal- W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position - Specific Gap Penalties and Weight Matrix Choice.* Nucl. Acids Res. 22, 4673–4680 (1994).

[230] N. Hulo, C.J.A. Sigrist, V. Le Saux, P.S. Langendijk-Genevaux, L. Bordoli, A. Gattiker, E. De Castro, P. Bucher, and A. Bairoch. *Recent improvements to the PROSITE database.* Nucl. Acids Res. 32, D134–D137 (2004).

[231] D. H. Anderson and H. W. Duckworth. *Mutation of Amino-Acids Thought to Polarize the Oxaloacetate Carbonyl in Citrate Synthase Severely Reduces but Does Not Abolish Activity of the Enzyme*. Biochem. Cell Biol.-Biochim.Biol. Cell. 67, 98–102 (1989).

[232] W.J. Man, Y. Li, C.D. Oconnor, and D.C. Wilton. *The Effect of Replacing the Conserved Active-Site Residues His-264, Asp-312 and Arg-314 on the Binding and Catalytic Properties of Escherichia-Coli Citrate Synthase*. Biochem. J. 300, 765–770 (1994).

Acknowledgments

I would like to thank all the people that have supported me during my PhD. My professor, Alfredo Di Nola, for his precious guidance during these three years. Danilo Roccatano for the work done together and because he always believed in me. Andrea Amadei for his always enthusiastic scientific support. Giorgio Colombo and Steven Hayward for our excellent collaborations. All the members of the group here in Rome, in particular Vincenza, Cecilia, Alex, Marco, Gianluca, Riccardo, Angela, Dagmar, Daniele, Massi, Martin and everyone whom I eventually forgot!
My father.
My mother.

List of Abbreviations

3D, Three Dimensional
C_α, Backbone Carbon Atom
CD, Circular Dichroism
CHC, Central Hydrophobic Core
cytc, Cytochrome c
ED, Essential Dynamics
EDS, Essential Dynamics Sampling
HB, Hydrogen Bond
HFIP, 1,1,1-3,3,3-hexafluoropropan-2-ol
IAPP, Islet Amyloid Polypeptide
LD, Langevin Dynamics
MC, Monte Carlo
MD, Molecular Dynamics
NMR, Nuclear Magnetic Resonance
NOE, Nuclear Overhauser Effect
PBC, Periodic Boundary Conditions
PCA, Principal Component Analysis
PDB, Protein Data Bank
PM, Perturbation Method
PME, Particle Mesh Ewald
PMF, Potential of Mean Force
PRD, Parallel Replica Dynamics
PrP, Prion Protein
PrP^C, Cellular Form of PrP
PrP^{Sc}, Scrapie Form of PrP
REMD, Replica Exchange Molecular Dynamics
RMSD, Root Mean Square Deviation
RMSF, Root Mean Square Fluctuation
R_g, Radius of Gyration
RG_RMSD, Rigid Body RMSD

SAS, Solvent Accessible Surface
SAXS, Small Angle X-ray Scattering
SPC, Simple Point Charge
TFE, 2,2,2-trifluoroethanol
TI, Thermodynamic Integration
TMD, Targeted Molecular Dynamics
UV, Ultraviolet