PER LARSSON

Computer simulations of protein folding and energetics in hydrogen bonding

Master’s degree project
Title (English)

Computer simulations of protein folding and energetics in hydrogen bonding

Title (Swedish)

Abstract

The lifetimes of the hydrogen bonds are a measure of the strength of the interaction, however, there is no obvious definition of the lifetime of a hydrogen bond. Here, different definitions used in the literature are compared. In addition, a novel way of calculating the strength of hydrogen bonds between a protein and the solvent is presented. The conclusion is that the method gives results that are within the experimental limits for hydrogen bond strength.

Keywords

Computer simulations, protein folding, hydrogen bond energetics, GROMACS

Supervisor

David van der Spoel
Uppsala universitet

Scientific reviewer

Johan Åqvist
Uppsala universitet

Project name

Sponsors

Language

English

Security

Secret until 2005-04

ISSN 1401-2138

Classification

Supplementary bibliographical information

Pages

43

Biology Education Centre

Biomedical Center

Box 592 S-75124 Uppsala

Tel +46 (0)18 4710000

Husargatan 3 Uppsala

Fax +46 (0)18 555217
Computer Simulations of Protein Folding and Energetics in Hydrogen Bonding

Per Larsson

October 5, 2004
Contents

Introduction and Purpose 1

1 Protein Folding Theory 2
  1.1 The Levinthal Paradox 2
  1.2 The New View of Protein Folding 3
  1.3 Trp-cage 4

2 Calculational Theory 7
  2.1 Computational Chemistry 7
    2.1.1 Force Fields 8
    2.1.2 Integration Algorithms 10
  2.2 Dynamics and Structure of Hydrogen Bonds 11
    2.2.1 Autocorrelation Functions 12
    2.2.2 Hydrogen Bond Lifetimes 12
    2.2.3 Derivation of the Eyring Equation 14

3 Methods 16
  3.1 Molecular Dynamics Simulations 16

4 Results and Discussion 18
  4.1 Trp-cage Simulations 18
    4.1.1 RMS Deviation from the NMR Structure 18
    4.1.2 Radius of Gyration 23
    4.1.3 Salt Bridge Distance 23
  4.2 Autocorrelation Times 26
  4.3 Hydrogen Bond Energies 26

5 Conclusions 29
  5.1 Hydrogen Bonding 29
  5.2 The Prospect of Miniproteins 30
  5.3 Future Outlook 31

Acknowledgements 31

References 32
CONTENTS

A  Autocorrelation Times  35
   A.1 Hydrogen Bond Lifetimes, $\tau_{HB}$  .................................. 35
   A.2 Relaxation Times, $\tau_{rlx}$  .............................................. 36
   A.3 Mean First Passage Times, $\tau_{mf}$  ...................................... 37

B  Hydrogen Bond Energies  38
   B.1 Hydrogen Bond Energies from $\tau_{HB}$  ............................... 38
Abstract

The mechanisms behind the ability of proteins to find their native structure, the so-called protein folding problem, remains one of the grand challenges in science. Despite much effort, the problem remains largely unsolved. Today, computers are used to simulate folding proteins using different models. These computer models together with the theory of molecular dynamics have allowed us to make progress, despite the challenges. One important contributor to the folding of proteins is hydrogen bonding, both internal bonds and bonds with the surrounding solvent. In this project, extensive computer simulations of Gly-X-Gly peptides and a small protein, the Trp-cage, in water are presented from which the hydrogen bonding is analysed. The lifetimes of the hydrogen bonds are a measure of the strength of the interaction, however, there is no obvious definition of the lifetime of a hydrogen bond. Here, different definitions used in the literature are compared. In addition, a novel way of calculating the strength of hydrogen bonds between a protein and the solvent is presented. The conclusion is that the method gives results that are within the experimental limits for hydrogen bond strength. The drawback is its tendency to overestimate the energies, due to the fact that the autocorrelation function also takes into account the diffusion of molecules necessary to put a hydrogen bond in place. Common protein folding determinants are also calculated for the Trp-cage.
Introduction and Purpose

For a protein to be biologically active, it must adopt a stable, unique structure within a very short time-span. Usually, the folding process takes place on the microsecond (for smaller proteins) to second (larger and more complex proteins) time scale. The question of how individual sequences achieve this is one of the most intriguing questions in structural biology. The advent and ever-increasing capacity of today’s computers has given birth to a new way of addressing this the so called protein folding problem. Many different mechanisms are thought to contribute to the folding of a protein, such as the hydrophobic effect and hydrogen bonding. Using computer simulations, it is possible to elucidate some of the factors that contribute to the folding process.

The purpose of this thesis work was twofold. It addresses the ability of the GROMACS [1] molecular dynamics package to simulate the Trp-cage miniprotein. To this end, common protein folding determinants such as RMSD and Radius of Gyration were used. In addition, the Trp-cage simulations were used together with simulations of small peptides to analyse the hydrogen bond dynamics between amino acid residues and solvent (water) using the theory of autocorrelation functions. It is shown that such correlations give different ways of calculating the energy of a hydrogen bond, and that these energy values differ but fall within reasonable limits. A novel analysis technique for energetics in hydrogen bonding is also presented and evaluated.
Chapter 1

Protein Folding Theory

There are two protein folding problems [4], one concerned with the kinetics of the actual folding process and one that deals with the prediction of the unique three-dimensional structure determined by the sequence of amino acids, as suggested by Anfinsen[3].

1.1 The Levinthal Paradox

The amino-acid sequence of a protein is encoded as the genetic information in a nucleotide sequence of DNA, which resides in the nucleus of a cell. DNA is transcribed into messenger RNA, which moves out of the nucleus and into the cytoplasm, and binds to the ribosome, where the proteins are synthesized. Hence, the tertiary structures of proteins seem to be dependent on various environmental factors within the cell. However, it was experimentally shown by Anfinsen that the three-dimensional structure of a protein is determined solely by the amino-acid sequence [3]. Earlier still, Levinthal had put forward his now very famous paradox about protein folding. The paradox can be stated as the observation that there is insufficient time to randomly search the entire conformational space available to a polypeptide chain, yet a protein folds [2]. In a simple form, given a protein with $N$ amino acids, the number of conformations that the protein can sample is of the order $10^N$, i.e. 10 degrees of freedom per residue. Considering an average rotational frequency around each bond [2] one can assume that a protein can sample of the order of $10^{14}$ structures per second. Hence, it would take this protein about $10^{26}$ seconds $\approx 10^{18}$ years to examine all the possible conformations. This is of course much longer than the present
age of the universe. Then why can proteins fold into their native conformations on the time scale of milliseconds to minutes? The obvious solution, and this was Levintinthal’s point, is that proteins have to fold through some directed process. The challenge then is not to resolve the paradox itself, but rather to address scientific questions such as the nature of the directed process.

The discovery made by Anfinsen gives rise to the thought that if the rules by which the sequence of amino acid residues influence the overall structure of the folded protein are known, it should be possible to predict such processes. However, this has not been accomplished to date, and it is remains an open question whether \textit{ab initio} prediction of protein tertiary structure is possible at all.

1.2 The New View of Protein Folding

The 'old view' of protein folding can be defined as describing folding akin to chemical reactions involving intermediates and transition states [5]. However, it is possible to reason that not many people thought in these simplistic terms. In a paper from 1976 [6], Levinthal and Honig \textit{et al.} wrote that 'We assume that proteins fold by following multiply defined branched pathways in which the first stage is the formation of local secondary structure governed by integrations that are near each other in the peptide chain. Subsequently, these structures, such as alpha-helices and antiparallel beta-strands, would interact, perhaps being modified in the process, to produce larger structural fragments which then undergo further assembly to yield the native conformation.' Of course, one might argue that the phrase 'branched pathways' is somewhat vague, especially given the lack of experimental knowledge at that time. However, is possible still to deduce from this the idea that even though there might not be a fixed sequence of events in protein folding, there is a general order of events that can be described in a phenomenological sense, in terms of the progressive accumulation of tertiary structure from secondary fragments. Ptitsyn [7] expressed a similar view a few years earlier, and Karplus and Weaver [8] introduced 'the diffusion collision model' which describes the coalescence of secondary structure units that are themselves unstable.
Recently, there has emerged a ‘new view’ of protein folding, due to both experimental and theoretical advances [4, 9]. The new view recognizes the importance of describing proteins in terms of statistical ensembles, rather than isolated molecules. To understand the concepts that have led to this change in perception about protein folding, it can be helpful to make comparisons with ideas developed to understand simple chemical reactions. These reactions are thought to take place on an energy surface from which reaction trajectories can be calculated. For proteins, which are larger and more complex with thousands of degrees of freedom, these multi-dimensional surfaces play an essential role. They can be described in terms of a folding funnel which embodies the idea that there is a large number of states available to unfolded proteins and far fewer to folded proteins. The width of the funnel is is related to the configurational entropy of the polypeptide chain, while the depth of the funnel depicts a free energy function that does not include the protein’s internal degrees of freedom. A funnel for protein folding implies that there is a decrease in energy and concomitant loss of entropy with increasing structure [9]. An essential element is thus the recognition that a bias toward the native state over the effective energy surface may govern the folding process. This has replaced the random search paradigm of Levinthal and suggests that there are many ways of reaching the native state in a reasonable time so that a specific pathway does not have to be postulated.

1.3 Trp-cage

Both experimentalists and theoreticians keep searching for faster and faster folders in order to bridge the gap between computer simulations and the actual time scale of protein folding. The protein used in this study was designed by Neidigh et al. by a series of truncations and mutations of a poorly folded 39-residue peptide [10]. Several constructs were made, resulting in a peptide that is >95 percent folded in water at physiological pH. The constructs optimize a novel fold, called the Trp-cage, and folding is cooperative and hydrophobically driven by the encapsulation of a Trp side chain in a sheath of Pro rings.

The 20-residue Trp-cage has the sequence NLYIQ WLKDG GPSSG RPPPS
and contains a short alpha-helix from residues 2 through 9, a $3_{10}$ helix from residues 11 through 14, and a C-terminal poly-proline II helix packing against the central tryptophan [22]. Thus, despite its relatively short amino acid sequence, the Trp-cage has multiple secondary structure elements and side-chain side-chain packing interactions.

The Trp cage is cooperatively folded. The fields of protein structure and folding are mature fields of endeavor [14], and one is led to ask two fundamental questions regarding the system reported by Neidigh et al. [10]. (i) Is the structure completely folded under accessible conditions, such as physiological pH and temperature? (ii) Is folding cooperative? This latter question is important because for most natural protein domains, all parts of the molecule undergo the folding transition in concert, as conditions, such as temperature or denaturant concentration, are varied. This type of cooperativity is regarded as a hallmark of native protein structures [12]. In such cases, only two 'states' of the protein - the folded and unfolded states - are present throughout the process. Both states are collections of conformations, also referred to as ensembles. The unfolded conformations are flexible and largely unrelated to one another, whereas the folded-state conformations are closely related. The work of Neidigh et al. [10] provides clear answers to these two questions.
Evidence for essentially complete folding into a specific tertiary structure near 0°C is provided by the following independent global measures of folding—that is, characteristic CD signals and 1H NMR chemical shifts—as a function of temperature. In both cases the measured parameters plateau at low temperature to values consistent with a highly folded state. Further support for complete folding in water comes from the observation that there is very little change in the CD spectrum of the final Trp-cage design when the sample contained TFE (2,2,2-trifluoroethanol), an agent that promotes the formation of α-helical and β-sheet structures. In contrast, intermediate designs showed substantial enhancements in folding upon addition of TFE. Although addition of TFE does not necessarily induce complete peptide folding [15], it seems plausible that failure to see a TFE effect indicates that the extent of folding is already quite high.

What is the evidence for cooperative folding? Sigmoidal dependence of spectroscopic measurements with changes in temperature or denaturant concentration is a classic signature of cooperativity. The authors show such temperature dependencies for both CD and NMR signals. They also use a novel method [16] to test for cooperativity: they compare the rates of change of characteristic 1H NMR chemical shifts through the thermal unfolding transition with the differences measured for these chemical shifts between the folded state and reference values for the unfolded state (folded - unfolded). A linear correlation between these two parameters, which is observed for the Trp cage, fits a cooperative unfolding transition between a single folded state and an unfolded ensemble. These comparisons support the hypothesis that the final Trp cage peptide undergoes a concerted unfolding transition.
Chapter 2

Calculational Theory

2.1 Computational Chemistry

Computer calculations using the laws of physics can be performed at many levels of accuracy. At the heart of much of today’s computational chemistry lies quantum mechanics and the Shrödinger equation. It is possible to solve the Schrödinger equation numerically, but while not very complicated in itself, exact solutions can only be found for very small and simple systems. In the case of more complicated systems, essentially larger than the hydrogen atom, it becomes necessary to look for either numerical approximations to quantum mechanics or another set of methods.

Numerically solving the Schrödinger equation for a system of arbitrary size requires an enormous amount of computer power, and so puts an upper limit to the size of systems that can be simulated. Theoretically (and hypothetically), at some stage in the development of computers, it will become possible to simulate any system. In practice, this obstacle will probably not be overcome any time soon, and we are left with smaller system.

As mentioned, it is necessary to look for other methods than pure quantum mechanics. The perhaps easiest way to go about this is to ignore the quantum nature of atoms completely, and fall back to classic Newtonian physics. The theoretical basis for this is the Born-Oppenheimer approximation, in which only atom nuclei are considered, and electrons are disregarded. To this end models have been developed whereby a molecule is represented as a collection of (interconnected) spheres. Most of the relevant motions in such a model, including
2.1 Computational Chemistry

bond stretching, bond rotations and torsional motions, are adequately described by classical mechanics.

These simplifications make it possible to study chemistry in silico at a much lower computational cost than would otherwise be the case. The size of the systems that can be simulated can be increased tremendously, and this makes it possible to apply computer calculations to systems like proteins. The concept of molecular dynamics was originally developed by Alder and Wainwright in the early 1950s as a technique to simulate a system of colliding particles [17]. A simple definition of molecular dynamics is thus that it simulates the motions of a system of $N$ spheres over time. Assuming that these spheres each have mass $m_i$, systematic application of Newton’s equations of motion will reproduce this time development. This calculation has the form

$$m_i \frac{d^2 r_i}{dt^2} = F_i$$

where $r_i(t)$ is the position of particle $i$ and $F_i$ the force acting on particle $i$. Thus, to obtain the dynamical behaviour of the system, it is necessary to solve this second order differential equation for every particle in the system. This solution is based upon the potential function $V(r)$, which is determined by the positions of the particles

$$F_i = \nabla_{r_i} V(r_1, ..., r_N).$$

By repeating this calculation many times a trajectory will be produced, containing the development of positions, velocities and forces on the atoms. If the potential function is selected carefully, this method gives a detailed representation of the evolution of the system in time.

2.1.1 Force Fields

The molecular representation introduced above was one which treated molecules as a set of interconnected spheres. The potential function $V(r)$, which governs the forces acting upon the system, was also introduced. Essentially, this potential function lies at the heart of molecular dynamics. It must be able to allow accurate calculations of molecular properties yet not consume too much
computational time. Several different force fields have been developed based on various types of experimental data.

Common for all of these is that the energy of a molecule is calculated as a sum of steric (bonds, bond angles and bond rotations) and non-bonded interactions close to each other in space (electrostatic interactions and van der Waals forces). The typical stretching behaviour of a bond is best described by the Morse potential [18]. While accurate, it is computationally less efficient than the simpler harmonic function, which is often used instead. Between atoms $i$ and $j$, the bond energy is then

$$V_{\text{bond}}(r_{ij}) = \frac{k_{ij}^b}{2} (r_{ij} - r_{ij}^0)^2,$$  \hspace{1cm} (2.3)

where $k_{ij}^b$ is a force constant that describes the rigidity of the bond and $r_{ij}^0$ is the bond length at equilibrium. Similarly, the angle between atoms $i$, $j$ and $k$ (where $i$ is bonded to $j$ and $j$ is bonded to $k$) can be described by

$$V_{\text{angle}}(\vartheta_{ijk}) = \frac{k_{ij}^\vartheta}{2} (\vartheta_{ijk} - \vartheta_{ijk}^0)^2$$  \hspace{1cm} (2.4)

where $\vartheta_{ijk}^0$ is the equilibrium value for the angle. The force constant $k_{ij}^\vartheta$ determines how hard it is to distort the angle. The last bonded interaction in most force fields is the dihedral angles. The dihedral angle is a rotation around around the central bond in a sequence of four atoms, usually modeled by a truncated Fourier series,

$$V_{\text{dihedral}}(\varphi_{ijkl}) = k_{jk}^\varphi [1 + \cos(n\varphi - \varphi_0)].$$  \hspace{1cm} (2.5)

where $\varphi_0$ represents the zero position of the angle.

All these interactions can be defined as bonded interactions in the sense that they are defined by the connectivity of the molecule. A significant property of these when making calculations is that their number only will increase linearly with the system size. The other type of interaction will be referred to as non-bonded interactions, and are dependent upon the distance between the atoms. It is common to consider these interactions having two components, van der Waals and electrostatic, describing a short range repulsion and a long
range attracting interaction. The repulsive component are modeled using the Lennard-Jones potential,

\[ V_{\text{Lennard-Jones}}(r_{ij}) = \frac{C_{12}^{ij}}{r_{ij}^{12}} - \frac{C_{6}^{ij}}{r_{ij}^{6}} \]  

(2.6)

where the parameters \( C_{12}^{ij} \) and \( C_{6}^{ij} \) depend on the system and govern the properties of the repulsion and attraction of the atoms. The potential between charged atoms is given by Coulomb’s law

\[ V_{\text{Coulomb}}(r_{ij}) = \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} \]  

(2.7)

where \( q_i \) and \( q_j \) are the charges on the atoms, \( \varepsilon_0 \) the permittivity of vacuum and \( \varepsilon_r \) the relative permittivity of the solvent.

### 2.1.2 Integration Algorithms

When solving equation 2.1, a numerical scheme has to be used. In all numerical methods, factors like round-off truncation errors has to be taken into account. To this end, a number of algorithms have been developed and the perhaps most widely used such algorithm was developed by Verlet and has the form:

\[ r_i(t - \Delta t) = r_i(t) - \Delta t \frac{d}{dt} r_i(t) - \frac{(\Delta t)^2}{2!} \frac{d^2}{dt^2} r_i(t) + \frac{(\Delta t)^3}{3!} \frac{d^3}{dt^3} r_i(t) + \ldots \]  

(2.8)

\[ r_i(t + \Delta t) = r_i(t) + \Delta t \frac{d}{dt} r_i(t) + \frac{(\Delta t)^2}{2!} \frac{d^2}{dt^2} r_i(t) + \frac{(\Delta t)^3}{3!} \frac{d^3}{dt^3} r_i(t) + \ldots \]  

(2.9)

which together with equation 2.1 lead to

\[ r_i(t + \Delta t) \approx -r_i(t - \Delta t) + 2r_i(t) + \frac{\Delta t^2}{m_i} F_i \]  

(2.10)

\[ v_i(t) \approx \frac{1}{2\Delta t} [r_i(t + \Delta t) - r_i(t - \Delta t)] \]  

(2.11)
A slightly better algorithm is a modified version of the Verlet algorithm known as the Leap-Frog algorithm. The equations are:

\[ r_i(t + \Delta t) \approx r_i(t) + \Delta t v_i(t + \frac{\Delta t}{2}) \]  
\[ v_i(t + \frac{\Delta t}{2}) \approx v_i(t - \frac{\Delta t}{2}) + \frac{\Delta t}{m_i} F_i \]

where the velocities are evaluated at half time steps as opposed to the original Verlet algorithm.

### 2.2 Dynamics and Structure of Hydrogen Bonds

Water is the single most important prerequisite for life [23]. It is questionable whether any macromolecule would be able to keep its structure and function without the presence of water. The small size of the water molecule allows it to act as a lubricant for conformational transitions in biomolecules [24, 25].

The rich variety of reactions involving water can be attributed to the geometry of the water molecule itself. A water molecule is triangular, not linear, and so there is an asymmetrical distribution of charge. The oxygen nucleus draws electrons away from the hydrogen nuclei, which leaves the region around those nuclei with a net positive charge. Thus, the water molecule is an electrically polar structure. The polarity and the hydrogen bonding capacity of the water molecule make it a highly interacting molecule. The hydrogen atoms of a water molecule can replace an amide hydrogen group as a hydrogen-bond donor, and the oxygen atom can replace a carbonyl oxygen as the acceptor. Interactions between water molecules is dominated by the electrostatic interactions of the dipoles. These interactions bring two water molecules so close together that a weak bond is formed between a hydrogen atom and an oxygen atom on the other molecule. A complex network of such hydrogen bonds can be formed as illustrated in figure 2.1.

At room temperature, a typical pair of these bonded molecules will separate and form new bonds with other neighbours in a time of roughly 4 ps according
2.2 Dynamics and Structure of Hydrogen Bonds

Figure 2.1: Simplified outline of hydrogen bonding pattern between water molecules.

Hydrogen bonds are generally weak bonds but when there are many, they maintain structure. DNA, for example, maintains its helical structure because of core hydrogen bonds. The breaking of hydrogen bonds between water molecules is what makes water a liquid (or vapor). The strength of the hydrogen bond in the linking of protein structures, particularly in a water environment, is of essential importance to predict several aspects of protein folding.

2.2.1 Autocorrelation Functions

Hydrogen bond lifetimes can be analysed by defining a binary function $h(t)$, which is 1 when a hydrogen bond is present and 0 otherwise. Then the autocorrelation function $C_h(t)$ of $h(t)$ can be computed and averaged over hydrogen bonds between the peptide sidechains and solvent according to

$$C_h(t) = \frac{\langle h(0)h(t) \rangle}{\langle h \rangle}$$  \hspace{1cm} (2.14)

2.2.2 Hydrogen Bond Lifetimes

Generally, the integral of $C_h(t)$ can be interpreted as a 'correlation time', but this interpretation depends on a physical model in which each event like hy-
Dynamics and Structure of Hydrogen Bonds

Hydrogen bonding is discrete with its own time constant. By working under this assumption, the lifetimes of the hydrogen bonds can be determined. The autocorrelations from equation 2.14 are so-called history independent correlations, because the value of the correlation function is independent of previous hydrogen bonds. However, several autocorrelation times can be calculated, and they all correspond to various lifetime definitions of hydrogen bonds. Here, three such definitions of lifetimes were used:

i) Hydrogen bond lifetimes, $\tau_{HB}$, defined as the integral of $C_h(t)$.

ii) Hydrogen bond relaxation times, $\tau_{rlx}$, defined as $C_h(t = \tau_{rlx}) = e^{-1}$

iii) Hydrogen bond mean first passage time, $\tau_{mfp}$, defined as the time for passage along the bond length from $t=0$ until the bond is broken.

With the exception of $\tau_{mfp}$, these times have been calculated using both average values over the entire simulation trajectories and the initial distribution of hydrogen bonds. This is a probability distribution that gives the probability of a hydrogen bond existing at time $t$, given that it existed at time $t=0$. From this distribution, it is possible to calculate the initial lifetime of a hydrogen bond by multiplying with time and integrating. Also, from this distribution, the uninterrupted relaxation time can be calculated from the relation

$$C_h(t = \tau_{rlx}) = e^{-1} = 1 - \int_0^{\tau_{rlx}} P(t)dt$$  \hspace{1cm} (2.15)

were $P(t)$ is the initial distribution of hydrogen bond lifetimes.

In this work, the default GROMACS definition of a hydrogen bond has been used. This is a geometric definition, and it states that for a hydrogen bond to be present the donor-hydrogen-acceptor angle is no greater than $30^\circ$ and the oxygen-oxygen cut-off distance at most 0.35 nm.
2.2 Dynamics and Structure of Hydrogen Bonds

2.2.3 Derivation of the Eyring Equation

To derive the Eyring equation, which is going to form the base for the energy calculations, and to show that it is valid, consider the unimolecular equilibrium

\[ \text{A} \xleftrightarrow{k_1, k_{-1}} \text{B} \]  

(2.16)

To simplify matters slightly, focus only on one of the reactions. This does not mean any loss of generality, since everything can still be applied to the reverse reaction as well. Thus, the reaction becomes

\[ \text{A} \xrightarrow{k_1} \text{B} \]  

(2.17)

At any given time, only a subset of the molecules of A have sufficient energy to undergo conversion to B. Assuming that these reactive molecules, which are populating the transition state, are in equilibrium with the unreactive molecules, the relative populations can be related by an equilibrium constant \( K^\dagger \):

\[ \text{[A]}^\dagger = \frac{k_1}{k_{-1}} \text{[A]} = K^\dagger \text{[A]} \]  

(2.18)

(2.19)

Now, under the assumption that only reactive molecules, \( \text{A}^\dagger \), undergo transformation to B, it is possible to equate the bulk and microscopic rates:

\[ k_1 \text{[A]} = k^\dagger \text{[A]}^\dagger \]  

(2.20)

and therefore to represent the bulk rate constant as:

\[ k_1 = k^\dagger \frac{\text{[A]}}{\text{[A]}^\dagger} = k^\dagger K^\dagger \]  

(2.21)

From the theory of statistical mechanics, equation 2.21 can now be re-written as

\[ k_1 = \frac{k_B T}{h} K^\dagger \]  

(2.22)
Now, using the relation

$$\Delta G^\dagger = -RT \ln K^\dagger$$ \hspace{2cm} (2.23)

it is possible to arrive at the Eyring equation

$$k_1 = \frac{k_B T}{h} \exp \frac{-\Delta G^\dagger}{RT},$$ \hspace{2cm} (2.24)

Luzar and Chandler have, from the decay of the hydrogen bond correlation function, observed that hydrogen bond kinetics is not governed by a single process [27]. However, by considering a combination of all the processes involved as a single process and integrating the correlation functions as above, it is possible to get the total correlation time $\tau_{HB}$ corresponding to the hydrogen bond lifetime. This gives a measure of the time constant related to making or breaking a hydrogen bond. Assuming that the association of a solvent molecule to a peptide can be described as an Eyring process, it is possible to directly associate a Gibbs energy of activation $\Delta G^\dagger_{cor}$ with the time constant:

$$\tau_{HB} = \frac{h}{k_B T} \exp \frac{\Delta G^\dagger_{cor}}{k_B T},$$ \hspace{2cm} (2.25)
Chapter 3

Methods

3.1 Molecular Dynamics Simulations

There are 20 structures in the NMR ensemble of Trp-cage. In order to get a larger statistical foundation, four of these, no 1, 3, 9 and 15, were selected as the starting point for MD simulations. The proteins were solvated in cubic boxes of two different sizes, containing approximately 2770 and 16015 single point charge water molecules respectively. This resulted in a total number of eight different trajectories. The systems were neutralized by addition of Na$^+$ and Cl$^-$ ions and then subjected to 1000 steps of steepest descent minimization. The temperature was controlled for the proteins, solvent and ions separately, using weak coupling to a bath of constant temperature [29], with a coupling constant of $\tau_T = 0.1\text{ps}$. All simulations were done using periodic boundary conditions. Following this, all systems were equilibrated for 200 ps to allow structural relaxation of the solvent molecules. All simulations were performed at 280 K. The resulting eight systems, four small and four big, were then simulated for 50 ns and 25 ns, respectively, using a time step of 2 fs. A group based twin-range cut-off of 1.0 and 1.4 nm for van der Waals interactions was used. The final simulations were carried out at the national supercomputer center in Linköping. The Gromos-96 43a1 forcefield was used [30].

Peptides Ace-Gly-X-Gly-NH$_2$ peptide conformations were generated using the Pymol program for initial building and then by manually editing the resulting PDB-files. Ace corresponds to a N-terminal capping acetyl fragment, and the residue X to each of the residues Asn, Tyr, Gln, Lys, Asp, Ser and Arg.
For each of the constructs, editing was done in such a way that the C-terminal sequence produced by Pymol, two methane fragments, was changed into a NH$_2$ terminus by removing appropriate atoms. The neutral termini was used to avoid interactions with polar side chains. While this protonation state not, in all cases, correspond to the true protonation state under physiological conditions, it does not introduce any serious artefacts [31]. All peptides were solvated in a cubic box with 1400 simple point charge [29] water molecules. Subsequently, the peptides were subjected to 100 steps of steepest descent energy minimization to remove any bad van der Waals contacts between the peptides and the solvent. Simulations were then performed at 270, 280 and 290 K using the SHAKE algorithm and a time-step of 4 fs. The simulations were 2 500 000 steps long or 10 ns each. As with the Trp-cage simulations, the Gromos-96 43a1 force field was used for the peptide simulations. The MD simulations took approximately 70 hours each using a single node in the Beowulf class cluster. Equilibration of the simulations was monitored by plotting the energy and density (data not shown).
Chapter 4

Results and Discussion

4.1 Trp-cage Simulations

When analysing long molecular dynamics trajectories, it is helpful to look at the relative values of several parameters that give an indication of the quality of the simulation and the stability of the protein. Some of these values, such as Radius of Gyration, Root Mean Square Deviation and Solvent Accessible Hydrophobic Surface Area can be plotted as a function of time, and, in a stable protein, are expected to fluctuate around an equilibrium value. It is also interesting to look at the dynamical processes taking place during the simulation, and, thus, distance matrices and a salt bridge involving Asp9 and Arg16 has been studied.

4.1.1 RMS Deviation from the NMR Structure

There are several simulations published already on the Trp-cage: one by Simmerling et al. [19] who ran a few 20- to 50-ns MD simulations with a modified AMBER99 forcefield and found a structure very close to the native one from many low potential energy structures; and one by Snow et al. [20], who estimated the folding rate by using thousands of short (nanoseconds) kinetics runs with the united-atom OPLS force field; another by Pitera and Swope [21] who used the replica exchange method and found a <0.1 nm RMSD\textsubscript{\textalpha} deviation structure from the simulated ensemble using the AMBER94 force field. All the studies used a continuum solvent model, generalised Born, to save computational cost. This study, as well as the study by Zhou [22], uses explicit solvent to model water.
4.1 Trp-cage Simulations

The root mean square deviation (RMSD) is a measure of how much a conformation deviates from an initial structure, which has to be known beforehand. This information usually comes from a crystal or NMR structure, as is the case here. The RMSD$_{C\alpha}$ deviation as well as the overall protein RMSD$_{prot}$ was calculated according to equation 4.1 and is presented in figures 4.1 and 4.2.

$$RMSD(t) = \left[ \frac{1}{N_S} \sum (r_i(t) - r_i(0))^2 \right]^{\frac{1}{2}}$$ (4.1)

Here $r_i(t)$ is the position of atom i at time t after superposition, $r_i(0)$ is the position of atom i in the reference structure and $N_S$ is the number of atoms in the system.

While the values obtained in this simulation are not as good as the ones calculated by previously with other force fields [19, 21, 22] they are still fairly good. A small and stable RMSD value (typically < 0.2nm) for protein backbone atoms or $C\alpha$-atoms is a useful quality control for protein simulations. It is not possible here to distinguish, in RMSD-terms, any differences between the small and large simulations. They all fluctuate some, and are usually not below 0.2 nm. However, they still lie in the range 0.25 - 0.3 nm over most of the trajectory. So it is possible to deduce from the RMSD that the simulations are 'all right'.

The discrepancy between this study and previous ones might arise due to several factors. The reason for the low RMSD obtained by Pitera and Swope [21] is probably that the refinement of the NMR structure by Neidigh et al. [10] was done using the same force field. In an NMR structure, the terminal residues are usually poorly defined, and might contribute to a slightly higher RMSD-value. However, from figures 4.1 and 4.2, there is no significant decrease in RMSD when the terminal residues have been omitted. In order to get the most out of the RMSD calculations, an average structure for each nanosecond in the trajectory was also calculated and RMSD based on that structure was also analysed. The results of this can be found in figure 4.3.
4.1 Trp-cage Simulations

Figure 4.1: Root mean square deviation (RMSD) calculated on the four small simulated systems. RMSD$_{prot}$ RMSD$_{C\alpha}$, as well as the RMSD for the protein without the terminal Asp1 ans Ser20 residues (RMSD$_{core-prot}$) are shown. Up left NMR1, up right NMR3, down left NMR9 and down right NMR15. Note the small improvement of the core-RMSD compared with overall protein RMSD. The RMSD$_{C\alpha}$ is lower still, indicating that the crystal structure backbone conformation is reproduced in the simulation.
4.1 Trp-cage Simulations

Figure 4.2: Root mean square deviation (RMSD) calculated on the four large simulated systems. The same RMSD-values as for the small simulations are shown.
4.1 Trp-cage Simulations

Figure 4.3: Root mean square deviation (RMSD) calculated on all eight simulated Trp-cage systems. For each nanosecond, an average structure has been calculated and RMSD-calculations was done using this structure.
4.1 Trp-cage Simulations

4.1.2 Radius of Gyration

The compactness of a structure can be understood quantitatively by monitoring the radius of gyration. So the time evolution of the radius of gyration can be regarded as a good measure of the dynamics of collapse of a protein. The radius of gyration, $R_g$, was calculated according to equation 4.2

$$R_g = \left( \frac{\sum_i r_i^2 m_i}{\sum_i m_i} \right)^{\frac{1}{2}} \quad (4.2)$$

where $m_i$ is the mass of atom $i$ and $r_i$ is the position of atom $i$ with respect to the center of mass of the protein. In the model protein studied here, the change in the radius of gyration with time is plotted in figures 4.5. The corresponding calculation for the large simulations did not reveal any significant differences, indicating that the radius of gyration is largely unaffected by the number of solvent molecules that surrounds the protein. Had the radius of gyration increased substantially at any point, that would have been something to worry about.

4.1.3 Salt Bridge Distance

The distance between residues Asp9 and Arg16 has been calculated. The reason for this is that the previous simulations mentioned earlier has shown the existence of a salt bridge between these residues. Furthermore, when designing the Trp-cage miniprotein, Neidigh et al. [10] explicitly designed it to have a salt bridge. It is unclear what role salt bridges actually have in protein folding, and there is much discussion. The analysis shows that the salt bridge comes back and forth into existence over the whole trajectory, with some individual variation in different simulations.
4.1 Trp-cage Simulations

Figure 4.4: Radius of gyration for the four small simulations.
4.1 Trp-cage Simulations

Figure 4.5: Distance between the sidechains of Asp9 and Arg16. The salt bridge is considered to exist when the distance falls below 0.4 nm.
4.2 Autocorrelation Times

As stated above, there exist a number of definitions of hydrogen bond lifetimes, a number of hydrogen bond autocorrelation functions and a number of ways in which such functions can be used. Here, simulations on the Trp-cage miniprotein and small peptides have been performed to calculate a number of different τ-values, namely the autocorrelation lifetime (τ_{HB}), the relaxation time (τ_{rlx}) and the mean first passage time (τ_{mfp}). Furthermore, these times have been calculated for different situations: over entire simulation trajectories (τ_{rlx}, τ_{HB}) and for uninterrupted hydrogen bonds (τ_{rlx}, τ_{HB} and τ_{mfp}). There seems to be no correlation between these different values of τ. The results indicate, perhaps not surprisingly, that calculations of hydrogen bond lifetimes involving correlation functions of any kind, are sensitive to the type of correlation function used as well as to the way in which the time-scales of the system are defined. All different correlation times, however, point towards the temporal evolution of hydrogen bonds taking place on a picosecond timescale, in good agreement with other results [11, 26]. Theoretically, there exist a true distribution of hydrogen bond lifetimes, but sampling this distribution might not be straightforward in all cases. All correlation times can be found in Appendix A.

4.3 Hydrogen Bond Energies

One of the main objectives of the work presented here was to investigate whether energetics in hydrogen bonding between a protein and solvent can be described by the Eyring equation, 2.25. If this should turn out to be the case, i.e. the hydrogen bonding strength from experiments and other theoretical methods should be reproduced, the method presented here represents a novel simulation analysis technique for hydrogen bonding energetics. To this end, the different correlation times are used, in addition, to calculate (Arrhenius) activation energies to see how well the Eyring equation performs.

The Arrhenius equation is well known and relates the activation energy of a process with its rate constant. It has the form \( \tau_{HB} = \tau_0 \exp(-E_A/kT) \) and it can be used to calculate (among other things) activation energies, E_A. In this study, the different correlation times all behave according to the Arrhenius
4.3 Hydrogen Bond Energies

Table 4.1: Hydrogen bond energies from the different correlation times and the Eyring equation. Number in parentheses indicate residues in Trp-cage and the energies represent average values for all eight simulated systems. The columns correspond to (from left to right) the energy from the Eyring equation, the Arrhenius activation energy from the trajectory averages ($\tau_{HB}$) and from the initial hydrogen bond probability distribution ($\tau_{HB}$ and $\tau_{mfp}$). All units in kJ/mol.

<table>
<thead>
<tr>
<th>Residue</th>
<th>$\Delta G_{cor}$</th>
<th>$E_A (\tau_{HB})$</th>
<th>$E_A (\tau_{HB})$</th>
<th>$E_A (\tau_{mfp})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asn-1</td>
<td>7.1</td>
<td>16.2</td>
<td>3.48</td>
<td>7.29</td>
</tr>
<tr>
<td>Asn</td>
<td>7.74</td>
<td>13.9</td>
<td>5.42</td>
<td>0.95</td>
</tr>
<tr>
<td>Tyr-3</td>
<td>10.52</td>
<td>13.6</td>
<td>2.64</td>
<td>10.47</td>
</tr>
<tr>
<td>Gln-5</td>
<td>5.96</td>
<td>6.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gln</td>
<td>7.64</td>
<td>13.2</td>
<td>2.18</td>
<td>3.62</td>
</tr>
<tr>
<td>Lys-8</td>
<td>8.69</td>
<td>13.0</td>
<td>3.82</td>
<td>6.64</td>
</tr>
<tr>
<td>Lys</td>
<td>7.29</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp-9</td>
<td>9.93</td>
<td>12.2</td>
<td>4.78</td>
<td>5.94</td>
</tr>
<tr>
<td>Ser-13</td>
<td>6.44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser-14</td>
<td>6.31</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser-20</td>
<td>5.93</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>9.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg-16</td>
<td>6.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg</td>
<td>7.56</td>
<td>15.5</td>
<td>3.48</td>
<td>10.02</td>
</tr>
</tbody>
</table>

Thus, they serve as a convenient way to see if the Eyring equation gives reasonable values. This is assuming that the same energy is calculated in both cases. To distinguish whether the calculated Arrhenius activation energy is a total energy difference, enthalpy or free energy, two situations have to be considered. First, the time constants are ensemble averages and second, the Gibbs free energy is defined at constant pressure of the simulation system which is also the case here. Subsequently, the activation energy is a Gibbs energy rather than anything else. The Arrhenius activation energies have only been calculated for the Ace-Gly-X-Gly-NH$_2$ peptides, because simulations of the Trp-cage at different temperatures, a necessity for an Arrhenius analysis, was beyond the timeframe of this work. A comparison of these different energies can be found in Table 4.1. Note that it would not be meaningful to use the values of $\tau_{rlx}$ in this way, since they measure a different process.

By comparing the calculated hydrogen bond energies, it is evident that differences exist. The Arrhenius activation energies from the trajectory averages
4.3 Hydrogen Bond Energies

of $\tau_{HB}$ are highest in all cases, and the Eyring energies fall in between these energies and the others (uninterrupted $\tau_{HB}$ and $\tau_{mfp}$). It is worth noting that the estimates of the mean first passage times (and consequently the corresponding energies) are less accurate than the rest of the calculations, because of computational and modelling shortcomings. Thus, these values might not be very accurate.

More importantly, perhaps, is that the Eyring equation with autocorrelations also takes into account some of the diffusion that is necessary for formation of a hydrogen bond. This process is not involved in the actual association of a water molecule with an amino acid residue, but one can imagine that the time it takes for a molecule to diffuse makes the autocorrelations longer. This means that the method would overestimate the energy strength of a hydrogen bond. As there does not seem to exist any correlations between the different $\tau$-values, it is perhaps too much to ask that the corresponding energy values should be the same.

In terms of overall quality, all methods perform reasonably well. The correct strength of hydrogen bonds for proteins in solution is known to vary greatly from $\approx 20$-$25$ kJ/mol for isolated bonds to $\approx 2$-$6$ kJ/mol for proteins in solution [37].

An couple of interesting features in the Trp-cage are worth noting. The hydrogen bond lifetimes $\tau_{HB}$ (Tables A.1 and A.2) are longer than the corresponding times for the peptides (Table A.3). A possible explanation for this is that the Trp-cage, which is bigger and has more structure, does not change conformation as often as the smaller peptide systems, and thus that hydrogen bonds between Trp-cage and specific solvent molecules exist longer. A consequence of the longer lifetimes is that the $\Delta G_{cor}$-values also are higher (Table B.1 and B.2), when compared with the values in Table B.3. It remains to be seen if this holds for $\Delta H_{cor}$ and $T\Delta S_{cor}$ as well. Another thing that can be attributed to the more rigid structure of the Trp-cage is that the relaxation times are higher, probably corresponding to slower overall dynamics or the presence of more tightly bound water molecules due to stronger bonds. There might not exist any relationship between the size of a protein and the relaxation of hydrogen bonds (or overall dynamics), but for the Trp-cage and the smaller peptides this seem to be the case.
Chapter 5

Conclusions

5.1 Hydrogen Bonding

There exist a number of definitions of hydrogen bond lifetimes, a number of hydrogen bond autocorrelation functions and a number of ways in which such functions can be used. We have used molecular dynamics simulations on small peptides to calculate a number of different $\tau$-values, namely the autocorrelation lifetime ($\tau_{HB}$), the relaxation time ($\tau_{rlx}$) and the mean first passage time ($\tau_{mfp}$). Furthermore, these times have been calculated for different situations: over entire simulation trajectories ($\tau_{rlx}, \tau_{HB}$) and for uninterrupted hydrogen bonds ($\tau_{rlx}, \tau_{HB}$ and $\tau_{mfp}$). We find no correlations between these different values of $\tau$. Our results indicate that, perhaps not surprisingly, calculations of hydrogen bond lifetimes involving correlation functions of any kind, are sensitive to the type of correlation function used as well as to the way in which the time-scales of the system are defined. Theoretically, there exist a true distribution of hydrogen bond lifetimes, but sampling this distribution might not be straightforward in all cases. However, the correlation times and energies fall within reasonable limits. The Arrhenius activation energies are lower for the uninterrupted hydrogen bonds than for the trajectory average. This is probably due to fast motions initially in the simulations, before the system has relaxed completely. The above calculations show that the relaxation of hydrogen bonds is faster initially in the simulations when compared to the average relaxation time over the entire simulation.

When it comes to the energetics of the hydrogen bonds, using the Eyring
5.2 The Prospect of Miniproteins

The development of the Trp-cage is an important achievement for many reasons. We still have a great deal to learn about why specific amino acid sequences fold as they do, from both thermodynamic and kinetic perspectives. Small folding motifs offer excellent systems for detailed analysis because the number of factors that influence folded state stability is reduced, which benefits both experimental and computational studies of folding. In addition, a short sequence facilitates the use of chemical synthesis to generate analogs using not only the 20 standard amino acids, but also unnatural side chains and backbone elements. Such molecules would allow one to explore the origins of conformational stability in ways that are impossible with the proteinogenic amino acids [32]. For example, replacing a methyl side chain (alanine) with ethyl can probe the contribution of incremental nonpolar surface burial to conformational stability, and replacing a backbone amide linkage with an ester can probe the contribution of specific hydrogen bonds. In addition to their use for fundamental structural studies, small, robust tertiary motifs provide ideal scaffolds upon which interesting functions may be built [33].

The Trp cage represents a new benchmark in protein design [34, 35]. It is hard to see how a well-defined tertiary fold could comprise fewer than 20 residues. The exciting challenge offered by the Trp cage is development of alternative folding patterns with comparable length and then endowing them with interesting functions.
5.3 Future Outlook

The subject of protein folding is still one of the most challenging areas in biophysical chemistry. Given the complexity of protein structures, diverse folding processes should be expected, including the possible role of certain folding-assisting domains within large proteins [36]. Furthermore, the complexity of in vivo folding processes is, as yet, another challenge. Therefore, full understanding of protein folding mechanisms is indeed a daunting task. Despite this, recent developments give rise to the belief that an eventual solution may lie ahead.

From a theoretical perspective, an immediate objective is to accurately replicate the complete folding process of small fast-folding proteins on computers, including atomic details. Such simulations results would itself be a strong testament to the accuracy of the method and parameters. The diversity of protein structures and the complexity of the in vivo folding process can, in principle, be dealt with by a combination of experiments and further simulations. We may also be able to answer questions such as whether or not a particular part of a protein is designed to assist the folding of the rest of the protein, as found in "intramolecular chaperones", and how this assistance occurs. But without the basic understanding of the folding process of single-domain small, fast-folding proteins, understanding of more complex folding processes will be more difficult to achieve. In terms of the hydrogen bonding energy, it would be interesting to repeat these simulations at different temperatures. This would give a way of calculating not only the $\Delta G$ but also the values of $\Delta H$, and, subsequently $T\Delta S$. 
Acknowledgements

First of all I would like to thank my supervisor David van der Spoel for allowing me to do this project. I really enjoyed it! Thank you for always answering my questions and for helping me out with computer related stuff. And for the new computer! My examiner, Johan Åqvist, for reading my report, especially over the week-end.

To rest of the people in the corridor for enjoyable coffeebreaks. To Alexandra, for guidance and interesting discussions about absolutely anything and nothing. To my student opponents, Karin von Post and Henrik Hallingbäck, for reading my report and for putting up with its many flaws. And to Tomas Tengby, Nina Glans and Täppas Fogelberg for not giving up on people.

Finally, to Jenny, for all your support. Kebnekaise nästa!
Bibliography

Appendix A

Autocorrelation Times

A.1 Hydrogen Bond Lifetimes, $\tau_{HB}$

Table A.1: Hydrogen bond lifetimes for the four small simulations (NMR ensemble structures number 1, 3, 9 and 15) as computed by integrating the GROMACS hydrogen bond autocorrelation functions. All units in ps.

<table>
<thead>
<tr>
<th>Residue</th>
<th>$\tau_{NMR1}(ps)$</th>
<th>$\tau_{NMR2}(ps)$</th>
<th>$\tau_{NMR3}(ps)$</th>
<th>$\tau_{NMR4}(ps)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asn-1</td>
<td>13.75</td>
<td>13.38</td>
<td>24.64</td>
<td>12.98</td>
</tr>
<tr>
<td>Tyr-3</td>
<td>27.98</td>
<td>22.92</td>
<td>28.75</td>
<td>26.96</td>
</tr>
<tr>
<td>Gln-5</td>
<td>9.02</td>
<td>13.62</td>
<td>10.27</td>
<td>10.97</td>
</tr>
<tr>
<td>Lys-8</td>
<td>11.62</td>
<td>12.68</td>
<td>11.03</td>
<td>11.42</td>
</tr>
<tr>
<td>Asp-9</td>
<td>18.89</td>
<td>23.14</td>
<td>18.58</td>
<td>17.67</td>
</tr>
<tr>
<td>Ser-13</td>
<td>11.63</td>
<td>15.87</td>
<td>16.34</td>
<td>14.62</td>
</tr>
<tr>
<td>Ser-14</td>
<td>20.51</td>
<td>13.24</td>
<td>10.12</td>
<td>11.36</td>
</tr>
<tr>
<td>Arg-16</td>
<td>11.32</td>
<td>7.03</td>
<td>12.07</td>
<td>17.41</td>
</tr>
<tr>
<td>Ser-20</td>
<td>9.65</td>
<td>17.52</td>
<td>9.30</td>
<td>13.61</td>
</tr>
</tbody>
</table>
### A.2 Relaxation Times, $\tau_{rlx}$

Table A.2: Hydrogen bond life times for the four large simulations (NMR ensemble structures number 1, 3, 9 and 15) as computed by integrating the GROMACS hydrogen bond autocorrelation functions. Reported values represent averages of four different calculations on subsets of the simulation. All units in ps.

<table>
<thead>
<tr>
<th>Residue</th>
<th>$\tau_{NMR1}(ps)$</th>
<th>$\tau_{NMR2}(ps)$</th>
<th>$\tau_{NMR3}(ps)$</th>
<th>$\tau_{NMR4}(ps)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asn-1</td>
<td>17.59</td>
<td>28.56</td>
<td>42.39</td>
<td>14.77</td>
</tr>
<tr>
<td>Tyr-3</td>
<td>22.76</td>
<td>26.59</td>
<td>28.31</td>
<td>27.57</td>
</tr>
<tr>
<td>Gln-5</td>
<td>11.67</td>
<td>10.37</td>
<td>11.78</td>
<td>12.41</td>
</tr>
<tr>
<td>Lys-8</td>
<td>11.94</td>
<td>12.14</td>
<td>11.75</td>
<td>12.80</td>
</tr>
<tr>
<td>Asp-9</td>
<td>34.42</td>
<td>20.97</td>
<td>19.32</td>
<td>19.65</td>
</tr>
<tr>
<td>Ser-13</td>
<td>12.97</td>
<td>16.08</td>
<td>15.47</td>
<td>11.29</td>
</tr>
<tr>
<td>Ser-14</td>
<td>15.70</td>
<td>12.08</td>
<td>12.63</td>
<td>12.64</td>
</tr>
<tr>
<td>Arg-16</td>
<td>15.25</td>
<td>11.84</td>
<td>13.00</td>
<td>14.25</td>
</tr>
<tr>
<td>Ser-20</td>
<td>9.57</td>
<td>9.64</td>
<td>10.15</td>
<td>10.73</td>
</tr>
</tbody>
</table>

Table A.3: Average hydrogen bond lifetimes $\tau_{HB}$ for Ace-Gly-X-Hly-NH$_2$ based on integration of $C_h(t)$ and corresponding activation energies after curve fit to Arrhenius behaviour $\tau_{HB} = \tau_0 \exp(-E_A/kT)$.

<table>
<thead>
<tr>
<th>Residue</th>
<th>$\tau_{HB}^{270}$</th>
<th>$\tau_{HB}^{280}$</th>
<th>$\tau_{HB}^{290}$</th>
<th>$E_A^{270}$</th>
<th>$E_A^{280}$</th>
<th>$E_A^{290}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg</td>
<td>5.81</td>
<td>4.40</td>
<td>3.61</td>
<td>15.5</td>
<td>1.78</td>
<td>1.60</td>
</tr>
<tr>
<td>Asn</td>
<td>6.79</td>
<td>4.77</td>
<td>4.14</td>
<td>16.2</td>
<td>2.14</td>
<td>1.90</td>
</tr>
<tr>
<td>Asp</td>
<td>14.55</td>
<td>12.22</td>
<td>9.74</td>
<td>13.0</td>
<td>3.28</td>
<td>2.96</td>
</tr>
<tr>
<td>Gln</td>
<td>6.10</td>
<td>4.57</td>
<td>4.02</td>
<td>13.6</td>
<td>1.92</td>
<td>1.86</td>
</tr>
<tr>
<td>Glu</td>
<td>14.73</td>
<td>12.81</td>
<td>9.30</td>
<td>14.9</td>
<td>3.36</td>
<td>3.03</td>
</tr>
<tr>
<td>Lys</td>
<td>8.80</td>
<td>7.16</td>
<td>5.87</td>
<td>13.2</td>
<td>1.98</td>
<td>1.98</td>
</tr>
<tr>
<td>Ser</td>
<td>10.80</td>
<td>8.86</td>
<td>7.43</td>
<td>12.2</td>
<td>3.21</td>
<td>3.03</td>
</tr>
<tr>
<td>Tyr</td>
<td>18.95</td>
<td>15.72</td>
<td>12.34</td>
<td>13.9</td>
<td>3.63</td>
<td>3.46</td>
</tr>
<tr>
<td>Trp</td>
<td>3.38</td>
<td>2.88</td>
<td>2.47</td>
<td>10.2</td>
<td>1.36</td>
<td>1.33</td>
</tr>
<tr>
<td>Thr</td>
<td>13.04</td>
<td>10.59</td>
<td>8.66</td>
<td>13.3</td>
<td>3.48</td>
<td>3.07</td>
</tr>
<tr>
<td>His</td>
<td>21.91</td>
<td>15.04</td>
<td>12.50</td>
<td>18.3</td>
<td>5.04</td>
<td>4.64</td>
</tr>
</tbody>
</table>

A.2 Relaxation Times, $\tau_{rlx}$
A.3 Mean First Passage Times, $\tau_{mfp}$

Table A.4: Relaxation times $\tau_R$

<table>
<thead>
<tr>
<th>Residue</th>
<th>$\tau_{270}^{rlx}$</th>
<th>$\tau_{280}^{rlx}$</th>
<th>$\tau_{290}^{rlx}$</th>
<th>$\tau_{270}^{rlx}$</th>
<th>$\tau_{280}^{rlx}$</th>
<th>$\tau_{290}^{rlx}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg</td>
<td>2.49</td>
<td>1.89</td>
<td>1.42</td>
<td>1.77</td>
<td>1.7</td>
<td>1.42</td>
</tr>
<tr>
<td>Asn</td>
<td>3.53</td>
<td>2.57</td>
<td>2.21</td>
<td>1.95</td>
<td>1.88</td>
<td>1.75</td>
</tr>
<tr>
<td>Asp</td>
<td>9.91</td>
<td>8.02</td>
<td>6.43</td>
<td>3.49</td>
<td>3.28</td>
<td>3.05</td>
</tr>
<tr>
<td>Gln</td>
<td>3.09</td>
<td>2.37</td>
<td>2.06</td>
<td>1.95</td>
<td>1.83</td>
<td>1.79</td>
</tr>
<tr>
<td>Glu</td>
<td>10.07</td>
<td>7.89</td>
<td>6.43</td>
<td>3.5</td>
<td>3.27</td>
<td>3.04</td>
</tr>
<tr>
<td>Lys</td>
<td>4.58</td>
<td>3.82</td>
<td>3.15</td>
<td>2.16</td>
<td>2.09</td>
<td>2.01</td>
</tr>
<tr>
<td>Ser</td>
<td>7.78</td>
<td>6.55</td>
<td>5.17</td>
<td>2.93</td>
<td>2.73</td>
<td>2.44</td>
</tr>
<tr>
<td>Tyr</td>
<td>12.92</td>
<td>10.61</td>
<td>7.81</td>
<td>2.84</td>
<td>2.77</td>
<td>2.59</td>
</tr>
<tr>
<td>Trp</td>
<td>0.97</td>
<td>0.89</td>
<td>0.81</td>
<td>1.49</td>
<td>1.47</td>
<td>1.43</td>
</tr>
<tr>
<td>Thr</td>
<td>9.41</td>
<td>7.85</td>
<td>6.26</td>
<td>2.86</td>
<td>2.55</td>
<td>2.37</td>
</tr>
<tr>
<td>His</td>
<td>16.03</td>
<td>12.8</td>
<td>10.75</td>
<td>5.94</td>
<td>4.98</td>
<td>4.77</td>
</tr>
</tbody>
</table>

A.3 Mean First Passage Times, $\tau_{mfp}$

Table A.5: Mean first passage time $\tau_{mfp}$ obtained by measuring the average time for passage of a hydrogen bond along the bond length reaction coordinate and fit to an Arrhenius behaviour. All units in kJ/mol.

<table>
<thead>
<tr>
<th>Residue</th>
<th>$\tau_{mfp}$</th>
<th>$\Delta G_{act}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg</td>
<td>18</td>
<td>10.02</td>
</tr>
<tr>
<td>Asn</td>
<td>23</td>
<td>7.29</td>
</tr>
<tr>
<td>Asp</td>
<td>35</td>
<td>6.64</td>
</tr>
<tr>
<td>Gln</td>
<td>20</td>
<td>10.47</td>
</tr>
<tr>
<td>Glu</td>
<td>34</td>
<td>6.19</td>
</tr>
<tr>
<td>Lys</td>
<td>18</td>
<td>3.62</td>
</tr>
<tr>
<td>Ser</td>
<td>33</td>
<td>5.94</td>
</tr>
<tr>
<td>Tyr</td>
<td>49</td>
<td>0.95</td>
</tr>
<tr>
<td>Trp</td>
<td>11</td>
<td>3.55</td>
</tr>
<tr>
<td>Thr</td>
<td>37</td>
<td>8.38</td>
</tr>
<tr>
<td>His</td>
<td>39</td>
<td>8.53</td>
</tr>
</tbody>
</table>
Appendix B

Hydrogen Bond Energies

B.1 Hydrogen Bond Energies from $\tau_{HB}$

Table B.1: Hydrogen bond energies between residues and solvent in the four small simulations calculated using equation 2.25. All energies in kJ/mol.

<table>
<thead>
<tr>
<th>Residue</th>
<th>$\Delta G_{NMR1}$</th>
<th>$\Delta G_{NMR3}$</th>
<th>$\Delta G_{NMR9}$</th>
<th>$\Delta G_{NMR15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asn-1</td>
<td>6.38</td>
<td>6.33</td>
<td>7.62</td>
<td>6.26</td>
</tr>
<tr>
<td>Tyr-3</td>
<td>7.90</td>
<td>7.46</td>
<td>7.96</td>
<td>7.82</td>
</tr>
<tr>
<td>Gln-5</td>
<td>5.54</td>
<td>6.36</td>
<td>5.79</td>
<td>5.92</td>
</tr>
<tr>
<td>Lys-8</td>
<td>6.04</td>
<td>6.22</td>
<td>5.93</td>
<td>6.00</td>
</tr>
<tr>
<td>Asp-9</td>
<td>7.05</td>
<td>7.48</td>
<td>7.01</td>
<td>6.91</td>
</tr>
<tr>
<td>Ser-13</td>
<td>6.04</td>
<td>6.68</td>
<td>6.74</td>
<td>6.51</td>
</tr>
<tr>
<td>Ser-14</td>
<td>7.23</td>
<td>6.30</td>
<td>5.76</td>
<td>5.99</td>
</tr>
<tr>
<td>Arg-16</td>
<td>5.99</td>
<td>5.06</td>
<td>6.12</td>
<td>6.88</td>
</tr>
<tr>
<td>Ser-20</td>
<td>5.67</td>
<td>6.89</td>
<td>5.59</td>
<td>6.36</td>
</tr>
</tbody>
</table>

Table B.2: Hydrogen bond energies between residues and solvent in the four large simulations calculated using equation 2.25. All energies in kJ/mol.

<table>
<thead>
<tr>
<th>Residue</th>
<th>$\Delta G_{NMR1}$</th>
<th>$\Delta G_{NMR3}$</th>
<th>$\Delta G_{NMR9}$</th>
<th>$\Delta G_{NMR15}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asn-1</td>
<td>6.90</td>
<td>7.94</td>
<td>8.82</td>
<td>6.53</td>
</tr>
<tr>
<td>Tyr-3</td>
<td>7.45</td>
<td>7.79</td>
<td>7.92</td>
<td>7.87</td>
</tr>
<tr>
<td>Gln-5</td>
<td>6.05</td>
<td>5.81</td>
<td>6.07</td>
<td>6.17</td>
</tr>
<tr>
<td>Lys-8</td>
<td>6.09</td>
<td>6.13</td>
<td>6.06</td>
<td>6.23</td>
</tr>
<tr>
<td>Asp-9</td>
<td>8.35</td>
<td>7.27</td>
<td>7.10</td>
<td>7.13</td>
</tr>
<tr>
<td>Ser-13</td>
<td>6.26</td>
<td>6.71</td>
<td>6.63</td>
<td>5.98</td>
</tr>
<tr>
<td>Ser-14</td>
<td>6.65</td>
<td>6.12</td>
<td>6.21</td>
<td>6.21</td>
</tr>
<tr>
<td>Arg-16</td>
<td>6.60</td>
<td>6.08</td>
<td>6.27</td>
<td>6.46</td>
</tr>
<tr>
<td>Ser-20</td>
<td>5.65</td>
<td>5.67</td>
<td>5.77</td>
<td>5.88</td>
</tr>
</tbody>
</table>
### Table B.3: Hydrogen bond energies for Ace-Gly-X-Gly-NH$_2$ peptides calculated using equation 2.25. All values in kJ/mol

<table>
<thead>
<tr>
<th>Residue</th>
<th>$\Delta G_{cor}$</th>
<th>$\Delta H_{cor}$</th>
<th>$T \Delta S_{cor}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg</td>
<td>7.56</td>
<td>13.17</td>
<td>5.61</td>
</tr>
<tr>
<td>Asn</td>
<td>7.74</td>
<td>13.87</td>
<td>6.12</td>
</tr>
<tr>
<td>Asp</td>
<td>9.93</td>
<td>10.72</td>
<td>0.78</td>
</tr>
<tr>
<td>Gln</td>
<td>7.64</td>
<td>11.35</td>
<td>3.70</td>
</tr>
<tr>
<td>Glu</td>
<td>10.04</td>
<td>12.55</td>
<td>2.51</td>
</tr>
<tr>
<td>Lys</td>
<td>8.69</td>
<td>10.85</td>
<td>2.16</td>
</tr>
<tr>
<td>Ser</td>
<td>9.18</td>
<td>9.88</td>
<td>0.70</td>
</tr>
<tr>
<td>Tyr</td>
<td>10.52</td>
<td>11.60</td>
<td>1.08</td>
</tr>
<tr>
<td>Trp</td>
<td>6.57</td>
<td>7.81</td>
<td>1.24</td>
</tr>
<tr>
<td>Thr</td>
<td>9.60</td>
<td>11.01</td>
<td>1.41</td>
</tr>
<tr>
<td>His</td>
<td>10.42</td>
<td>16.01</td>
<td>5.60</td>
</tr>
</tbody>
</table>