Ab initio Prediction of the Conformation of Solvated and Adsorbed Proteins

by

Milan Mijajlovic

A thesis submitted to The University of Edinburgh for the degree of

Doctor of Philosophy

The University of Edinburgh

2008

Abstract

Proteins are among the most important groups of biomolecules, with their biological functions ranging from structural elements to signal transducers between cells. Apart from their biological role, phenomena related to protein behaviour in solutions and at solid interfaces can find a broad range of engineering applications such as in biomedical implants, scaffolds for artificial tissues, bioseparations, biomineralization and biosensors. For both biological and engineering applications, the functionality of a protein is directly related to its three-dimensional structure (*i.e.* conformation). Methods such as homology and threading that depend on a large database of existing experimental knowledge are the most popular means of predicting the conformation of proteins in their native environment. Lack of sufficient experimentally-derived information for non-native environments such as general solutions and solid interfaces prevents these knowledge-based methods being used for such environments. Resort must, instead, be made to so-called *ab initio* methods that rely upon knowledge of the primary sequence of the protein, its environment, and the physics of the interatomic interactions. The development of such methods for non-native environments is in its infancy – this thesis reports on the development of such a method and its application to proteins in water and at gas/solid and water/solid interfaces. After introducing the approach used – which is based on evolutionary algorithms (EAs) – we first report a study of polyalanine adsorbed at a gas/solid interface in which a switching behaviour is observed that, to our knowledge, has never been reported before. The next section reports work that shows the combination of the Langevin dipole (LD) solvent method with the Amber potential energy (PE) model is able to yield solvation energies comparable to those of more sophisticated methods at a fraction of the cost, and that the LD method is able to capture effects that arise from inhomogenities in the water structure such as H-bond bridges. The third section reports a study that shows that EA performance and optimal control parameters vary substantially with the PE model. The first three parts form the basis of the last part of the thesis, which reports pioneering work on predicting *ab initio* the conformation of proteins in solutions and at water/solid interfaces.

Acknowledgments

I would like to thank my supervisor Dr Mark Biggs for the invaluable guidance throughout my PhD studies and for the precious advices he provided in research as well as in everyday life.

Many thanks to Dr Dusan Djurdjevic, who helped building the framework for the project and introduced me to the computational details of protein simulations. I would also like to thank Dr Alex Buts for sharing his great theoretical knowledge and helping in building new mathematical models.

Finally, many thanks to the members of my family back in Serbia, who have provided moral support in difficult moments, and to all of my flatmates who had to bear with me all these years.

Declaration

I declare that this PhD thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

In Edinburgh,

Milan Mijajlovic

List of Publications Produced from this Thesis

Chapter 4: M. Mijajlovic, D. P. Djurdjevic, M. J. Biggs, “The behavior of an EA used in *ab initio* protein fold prediction can vary significantly with the potential energy model used”, submitted to *Evol. Comput.*

Chapter 5: M. Mijajlovic, M. J. Biggs, “Study of conformational switching in polyalanine at solid surfaces using molecular simulation”, *J. Phys. Chem. C* **2007**, *111*, 15839-15847.

Chapter 6: M. Mijajlovic, M. J. Biggs, “On use of the Amber potential with the Langevin dipole method”, *J. Phys. Chem. B* **2007**, *111*, 7591-7602.

Table of Contents

Chapter 1. Introduction

Proteins are biomolecules that underpin life. The function of these proteins – which can vary from structural over immunological to material and signal transporting (Bogen, 1968; Goodsell, 1996; Rappé and Casewit, 1997; Siegel et al., 2006) – is linked directly to the three dimensional (3D) conformation of the proteins, which in the native state is termed the "tertiary structure". This tertiary structure is dictated by the amino acid sequence (i.e. the primary structure) of the protein, and the physics of the intra-protein and protein-environment interactions (Anfinsen, 1973; Rappé and Casewit, 1997).

There has been a vast effort aimed at understanding the behaviour of proteins in their native environment such as in solutions and within biological membranes (Forrest and Sansom, 2000; Scharnagl et al., 2005). The experimental efforts are reflected in, for example, the Protein Data Bank (PDB) (Berman et al., 2000) and numerable Nobel prizes.[1] Computational methods are also making an increasing contribution to understanding protein structure and function in the native environment. The worth of such computational work is demonstrated by its pivotal role in elucidating the mechanism of development of neurodegenerative disorders, such as Alzheimer's (Nguyen and Hall, 2006), and, increasingly, in the design of drugs to treat various diseases (Frecer et al., 2004).

As demonstrated by the fields of biomedical and tissue engineering, bionanotechnology and bioprocessing amongst others, proteins are also found at the interface between the native and inorganic worlds (Kasemo, 2002). For example, protein adsorption is the first step in the body's response to inorganic implants such

[1] For example, for the Nobel Prize in Chemistry, the following received awards based on their protein structure related work: Frederick Sanger (1958), Max Ferdinand Perutz and Sir John Cowdery Kendrew (1962), and Sir Aaron Klug (1982). Prizes in other categories were also awarded for work relating to protein structure and function in their native environment.

as artificial heart valves, as shown in Figure 1.1 (Kasemo, 2002; Ratner and Bryant, 2004). As this can lead to complications and even life-threatening reactions (e.g. emboli), technologies based on understanding of protein behaviour at solid surfaces are currently being developed to eliminate such responses (Ratner and Bryant, 2004). Similar approaches are also being used in the next generation tissue scaffolds to improve spatial control over cell adhesion, which is essential for producing all but the simplest tissue (Shin et al., 2003).

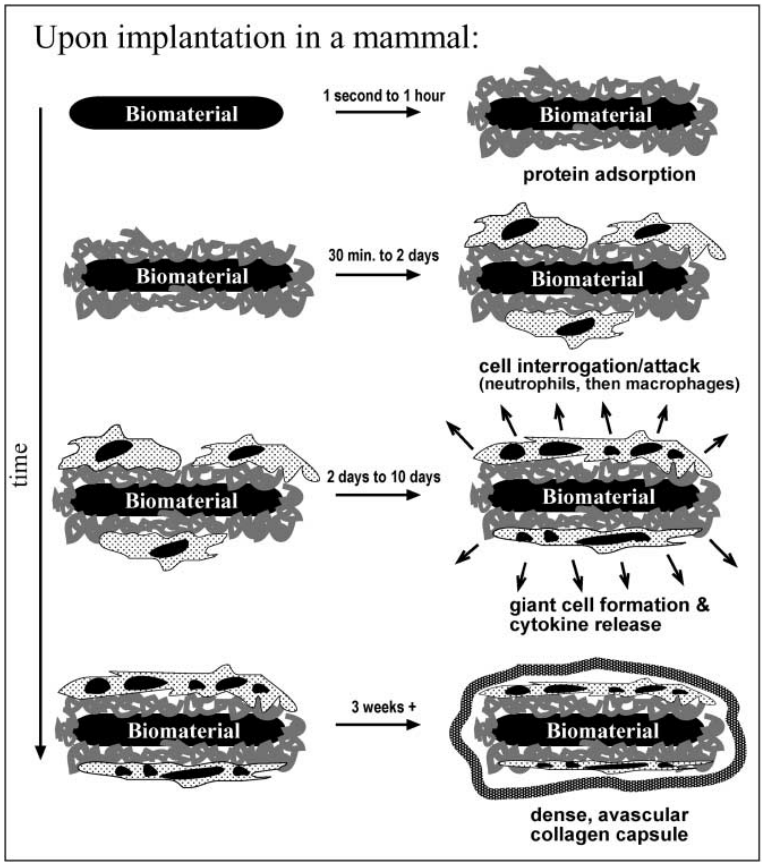


Figure 1.1 Protein mediated immunological response to implants – from (Ratner and Bryant, 2004).

Proteins are also found at solid surfaces in biosensors and bioarrays as sensing elements, analytes and foulants, as depicted in Figure 1.2 (Castillo et al., 2004; Hultschig et al., 2006). Biosensors are attractive as they can be easily miniaturised and respond rapidly, making them ideal for use outside the lab (e.g. at home), as *in vivo* sensors (e.g. glucose monitors of diabetics), for continuous monitoring of processes in industry and the environment, and at potential biohazard sites (Castillo et al., 2004; Sapsford et al., 2004). The high throughput capacity of protein arrays, on

the other hand, means they are playing an increasing role in diagnosis and drug discovery (Hultschig et al., 2006). Protein adsorption and migration on solid surfaces are also central to bioseparations (Przybycien et al., 2004) and fouling in the processes (e.g. in food industry) and beyond (Flemming, 2002).

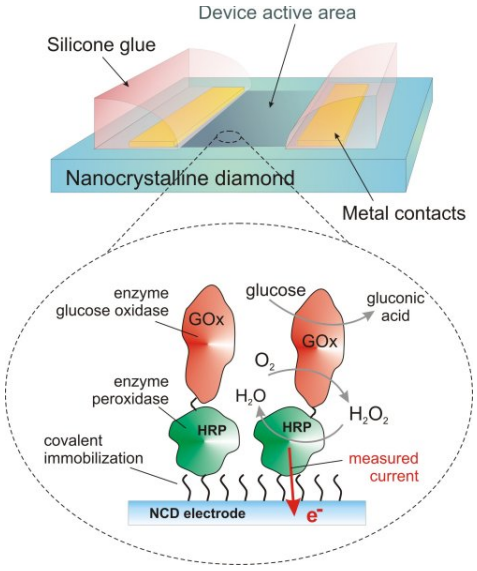


Figure 1.2 Proteins on the surface of biosensors – from (Härtl et al., 2004).

Proteins at solid interfaces are also essential to nature – examples include antifreeze proteins (AFP), shown in Figure 1.4, that allow some species to survive at sub-zero temperatures by binding to small ice crystals to inhibit their growth (Liou et al., 2000), and proteins involved in biomineralization, a process responsible for egg shell for example (Weiner and Addadi, 1997). Such processes are inspiring new "biomimetic" technologies. For example, mimicking AFPs, a number of groups have developed peptides that can control crystal growth to obtain desired crystal characteristics (Seeman and Belcher, 2002; Sarikaya et al., 2003). It is believed these peptides can also be used to self-assemble nanoscale entities to form complex multiscale structures in a manner similar to biomineralization (Seeman and Belcher, 2002; Sarikaya et al., 2003) and systems such as nanoelectronic elements and circuits illustrated in Figure 1.5 (Katz and Willner, 2004).

As in the native state, the behaviour of a protein at a solid surface and the response of the surface to the protein depend on the 3D conformation of protein. This is clearly seen in, for example, the anti-freeze protein shown in Figure 1.4 but is also evident in applications such as biosensors where sensing of a protein depends upon

the three-dimensional conformation of the binding site (Kasemo, 2002). It is clear, therefore, that understanding of the 3D conformation of proteins at solid surfaces is as important as in the native context.

Figure 1.3 Biofouling of heat exchangers in process industry – from (Flemming, 2002)

The capacity to experimentally determine the conformation of a protein at a solid interface is far more limited compared to their conformation in the native state or crystal. In particular, it is not possible to determine the 3D conformation of proteins at solid surfaces at an atomistic level but, rather, at best details such as secondary structure measures and the orientation of the peptide to the surface (e.g. (Giacomelli et al., 1999; Vermeer and Norde, 2000)). Given these experimental limitations and challenges for proteins at solid surfaces, modelling has an even more important role to play than in the study of the native state – it is this which motivated the work reported in this thesis.

Study of proteins on solid surfaces using molecular methods is still in its infancy. Of the limited work to date, much is based on simplified models, e.g. reduced molecular models (Zhdanov and Kasemo, 1997, 1998a), or rigid structures (Lu and Park, 1989; Lu et al., 1992). The few studies that use realistic models have used molecular dynamics (MD) (Raffaini and Ganazzoli, 2003, 2004a), Monte Carlo (MC) (Song and Forciniti, 2001; Mungikar and Forciniti, 2004) or *local* molecular mechanic simulations (Oren et al., 2005) – all these methods are limited in their ability to identify the likely structure of proteins on solid surfaces either because of algorithmic limitations (e.g. local molecular mechanics and standard MC) or computational expense such as in the case of the more sophisticated MC methods

and MD. The work reported here was, therefore, concerned with developing a computationally rapid means of predicting the 3D conformation of proteins on solid surfaces in the presence or otherwise of a solution phase.

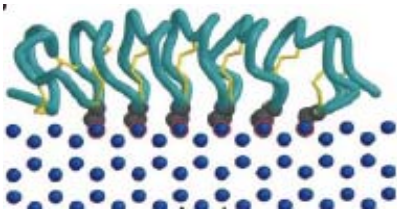


Figure 1.4 Interaction of anti-freeze protein with water molecules from ice crystal – from (Liou et al., 2000).

The approach developed involves using an evolutionary algorithm (EA) to determine, in principle, the global free energy (FE) minimum associated with a protein at a solid interface in the presence of a solvent or otherwise. The protein is modelled at an atomistic level and the interactions both between the atoms within the protein and the protein and its environment are modelled using physics-based potential energy (PE) models. This thesis describes the work undertaken in developing this approach.

There are many possible PE models available for biomolecular systems (Ponder and Case, 2003; Mackerell, 2004). Previous work by the group at Edinburgh showed that Amber (Cornell et al., 1995), a well established PE model, can be used to predict protein structures using EAs (Djurdjević, 2006). It was not, however, clear if similarly good predictions could be achieved with less computational effort using other models – the first part of the study reported here focused, therefore, on determining the EA performance for a number of common PE models. This study, which, as far as we are aware, is the first of its kind, showed that EA performance can vary significantly with the PE model.

The second aspect of the work reported here is the very first example of the use of an EA to predict the 3D conformation of a protein at a gas-solid interface. In this work, we discovered that as the surface energy is increased, polyalanine does not undergo a gradual conformational change but, rather, switches between distinct conformations at specific surface energies that depend on the size of the polyalanine molecule. Detailed analysis of the results revealed that this novel behaviour – which could be exploitable in nanotechnologies and be of relevance to disease processes –

arises from the symmetry of the polyalanine molecule and its ability to support hydrogen bonds.

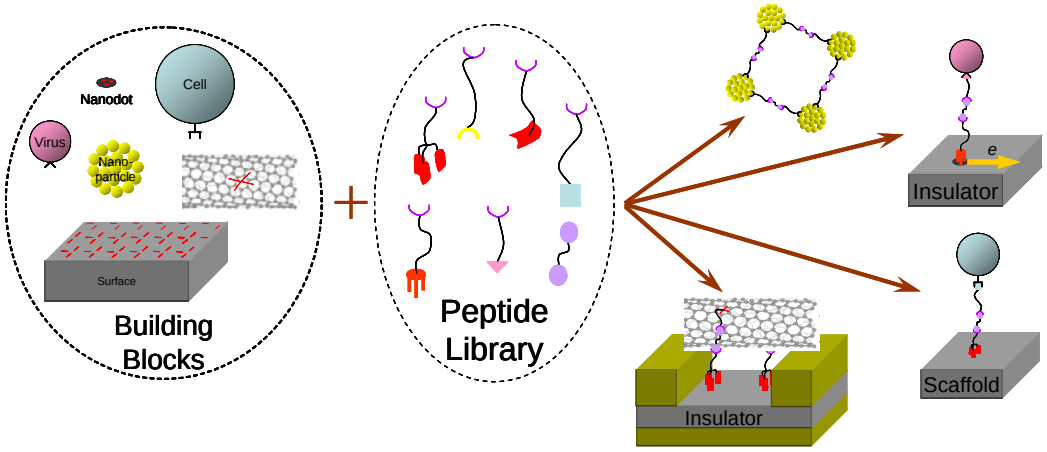


Figure 1.5 Protein facilitated assembly of elements in nanoelectronics – from (Katz and Willner, 2004)

There are a large number of ways in which solvents can be treated in biomolecular simulations (Orozco and Luque, 2000) ranging from simple implicit models such as electrostatic screening (Blaney et al., 1982) through to fully explicit solvent molecules involving multiple sites, such as TIP3P and TIP4P (Jorgensen et al., 1983). In the case of proteins at solid surfaces, phenomena such as solvent structuring between the protein and solid surface and hydrogen-bond bridging are likely to be important (Beglov and Roux, 1995; Bujnowski and Pitt, 1998) – it is, therefore, important that the solvent model used be able to capture such effects. Implicit models cannot capture these effects, indicating explicit models are necessary. However, the often used fully explicit models are computationally very expensive. We, therefore, investigated a semi-explicit model called Langevin dipoles (Warshel and Levitt, 1976; Florián and Warshel, 1997) to determine if this treatment of the solvent can capture complex phenomena at low computational cost compared to the traditional fully explicit methods. We show for the first time that Langevin dipoles combined with the Amber PE model can in fact predict the solvation energies of bio-related molecules as accurately as the state-of-the-art explicit methods at just a fraction of their cost.

The final part of the work undertaken here brings together all the previous elements of the work to study met-enkephalin, a pentapeptide, firstly in water and then, finally, at a water-graphite interface – both studies are firsts as far as we are

aware. This work revealed that water structuring and hydrogen bond-bridges play a role in the structure of proteins at solid surfaces as anticipated.

The thesis is structured as follows. A review of previous models for proteins at solid interfaces is first undertaken. This is then followed by a summary of the basic methods used in the work undertaken here, whilst further details are contained in appendices. The next four chapters report each of the studies undertaken here as outlined above. The final chapter summarizes the work and discusses some possible future work.

Chapter 2. Literature Review

2.1. Introduction

Given the overall aim of this project, the most obvious previous work of relevance here are studies which use computational methods to investigate conformation of proteins in solutions and at solid-fluid interfaces. Before the relevant numerical methods are introduced, however, it is helpful to provide a brief introduction into their theoretical background, i.e. into the statistical mechanics of proteins. Since the most common techniques applied in molecular simulations of proteins in solutions and at solid surfaces are Monte Carlo and molecular dynamics, these methods are, then, introduced in their basic form and with some recent improvements aimed at increased efficiency. Subsequently, an alternative approach to protein structure prediction, based on the Anfinsen hypothesis (Anfinsen, 1973) is discussed along with its advantages and limitations. This is followed by a more detailed description of the previous molecular simulation studies of proteins at solid surfaces. Emphasis is, then, placed on water representation, as water is the most common solvent in which proteins can be found. Finally, as met-enkephalin peptide is frequently considered here, a review of both experimental and simulation based methods for determination of its conformation is provided.

2.2. Statistical Mechanics of Proteins

The main goal of our work is to develop a method for the prediction of protein 3D structure. A standard procedure applied for this purpose is to explore the free energy surface of proteins (discussed in more details in Chapter 3). In order to relate microscopic properties (e.g. protein molecular structure) with thermodynamic properties of a system, such as the free energy, it is a common practice to revert to statistical mechanics. The approach used in statistical mechanics is to express

mechanical (i.e. temperature independent) thermodynamic properties of a macroscopic system in terms of the average of corresponding properties in the individual microscopic states (McQuarrie, 1976). The internal energy of a protein can, thus, be represented as an average of the potential energies of individual observed protein conformations. In order to perform sampling of various microscopic states of the system, a concept of ensembles is commonly used in statistical mechanics.

2.2.1. Canonical Ensemble

The concept of an ensemble of systems, introduced by Gibbs, represents a collection of a very large number of systems in different microscopic states constructed in a way that each microscopic system conforms to a set of macroscopic thermodynamic properties, such as the temperature or the volume of the system (McQuarrie, 1976). The most often used ensemble in statistical mechanics is the canonical ensemble, characterised by the constant values of number of particles, volume and the temperature of the system (McQuarrie, 1976). Since protein 3D structure studies commonly operate with a single molecule on a constant temperature, the canonical ensemble can also be used in statistical mechanical methods for protein structure prediction.

The macroscopic state of the entire ensemble is specified by the number of systems in individual microscopic states. If the number of systems in a state i is denoted as a_i, the macroscopic state of the ensemble may be represented as a multidimensional vector $\mathbf{a}$, $\mathbf{a} = \{a_1, a_2, a_3, ..., a_i, ...\}$. The number of ways in which a particular value of $\mathbf{a}$ may be obtained is denoted as $W(\mathbf{a})$. If the total number of systems in the ensemble is N, the probability of any one of them to be in a state i is (McQuarrie, 1976)

$$P_i = \frac{1}{N} \frac{\sum_{\mathbf{a}} a_i W(\mathbf{a})}{\sum_{\mathbf{a}} W(\mathbf{a})} \tag{2.1}$$

where the summation is conducted over all possible macroscopic states $\mathbf{a}$. The canonical ensemble average of a mechanical property M can then be calculated from the value of the property in all individual systems and probabilities to observe the systems in particular microscopic states (McQuarrie, 1976)

$$\langle M \rangle = \sum_i M_i P_i \tag{2.2}$$

The probability P_i is commonly replaced by a weight function w_i (Allen and Tildesley, 1989)

$$w_i = Q_{ens} P_i \tag{2.3}$$

where Q_{ens} is the partition function, defined as the sum of weight functions over all possible microscopic states (Allen and Tildesley, 1989)

$$Q_{ens} = \sum_i w_i \tag{2.4}$$

In the canonical ensemble, the weight function is expressed as a function of the Hamiltonian of the system, H (Allen and Tildesley, 1989)

$$w_i = \exp\left(-\frac{H_i}{k_B T}\right) \tag{2.5}$$

where the Hamiltonian is a sum of kinetic and potential energy for each particle of the system (Allen and Tildesley, 1989), while k_B is the Boltzmann factor and T temperature in K. The weight function of the canonical ensemble is also referred to as the Boltzmann factor (Frenkel and Smit, 1996). According to equation (2.4), the partition function of the canonical ensemble, Q_{NVT}, is calculated as (Allen and Tildesley, 1989)

$$Q_{NVT} = \sum_i \exp\left(-\frac{H_i}{k_B T}\right) = \frac{1}{n!}\frac{1}{h^{3n}} \int \exp\left(-\frac{H(\mathbf{r},\mathbf{p})}{k_B T}\right) d\mathbf{r} d\mathbf{p} \tag{2.6}$$

where the sum over microstates is replaced with an integral over all possible particle positions, $\mathbf{r}$, and momenta, $\mathbf{p}$, of a system with n particles, while h is the Planck constant.

Once the value of the partition function is known, it can be used to calculate the free energy, G, of a macroscopic system (Allen and Tildesley, 1989)

$$G = -k_B T \ln Q_{NVT} \tag{2.7}$$

The integration of equation (2.6) in order to obtain the partition function is, however, a nontrivial problem and it has to be performed using molecular simulation methods. Two molecular simulation techniques commonly used in protein simulation studies are Monte Carlo (MC) and molecular dynamics (MD). These two techniques will be described in more details in the next section.

2.3. Monte Carlo and Molecular Dynamics in Protein Simulations

Monte Carlo and molecular dynamics are standard molecular simulation approaches for exploration of free energy surface of proteins and other molecules or molecular ensembles. They rely on calculation of interatomic potentials (MC) or forces between the atoms (MD), thus operating on a potential energy (PE) surface of the system. However, in addition to this, they are also equipped with a mechanism for generating structures with similar energies. As the entropy of a macroscopic system is directly related to the number of the unique microscopic states accessible to it (McQuarrie, 1976; Frenkel and Smit, 1996), counting of the unique protein 3D structures in a simulation gives MC and MD the ability to incorporate entropic contribution of an ensemble of the structures. In other words, by combining PE with entropic contributions, both MC and MD are well suited for the exploration of free energy surface of proteins.

2.3.1. Basic Monte Carlo Implementation

According to the Boltzmann distribution (Frenkel and Smit, 1996), a probability for a thermodynamic system to be found in a microscopic state *i* with the total energy E_i is equal to

$$P_i = \frac{\exp\left(-E_i/k_BT\right)}{\sum_j \exp\left(-E_j/k_BT\right)} \tag{2.8}$$

where the denominator represents the summation of Boltzmann factors, $\exp\left(-E_j/k_BT\right)$ over all possible quantum states. This sum is also known as the partition function, *Q*. Using equation (2.8), the average energy of the system, $\langle E \rangle$, may be calculated by summing up energies of all possible quantum states multiplied with their corresponding probabilities (Frenkel and Smit, 1996)

$$\langle E \rangle = \sum_i E_i P_i = \frac{\sum_i E_i \exp\left(-E_i/k_BT\right)}{\sum_i \exp\left(-E_i/k_BT\right)} \tag{2.9}$$

In a similar fashion, thermodynamic average of an arbitrary variable *A*, $\langle A \rangle$, can be calculated from the equation (2.10)

$$\langle A \rangle = \sum_i A_i P_i = \frac{\sum_i A_i \exp\left(-E_i / k_B T\right)}{\sum_i \exp\left(-E_i / k_B T\right)} \tag{2.10}$$

where A_i is the value of variable A in a quantum state i.

The total energy of the system may be represented as a sum of the kinetic and potential energy contributions, while the sums over quantum states may be replaced by integrals over coordinates and momenta of all atoms (Frenkel and Smit, 1996). In a system of N atoms, the average value of A is, thus, calculated as

$$\langle A \rangle = \frac{\int A\left(\mathbf{r}^N, \mathbf{p}^N\right) \exp\left[-\left(\sum_{i=1}^{N} p_i^2 / 2m_i + U\left(\mathbf{r}^N\right)\right) \Big/ k_B T\right] d\mathbf{r}^N d\mathbf{p}^N}{\int \exp\left[-\left(\sum_{i=1}^{N} p_i^2 / 2m_i + U\left(\mathbf{r}^N\right)\right) \Big/ k_B T\right] d\mathbf{r}^N d\mathbf{p}^N} \tag{2.11}$$

where p_i is the momentum and m_i the mass of atom i. $\mathbf{r}^N$ contains coordinates of all N atoms, while $\mathbf{p}^N$ consists of their momenta. The kinetic energy part of the integral is a quadratic function of the momenta and can be solved analytically. It is the potential energy of the system that requires application of numerical methods, such as Monte Carlo, in order to be solved.

Being only a function of the system configuration, the part of the variable A calculated through PE integrals is referred to as configurational, A_{conf}. The thermodynamic average of the configurational part of the variable A is then calculated as

$$\langle A_{conf} \rangle = \frac{\int A\left(\mathbf{r}^N\right) \exp\left[-U\left(\mathbf{r}^N\right) / k_B T\right] d\mathbf{r}^N}{\int \exp\left[-U\left(\mathbf{r}^N\right) / k_B T\right] d\mathbf{r}^N} \tag{2.12}$$

where the denominator is denoted as the configurational part of the partition function, Z (Frenkel and Smit, 1996). Whilst numerical solving of the partition function, as well as of the integral in the numerator of equation (2.12), is still computationally intractable problem, Metropolis et al. (1953) have developed a method for efficient sampling of the ratio of the two integrals.

Monte Carlo is, in principle, a random walk method, i.e. the equation (2.12) would be solved by randomly creating structures and calculating their potential energies. In the Metropolis scheme, however, the so called importance sampling of structures is utilised. In this approach, new structures are not generated completely

randomly, but with a relative probability proportional to the Boltzmann factor (Frenkel and Smit, 1996). The simulation is initiated by creating the system in a random configuration $\mathbf{r}^N(o)$, characterised with a finite value of the Boltzmann factor, $\exp\left(-U\left(o\right)/k_BT\right)$. In protein simulations, this condition corresponds to a random conformation without overlaps. In the next step, a new configuration of the system, $\mathbf{r}^N(n)$, is created by randomly displacing atoms of the previous configuration. The Boltzmann factor of the new configuration is calculated as $\exp\left(-U\left(n\right)/k_BT\right)$. The numerical core of the Metropolis scheme consists of evaluation of the probability of transition between the old and the new configuration. In the original implementation of the method (Metropolis et al., 1953), a move from the old, *o*, to the new conformation, *n*, is always accepted if it does not lead to increase in the potential energy. If $U\left(n\right)>U\left(o\right)$, however, the move is accepted with the probability that is calculated from the difference of the Boltzmann factors of the two configurations. If probability of accepting a random move from configuration *o* to configuration *n* is denoted as $\pi\left(o\to n\right)$, the Metropolis scheme can be summarised with the following formula

$$\pi\left(o\to n\right)=\begin{cases}1 & ,U\left(n\right)\le U\left(o\right)\\ \exp\left(-\dfrac{U\left(n\right)-U\left(o\right)}{k_BT}\right) & ,U\left(n\right)>U\left(o\right)\end{cases} \tag{2.13}$$

In summary, a molecular simulation of a protein with the basic MC algorithm is conducted by first generating a random nonoverlapping protein conformation, which is then subjected to random incremental conformational changes, where each change is accepted with the probability defined by equation (2.13), in which *U* is the potential energy of generated protein conformations, calculated using one of the PE models (discussed in more details in Appendix B). An example of application of such an algorithm is the study of met-enkephalin conformation by Li and Scheraga (1987), in which protein conformation was represented through a set of dihedral angles and MC moves consisted of random changes in these angles. Other early MC studies of proteins, such as that of Krigbaum and Lin (1982) or Kolinski et al. (1986) relied on application of simplified, lattice models of proteins, but used the same MC procedure outlined by Metropolis et al. (1953).

2.3.2. Basic Molecular Dynamics Implementation

Whilst Monte Carlo samples configurations of a protein (or any other system) using random moves, molecular dynamics (MD) is a deterministic method in which new configurations are obtained using Newton's laws of motion (Frenkel and Smit, 1996). The central part of an MD algorithm is the calculation of forces that act on all protein atoms. The resultant force acting on an atom *i* is calculated as a gradient of its potential energy (van Gunsteren and Berendsen, 1990)

$$\mathbf{F}_i = -\frac{\partial U_i\left(\mathbf{r}^N\right)}{\partial \mathbf{r}_i} \tag{2.14}$$

where the potential energy, *U*, is calculated from the same PE models used for the calculation of potential energies in MC simulations. Once the force acting on each atom is obtained, it is used to calculate the acceleration of the atoms according to Newton's second law of motion

$$\mathbf{a}_i = \frac{\mathbf{F}_i}{m_i} \tag{2.15}$$

Accelerations are then used to calculate displacements of atoms after an arbitrary time step according to the definition of acceleration as the second derivative of position over time

$$\mathbf{a}_i = \frac{\mathrm{d}^2\mathbf{r}_i\left(t\right)}{\mathrm{d}t^2} \tag{2.16}$$

The evolution of position vector, $\mathbf{r}_i$, in time is obtained through integration of the equation (2.16) using one of the established numerical methods, such as Verlet algorithm (Verlet, 1967; Frenkel and Smit, 1996). The atomic displacements obtained this way are the effectors of conformational changes in protein. Analogous to the MC approach, these incremental structural changes enable sampling of conformations around local minima, thus adding the entropic contribution to the calculated intramolecular potential energy. Similarly, in an MD simulation, relevant thermodynamic properties may be obtained by averaging over trajectory, which is analogous to ensemble averaging in MC, as expressed by equation (2.10).

Although their reliance on natural laws of motion implies that MD simulations mimic the real movement of molecules, they are still not as reliable as laboratory experiments. Their ability to retrieve information on atomic level, which is often

inaccessible to experimental methods, makes them ideal as complementary methods to experiments (van Gunsteren and Berendsen, 1990). MD simulations are, for instance, often used in combination with NMR data, to determine the 3D structure of proteins (Clore and Gronenborn, 1987; Kaptein et al., 1988; van Gunsteren and Berendsen, 1990). Analogously, MD has been applied in refinement of protein structures obtained from crystallographic data (Brünger et al., 1987). In addition to this, a large number of studies have used MD approach to predict the 3D structure of proteins without restraints imposed by experimental data (McCammon et al., 1977; Tirado-Rives et al., 1993; Huston and Marshall, 1994; Zhang and Hermans, 1994; Lin and Baumgaertner, 2000; Hénin et al., 2005).

2.3.3. Algorithmic Improvements in Monte Carlo and Molecular Dynamics

Basic implementations of MC and MD methods have found widespread use in simulation of gases and liquids with simple, mono- or diatomic molecules. These systems are characterised with low energy barriers, which may easily be traversed by simple molecular simulation algorithms, thus allowing exhaustive sampling of the configurational search space. Conformational changes in proteins are, however, often characterised with very high energy barriers, which poses serious obstacle for representative sampling by simple MC and MD implementations. The algorithms often get "trapped" inside basins of local energy minima, thus spending too much computational time in sampling the structures of lower interest. Several methods have been developed with the aim of improving the ability of MC and MD to cross the energy barriers with higher efficiency.

Simulated annealing (Kirkpatrick et al., 1983) is a technique that can be combined with both MC and MD approaches. Annealing is a treatment usually used in metallurgy for removing imperfections in crystal structure. Material is exposed to a high temperature, and then slowly cooled down. Heating perturbs the crystal structure, thus removing the imperfections in it, while slow cooling allows the crystal to settle down in a more favourable state. Analogously, incremental increasing and subsequent decreasing of the temperature of a simulated molecular system allows its reconfiguration into a lower energy minimum (Rappé and Casewit, 1997). The theoretical background of the simulated annealing lies in the increased ability for crossing energy barriers on higher temperatures. For an energy barrier of the height

U_B, the probability for a molecule to cross it is expressed through the Boltzmann factor, $\exp\left(-U_B/k_BT\right)$. If a local energy minimum is surrounded by high barriers, the probability for a transition move may be very low on the room temperature. Temperature increase, however, increases the Boltzmann factor, thus allowing higher barrier crossing probability. In order to perform sampling on temperatures of interest, the system is incrementally cooled down and simulated long enough to allow a steady state to be reached for each of the temperatures (Kirkpatrick et al., 1983). If the temperature reduction is slow enough, the system is able to progress towards the global energy minimum without being trapped in higher-energy local minima (Rappé and Casewit, 1997). Simulated annealing in protein structure studies has been utilised to enhance conformational sampling in both MC (Liu et al., 2000; Nachman et al., 2002; Gordon et al., 2003) and MD algorithms (Esteve et al., 2001; D'Amelio et al., 2003; Doucet and Pelletier, 2007).

A method that is, in a phenomenological sense, similar to simulated annealing is the parallel tempering approach (Hansmann, 1997). Parallel tempering algorithm uses the same principle of enhanced efficiency for crossing energy barriers on higher temperatures. There are, however, some important differences between the two methods. Whilst simulated annealing operates with a single system, which is subjected to consecutive increase and decrease of temperature, parallel tempering approach makes several copies of the initial system and simulates each copy in parallel on different temperatures. The systems are, however, not completely independent from each other as the conformations are allowed to be exchanged between them (Hansmann, 1997). Parallel tempering operates with two main configurational moves. The first is the local update, in which a copy of the system is subjected to a regular MC or MD move without any effect from other instances of the system. Periodically, a global update is performed, in which the exchange of conformations between two copies is tried

$$\begin{aligned}\mathbf{r}^N_{new}\left(T_i\right)=\mathbf{r}^N_{old}\left(T_j\right);\mathbf{p}^N_{new}\left(T_i\right)=\mathbf{p}^N_{old}\left(T_j\right)\\ \mathbf{r}^N_{new}\left(T_j\right)=\mathbf{r}^N_{old}\left(T_i\right);\mathbf{p}^N_{new}\left(T_j\right)=\mathbf{p}^N_{old}\left(T_i\right)\end{aligned} \tag{2.17}$$

The update is global as, unlike incremental conformational changes in a single MC and MD step, a whole conformation is reshuffled if the move has been successful

(Hansmann, 1997). This substantial change of conformation is the main driving force for barrier crossing. The transition probability of a global update move is, analogously to configurational changes in a simple MC algorithm, based on Metropolis criterion (Metropolis et al., 1953; Hansmann, 1997)

$$\pi(o \to n) = \min\left(1, \exp\left(-\frac{U\left(\mathbf{r}_j^N, \mathbf{p}_j^N\right)}{k_B T_i} - \frac{U\left(\mathbf{r}_i^N, \mathbf{p}_i^N\right)}{k_B T_j} + \frac{U\left(\mathbf{r}_i^N, \mathbf{p}_i^N\right)}{k_B T_i} + \frac{U\left(\mathbf{r}_j^N, \mathbf{p}_j^N\right)}{k_B T_j}\right)\right) \quad (2.18)$$

where o and n denote the old and the new configurations, respectively. The old configuration is the one in which configurations i and j are on temperatures T_i and T_j, respectively, whilst the new configuration n corresponds to the same configurations with swapped temperatures. Like simulated annealing, parallel tempering has been used as a sampling improvement technique both in MC (Mitsutake et al., 2001; Rathore and de Pablo, 2002; Podtelezhnikov and Wild, 2005) and MD based (Sugita and Okamoto, 1999; Cheng et al., 2005; Rathore et al., 2005) analysis of protein and peptide 3D structure. MD algorithms improved by parallel tempering sampling are also referred to as replica exchange molecular dynamics (REMD) methods (Sugita and Okamoto, 1999). Due to the very low level of interactions between simulated replicas of the system, both MC and MD parallel tempering methods are very suitable for running on parallel CPU architectures (Mitsutake et al., 2001) as each CPU can be assigned a single instance of the system on a different temperature and CPU communication is performed only when replica exchanges are tried.

The weighting factor of global update move in parallel tempering – equation (2.18) – involves changes in both the potential energy and temperature, which makes it a non-Boltzmann factor. Sampling techniques which use non-Boltzmann weighting factors are referred to as generalised ensemble methods (Sugita and Okamoto, 1999). Thus, parallel tempering can be viewed as one of the generalised ensemble methods for protein structure prediction (Hansmann and Okamoto, 1999). One of the first incarnations of generalised ensemble techniques is the well known umbrella sampling (Torrie and Valleau, 1977). Umbrella sampling is used in calculation of free energy difference between two states of a system. A single MC or MD simulation with the basic algorithm on lower temperatures would sample the region around only one of them. In umbrella sampling method, however, sampling of both

states is achieved by replacing the Boltzmann factor with a modified weight function that favours parts of conformational space accessible to both states (Torrie and Valleau, 1977; Frenkel and Smit, 1996). Umbrella sampling, as other generalised ensemble techniques, can be used both within the MC and MD frameworks. Examples of its application in protein studies include analysis of met-enkephalin conformation (Bartels and Karplus, 1998) and the conformational study of the Betanova protein (Bursulaya and Brooks, 1999).

The most prominent of the generalised ensemble methods is the multicanonical algorithm (Berg and Neuhaus, 1992), also known as entropic sampling (Lee, 1993; Hansmann and Okamoto, 1999). In the canonical ensemble (in which the weight of conformations is equal to the Boltzmann factor), low temperatures are convenient for sampling around local minima, but they do not allow efficient barrier crossing. High temperatures, on the other hand, allow easy barrier crossing, but sampling of low energy regions deteriorates due to increased ability for escaping from them (Nakajima et al., 1997). The multicanonical algorithm alleviates this problem by introducing an artificial flat energy distribution, which increases efficiency of barrier crossing without affecting the efficiency of sampling the low energy barriers (Nakajima et al., 1997). Multicanonical algorithm in protein structure analysis has been combined with MC (Hansmann et al., 1996; Mitsutake et al., 2000) as well as with MD method (Hansmann et al., 1996; Nakajima et al., 1997).

Whilst discussing modifications and improvements to basic MC and MD algorithms, it should be noted that some regard genetic algorithms as a modified MC method (Hansmann and Okamoto, 1999). According to this view, the main difference between the two is that, unlike MC, genetic algorithms do not operate on a single configuration trajectory, but on a population of configurations from various regions of the search space. Genetic algorithms have, however, evolved into an independent search method and, within a broader category of evolutionary algorithms, will be an object of the study in this thesis.

2.4. Anfinsen's Hypothesis for Protein Structure

Monte Carlo and molecular dynamics methods are very convenient tools for sampling an ensemble of protein structures characterised with low potential energies. On many occasions, however, native conformation of a protein is characterised by a

single most stable conformation, in which case a more efficient numerical methods may be applied. In order to analyse alternative approaches in protein structure prediction, the free energy of a protein is first analysed in more detail.

2.4.1. Protein Free Energy and Entropy

The free energy of a macroscopic system described by canonical ensemble may be expressed in terms of potential energy and entropy of the system (McQuarrie, 1976)

$$G = U - TS \tag{2.19}$$

Potential energy of the system, U, is a consequence of interaction between the constituent particles. In protein simulations, it is usually calculated through the application of empirical potential energy models or force fields (discussed in detail in Appendix B). The entropy of a protein is a more difficult concept.

Using the statistical mechanics definition, the entropy of a macroscopic system, S, is proportional to the logarithm of the number of microscopic states available to the system (McQuarrie, 1976)

$$S = k_B \ln \Omega(T) \tag{2.20}$$

where $\Omega(T)$ is the total number of states accessible at the temperature T. In protein studies, individual microscopic states are defined as unique protein conformations and the entropy S is denoted as conformational or configurational entropy.

Trajectories between different protein conformations often involve transition over energy barriers of various heights. In order to include conformations from both sides of a barrier into the entropy calculation, the barrier has to be easily traversable by thermal fluctuations. This means that on the sampling temperature T, the protein has to be able to cross the barrier with high probability. The probability for barrier crossing, π_b, is equal to the Boltzmann factor

$$\pi_b = \exp\left(-\frac{\Delta U_b}{k_B T}\right) \tag{2.21}$$

where ΔU_b is the barrier height. It is obvious that for low energy barriers and high temperatures this probability increases, thus increasing the entropy of a protein. Low temperatures, however, increase the magnitude of the fraction, resulting in decrease of barrier crossing probability and the entropy of the system. In the extreme case, for

$T \to 0\,\mathrm{K}$, the protein becomes "frozen" in a single local minimum as the probability for conformational change becomes 0 even for the lowest of barriers: $\pi_b = \exp(-\infty) = 0$. Being unable to occupy more than a single conformation, the entropy of such a protein is 0 according to equation (2.20).

The approach that is commonly applied in molecular simulations of proteins is to decompose conformational entropy into two contributions: that of the local fluctuations in the neighbourhood of a well defined 3D structure and the contribution that corresponds to the existence of multiple distinct structures (Karplus et al., 1987). The potential energy surface is then modelled as a set of multidimensional harmonic wells separated by energy barriers. If the total number of distinct local minima (harmonic wells) is *N*, the total conformational entropy, *S*, is calculated as (Karplus et al., 1987)

$$S = \sum_{i=1}^{N} w_i S_i^v - k_B \sum_{i=1}^{N} w_i \ln w_i \tag{2.22}$$

where w_i is the Boltzmann factor associated with the well *i*, while S_i^v is its vibrational entropy, or entropy associated with the local fluctuations around the minimum. For the one-dimensional fluctuation (e.g. pendulum or rotation around a single chemical bond), the vibrational entropy can be expressed as a function of the vibrational frequency of the minimum *i*, v_i (Karplus et al., 1987)

$$S_i^v = \frac{h\nu_i}{T\left(\exp\frac{h\nu_i}{k_B T} - 1\right)} - k_B \ln\left[1 - \exp\left(-\frac{h\nu_i}{k_B T}\right)\right] \tag{2.23}$$

Since each bond in a protein molecule is an independent oscillator, this expression is replaced with a more general equation in protein studies in which explicit calculation of entropy is needed (Karplus et al., 1987).

2.4.2. Anfinsen's Hypothesis

Whilst there are instances in which proteins do not have a clearly defined 3D structure, i.e. they can be found in a range of conformations corresponding to various local minima on the potential energy surface, a distinct feature of many natural proteins is to occupy a single distinct conformation. In such cases, the Anfinsen's hypothesis (Anfinsen, 1973) is valid. According to this hypothesis, the native

conformation of a protein is a single folded conformation in which the free energy of the protein and its environment is at the minimum. When a molecule is found in a single conformation, its entropy is zero according to the previous discussion. Accordingly, Monte Carlo and molecular dynamics may be replaced by simpler and more efficient algorithms for minimisation of the potential energy surface. This assumption has been used throughout this thesis and will be discussed in more detail in Chapter 3 and the following chapters.

It should be noted, however, that Anfinsen's thermodynamic hypothesis has certain limitations. In particular, there are proteins and peptides that do not conform to a single conformation rule. An example of such molecules is the well known met-enkephalin, studied in Chapters 4 and 7 of this thesis. Short peptides, such as met-enkephalin, are characterised by an ensemble of different conformations with similar values of potential energy, rather than a single most stable 3D structure. Furthermore, the barriers between these local minima in potential energy landscape are also low, which means that molecule easily traverses from one conformation to the other. In other words, multiple conformations have similar probabilities of existence. Since this effectively increases the disorder of the system, we may say that for these peptides, entropy is increased and starts to play an important role in the free energy of the system even on the room temperature. Whilst acknowledging this role, it should be stressed that the approach used throughout this thesis, as only the initial stage in the study, deliberately neglects the entropic contribution. Some of the ways for including the entropy into the calculation will be discussed in the last chapter, devoted to future work.

2.5. Molecular Simulations of Proteins at Solid Surfaces

Systems of interest in protein adsorption consist of several elements. Whilst the protein and the solid surface are the most obvious physical elements of the system, water and other solvents may also be present. Numeric elements include various mathematical models of the physical elements, as well as methods for calculation of free energy of the system. Table 2.1 provides a brief overview of various physical and numeric elements applied in relevant protein adsorption studies. Classification of molecular simulation methods can be performed using any of these elements. This

review classifies the methods based on the complexity of mathematical model utilised in protein representation.

2.5.1. Simple Geometry for Protein Molecule Models

The simplest models utilised in protein adsorption studies are those that do not recognise any intramolecular details. This depiction is, of course, very efficient from perspective of computational cost since it represents a protein as a simple particle that can be described with a small number of parameters. An example of such an approach can be found in an early study of protein adsorption kinetics (Zhdanov and Kasemo, 1998a). Zhdanov and Kasemo have used Monte Carlo simulations in which adsorbed proteins are represented as disks with variable radius. Surprisingly, although the model does not deal with any structural details, it is still able to represent adsorption induced deformation of proteins by varying the radii of their corresponding disks.

Using a similar philosophy, other simplified protein models have been developed, although with somewhat more detailed internal structure. Yet, despite increasing the number of degrees of freedom, these models were not able to capture any conformational changes. Gorba et al. (2004) have used a very simple rigid sphere representation of cytochrome c molecules in their Brownian dynamics study of protein behaviour on charged surfaces. Compared to a simple disk representation of Zhdanov and Kasemo, their protein model featured a hard sphere with a charge and a dipole in its center. Utilisation of this model allows monitoring of energetic changes related to protein orientation, but hard sphere puts a limitation on analysis of conformational changes. Similar model for adsorption of lysozyme

Table 2.1 Overview of simulations of protein adsorption

Reference	Protein					Sol[f]	Surface			PE model[j]	Method[k]
	Name[a]	Size[b]	Model[c]	DOF[d]	Initial state[e]		Nat[g]	Pot[h]	Model[i]		
(Roush et al., 1994)	Rat cytochrmome b_5	13k	A	R^1, H	HOM	I	AEM		A	ES	MM
(Noinville et al., 1995)	1ALC and 7LYZ	123, 129	A	R^2, T^2, H	PDB	I	AEM	Amber	A	Amber	MM
(Juffer et al., 1996)	1CUS + 15 variants	200	A	R^1, H	PDB	I	CS		S	ES	MC
(Bujnowski and Pitt, 1998)	Leu-enkephalin	5	A	R^1, H		E	CPE	CVFF	A	CVFF	MD
(Zhdanov and Kasemo, 1998b)	27-mer	27	ONB	H, T^2, B	N	N	SA		S	-	MC
(Castells et al., 2002)	27-mer	27	ONB	H, T^2, B	N	N	SA		S	-	MC
(Shang and Geva, 2005)	128-mer	128	OFFB	B	RAND	I	SA, SR		A	A	LD
(Knotts IV et al., 2005)	(Residues 10-55) 1bdd	46	OFFB	B	RAND	N	SA, SN		S	A	MC
(Griffin et al., 2005)	2 × β-barrel HT model	46	OFFB	B	GR (SA)	N	SR; SN		S	A	REMD
(Skepö et al., 2006)	Proline-rich protein 1	150	OFFB	B	RAND	I	NCS		A	A	MC
(Friedel et al., 2006)	2 × β-barrel HT model	46	OFFB	B	GR (SA)	N	SR		S	A	RELD
(Ravichandran et al., 2001)	7LYZ	129	A	H, T^2, H	PDB	I	PCS		A	Amber	BD
(Zhou et al., 2003)	1IGY and 1IGT	1294, 1316	UR	R^3, H, T^2	PDB	I	S-Au	LJ+ES	A	CHARMM	MC
(Sun et al., 2005)	7LZY	129	UA	R^3, H	PDB	I	S-Au	LJ+ES	A	A/GROMACS	E-MM
(Song and Forciniti, 2001)	N-DDIIDDII-C	8	UA	R^3, H, T^2	LR (MC); α-helix	E	FCC	LJ+ES	A	GROMACS	MC
(Mungikar and Forciniti, 2004)	N-$(DDII)_n$-(C)	n = 2, 4, 5	UA	R^3, H, T^2, B	LR (MC) ; α-helix	E	FCC	LJ+ES	A	GROMACS	MC
(Mungikar and Forciniti, 2006)	N-DDIIDDII-C	8	UA	R^3, H, T^2, B	LR (MC) ; α-helix	E×2	FCC	LJ+ES	A	?	MC
(Braun et al., 2002)	GBP1, 2 and 3	84, 84, 94	UA	R^3, H, T^2, B	HOM	E	Au	LJ	S	CHARMM26	LD
(Raffaini and Ganazzoli, 2003)	Fragments of 1AO6	107, 126	A	R^3, H, T^2, B	PDB	I&E	G	CVFF	A	CVFF	MD
(Raffaini and Ganazzoli, 2004a)	1FBR	93	A	R^3, H, T^2, B	PDB	I&E	G	CVFF	A	CVFF	MD
(Kantarci et al., 2005)	4×9 residue peptides	9	A	R^3, H, T^2, B		E	Pt	CVFF	A	CVFF	MD
(Oren et al., 2005)	5 septa-peptides	7	A		Relaxed	V	Pt	LJ	A	CHARMM22	MM
(Carravetta and Monti, 2006)	4×dipeptides	2	A	R^3, H, T^2, B		E	TiO_2	DQM-LJ	A	Amber	MD
(Cormack et al., 2004)	BPTI	58	A		PDB-relaxed	E	MgO	LJ+ES	A	CVFF	MD-LM

a. Common name or PDB code.
b. Size of protein in residues or, if this is not available, approximate molecular weight.
c. Fidelity of model: A = fully atomistic; U = united atom; UR = united residue; OFFB = off-lattice bead model; ONB = on-lattice bead model.
d. The degrees of freedom considered: R^n = rotation about n axes; T^m = translation parallel to surface in m dimensions; H = distance from surface; B = backbone dihedral angles; O = other internal degrees of freedom such as bond angle and length stretch.
e. Initial state of protein: PDB = directly from PDB database; N = native state. HOM = defined from homology; LR = locally relaxed; GR = globally relaxed (method in brackets: SA = simulated annealing); RAND = completely random.
f. Solvent treatment: N = none; I = implicit; E = explicit.
g. Surface nature: G = graphite; Si = silicon; AEM = anion-exchange membrane surface; S-Au = SAM-Au; CS = charged surface; NCS = negatively charged surface; PCS = positively charged surface; CPE = crystalline polyethelene; SA = simple attractive; SR = simple repulsive; SN = simple neutral.
h. LJ = Lennard-Jones
i. Model of surface: S = smooth; A = atomic.
j. Energy model: P = Physics-based (with name in bracket); K = knowledge based; A = ad-hoc; ES = electrostatics only.
k. Method: MM = molecular mechanics; E-MM = exhaustive molecular mechanics; MC = Monte Carlo; MD = molecular dynamics; LD = Langevin dynamics; REMD = replica exchange MD; RELD = replica exchange LD. BD = Brownian dynamics

molecule on charged surfaces has been used by Carlsson and co-workers (Carlsson et al., 2004). The protein was represented as a hard sphere, with point charges added beneath the hard-sphere surface for each charged amino acid residue.

2.5.2. Lattice Model of Protein Molecules

Pioneered by Dill and co-workers in the field of protein folding (Lau and Dill, 1990; Chan and Dill, 1994), lattice models of proteins have established a prominent place in protein adsorption studies. By constraining amino acid residues to nodes of a cubic lattice, these models reduce degree of freedom for movement of a protein, thus, boosting computational performance. However, this limitation also means that it is impossible to sample all protein conformations and that those conformations that are sampled are going to be represented with lower accuracy. It is, therefore, common to see utilisation of this model only in studies concerned with the fundamentals of folding, while many applications with real proteins require finer grained representation. Application of lattice models in theoretical investigations also means that they can operate with idealised proteins in which only small number of the kinds of residues are present (e.g. polar and hydrophobic). Most of the protein adsorption studies based on the lattice model have been conducted with this simplification. It is also noticeable that all lattice-based protein adsorption studies have used Monte Carlo minimisation method, which is justifiable as high level of discretisation would render molecular dynamics inoperable.

The first studies of proteins on solid surfaces with the use of lattice models have been conducted by Zhdanov and Kasemo (Zhdanov and Kasemo, 1997, 1998b, 2000, 2001), who used the model to extend their studies of protein adsorption kinetics, as well as to conduct theoretical investigation of metastable states in protein denaturation and phenomenon of protein packing during adsorption from solutions of high concentrations. In another theoretical study, Castells et al. (2002) have shown how surfaces with different affinities toward hydrophobic and hydrophilic residues can induce different conformational changes. This result, although not directly applicable to real proteins, clearly indicates that adsorption induced conformational changes of proteins depend on the nature of both the protein and adsorbing surface.

2.5.3. Mesoscopic Protein Models

Mesoscopic models can be characterised by use of simple particles in representation of specific parts of a protein. In this sense, they are an extension to lattice models, which use separate elements for representing individual residues. On the other hand, mesoscopic models are not constrained to lattice nodes, while their particles are not restricted to single residues and can contain larger parts of the molecule with unique behaviour or structure (e.g. larger hydrophobic patches or whole helices).

An example of a mesocsopic approach applied in studies of protein adsorption is a molecular mechanics based investigation of albumin adsorption on pyrolytic carbon conducted by Mantero et al. (2002). While the albumin molecule has been separated into a hydrophilic and a hydrophobic helix, each of which has been modelled as nondeformable spheres, surrounding water was represented using an explicit model. Parameterisation of helical parts has been performed starting from their all-atom models using a physics based atomistic potential energy model.

The multiscale approach of Mantero et al. has obvious advances over lattice representation since parameterisation from atomistic models allows it to be used in modelling of real proteins. However, a major disadvantage of this model is its limitation in predicting conformational changes. By using hard spheres for individual helices, this representation does not allow any denaturation, which limits the application only to adsorption of proteins that do not demonstrably denature, with the main aim of determining their orientation on the surface.

Similar mesoscopic characteristics can be observed in colloid model applied by Zhou and co-workers in studies of immunoglobuline adsorption on charged surfaces (Sheng et al., 2002; Zhou et al., 2004). The Y-shaped antibody has been represented using a 12-bead model and subjected to Monte Carlo energy minimisation in search for optimal position of the molecule over the surface. The model did not allow any flexibility within beads and between them (i.e. the protein was rigid).

Zhou and co-workers have also diversified their methodology by application of a united-residue model in simulations of immunoglobuline adsorption (Zhou et al., 2003). United-residue models can be classified as mesoscopic since they represent whole residues as structureless particles. However, they are far finer grained than

other mesoscopic models described above. In the united-residue model of Zhou et al., each residue is reduced to a sphere centered at the position of the corresponding C_α atom. New van der Waals parameters (i.e. parameters for a whole residue), have been derived from atomistic simulations using the CHARMM potential energy model (MacKerell et al., 1998). In theory, allowing flexibility in movement of individual residues can be used to represent conformational changes. However, Zhou et al. have again restrained the antibody molecule to its rigid conformation and monitored only changes in the position of adsorbed molecule (Zhou et al., 2003).

In recent years, Knotts et al. have conducted a molecular simulation study of a bacterial protein on two types of surfaces (Knotts IV et al., 2005). They have used a bead-residue representation based on a Gō-like model of proteins (Abe and Gō, 1981; Gō and Abe, 1981; Hoang and Cieplak, 2000). Analogously to the model of Zhou and co-workers (Zhou et al., 2003), each residue was represented as a single bead. However, individual beads were not rigidly bound to each other, but connected with a spring, instead. This has allowed simulation of conformational changes – a feature that previously described mesoscopic models have not achieved. Similar to this was a study conducted by Skepö et al. (2006), in which a united-residue model was used to study conformational changes of proline-rich protein upon adsorption to a negatively charged surface.

2.5.4. Protein Models with Atomistic Details

Only in the last two decades have advances in computer technology allowed significant increase in number of protein adsorption studies based on full atomistic models. Although there are still models that utilise united-atom representations in which methyl groups (CH_n) are modelled as individual beads (Song and Forciniti, 2001), there is an increasing number of protein adsorption studies that rely on application of all-atom models. The number of degrees of freedom is significantly greater in such models, which enables much higher flexibility and accuracy than in the previously described models, but, of course, with computational cost implications.

Energy of adsorption systems in atomistic models is obtained by summing up individual interactions between all constituent atoms. Since the number of pair interactions is somewhere between a linear and a quadratic function of the number of

individual particles, it is clear that increase in the number of atoms can lead to significant growth in computational cost with protein size. It is, therefore, necessary to model these interactions with simple potential energy functions (Wilson et al., 2004). A number of empirical potential energy functions (also known as force fields) have been developed for work with proteins and other biomolecules. Some of the popular potential energy (PE) models used in protein adsorption studies are Amber (Cornell et al., 1995), OPLS (Jorgensen et al., 1996) and CVFF (Dauber-Osguthorpe et al., 1988). Application of empirical force fields significantly reduces computational time in comparison to some more detailed approaches, such as quantum mechanical energy calculation.

A group of very early reports on protein adsorption with the application of all-atom models has been published by Lu and co-workers in the early nineties (Lu and Park, 1989; Lu et al., 1992; Lu, 1993) These authors have investigated adsorption of large, biologically relevant proteins, such as lysozyme and haemoglobin on surfaces of polymers. Although polymer surfaces can not be regarded as solid in a strict sense, Lu et al. have treated them as such by using continuum surface representation, which recognises neither individual surface molecules nor their movements. Despite increasing the level of protein description, Lu et al. have kept the molecule rigid, thus studying only position and orientation of the protein on polymer surfaces. A crude attempt to study protein conformational changes during adsorption on polymer surfaces has been proposed by Lu who described conformational change of glucagons from α-helix to extended β-strand by calculating the energies of interaction of these two conformations with polyethylene surface (Lu, 1993). Still, this approach is far from prediction of conformational change based only on protein primary structure and its environment as it implies previous knowledge of adsorbed structure.

All-atom protein models have gained popularity in the last decade, especially in studies of initial stages of protein adsorption, in which conformational changes can be neglected, thus allowing monitoring of changes in orientation only (Noinville et al., 1995; Asthagiri and Lenhoff, 1997; Ravichandran et al., 2001). Keeping the protein rigid significantly reduces the number of necessary energy calculations as intramolecular potential energy remains constant throughout the simulation and is,

therefore, irrelevant in energy minimisation procedures. Since the energy of interaction between the surface and the protein still changes with the orientation, the number of interactions increases with the size of the protein. However, with rigid conformation, the scaling of number of interactions with the number of atoms is only linear (Wilson et al., 2004), which is a significant improvement over quadratic scaling evident in flexible molecules.

Some of the models for study of proteins at solid interfaces exhibit partial rigidity. Bujnowski and Pitt (1998) have, for example, conducted an investigation of water structure around enkephalin in proximity of a polymer surface. Their model keeps the backbone rigid by fixing positions of the α-carbons. This means that the bulk orientation of the protein to the surface also remains constant. At the same time, side chain atoms are allowed to move freely (with constraints implied by the potential energy model).

2.5.5. All-Atom Protein Models with Conformational Changes

Model of Bujnowsi and Pitt (1998) and other all atom approaches described above are very useful for specific purposes, such as description of initial stages of protein adsorption. However, their inability to deal with conformational changes is a serious limitation for their universal application. Along with orientation of adsorbed proteins, their conformation is a major factor that determines their biological activity (Wilson et al., 2004). Experimental studies have shown significant reduction in enzyme activities during their adsorption, which can be explained through changes in conformation (Kondo et al., 1996). It is, therefore, obvious that in many instances, only fully atomistic representations allowing conformational changes can give complete insight into the protein adsorption phenomena.

Simulation of conformational changes is usually performed using molecular dynamics or Monte Carlo method for potential energy minimisation. Raffaini and Ganazzoli have, for example, applied molecular dynamics in simulation of adsorption of albumin subdomains on the surface of graphite (Raffaini and Ganazzoli, 2003). Although conformational changes have been restricted only to isolated domains of the protein, their work has succeeded in explaining albumin adhesion on the graphite surface under flow. A model which would be based only on changes in orientation would not be able to elucidate this phenomenon. Similar

results have been obtained with adsorption of a fibronectin module (Raffaini and Ganazzoli, 2004a), although it was shown that conformational changes of the module were of a lower magnitude than those of albumin fragments. In a subsequent study, Raffaini and Ganazzoli have shown that surface induced conformational changes during adsorption on graphite are directed in a way that will increase the surface of protein exposed to the surface, but also enabling lateral interactions between the residues, thus forming parallel strands adsorbed on a graphite surface (Raffaini and Ganazzoli, 2004b). In an attempt to investigate the effect of the nature of adsorbing surface on the behaviour of proteins, Raffaini and Ganazzoli (2006) have undertaken a study of protein adsorption on hydrophilic, poly(vinyl alcohol) surface. It was shown that, despite formation of new hydrogen bonds between the protein and the hydrophilic surface, the extent of protein conformational changes is much lower than during adsorption on hydrophobic graphite surfaces.

Another study of a fibronectin fragment on a different surface has been conducted by Wilson and co-workers (Wilson et al., 2004). The surface was composed as self-assembled monolayer (SAM). Although SAMs generally have high flexibility and are, therefore, more complex than real solid surfaces, Wilson et al. have treated the surface residues as fixed, thus converting SAM into a proper solid surface (from perspective of molecular simulations). By introducing small chemical modifications to surface residues, the group has shown that the degree of denaturation (i.e. conformational change) depends on a kind of solid surface on which the protein is adsorbed. This result is analogous to the findings of Castells and co-workers (Castells et al., 2002) who have reached a similar conclusion using a much simpler, lattice representation of a protein.

Noinville et al. (2003) have used a combination of experimental study and molecular dynamics to investigate adsorption of dermaseptin on a synthetic surface. Although the surface in molecular simulations was constructed from ethane molecules, their positions were fixed and the surface treated as solid. Analogously to the findings of Castells and co-workers (Castells et al., 2002) and Wilson et al. (2004), it was found that dermaseptin molecule undergoes different conformational changes depending on the hydrophobic character of the surface.

Several groups have also studied adsorption of proteins and short peptides on metal surfaces (Braun et al., 2002; Bizzarri et al., 2003; Imamura et al., 2003; Bizzarri, 2006; Yang and Zhao, 2007). Adsorption on metal surfaces is, however, obscured by protein and metal polarisation (Imamura et al., 2003) or even by formation of chemical bonds between the protein and the surface atoms (Bizzarri, 2006).

While Monte Carlo simulations are very suitable for lattice-based models of protein adsorption, it is apparent that molecular dynamics is a favoured approach for all-atom models that include conformational changes. Examples of Monte Carlo minimisation applied in the all-atom flexible models include studies conducted by Mungikar and Forciniti (Mungikar and Forciniti, 2004, 2006), in which adsorption of peptides of various length on charged solid surfaces is simulated in the presence of explicitly represented water.

2.5.6. Summary of Molecular Simulation Methods for Protein Adsorption

The main conclusion that can be drawn from the previous protein adsorption simulation studies is the necessity to conduct all-atom simulations with capability to perform conformational changes in order to capture all the relevant effects and changes during the adsorption process.

It is also apparent that most of the methods applied in the field of protein adsorption are based on classical molecular simulation approaches: Monte Carlo and molecular dynamics. Both of these techniques are, however, hampered with high computational costs and inability to locate structures of interest in a reasonable time. It is, therefore, out intention to find an alternative approach for exploration of free energy surfaces of proteins in search of their optimal conformation.

Within a multitude of methods for optimisation of complex functions, evolutionary algorithms (EA) have shown high robustness with affordable computational cost (Goldberg, 1989). Although already applied in identification of conformation of proteins in their native state (Shulze-Kremer, 1992; Le Grand and Merz, 1993, 1994), research to date has not revealed any EA based studies of conformation of adsorbed proteins. This thesis presents development and testing of an EA based method for prediction of adsorption induced conformational changes.

2.6. Solvent Models in Protein Simulations

Despite decades of development, representation of water in biomolecular systems is still active area of research. Part of the reason is the significance that water plays in biological systems. However, even more important is the difficulty encountered in water molecule modelling. Despite its apparent simplicity, the water molecule is known to be highly polarizable, which requires development of sophisticated models for its representation. Sophisticated models are, however, computationally expensive, especially if they are applied in simulations of biomolecules, which are, due to their size, often surrounded by many thousands of solvent molecules.

Methods used in molecular simulations of solutions involving proteins can broadly be divided into two categories: implicit and explicit. The required level of solvent description depends on the needs and expected outcomes of the simulation. If the object of a study is investigation of solvent restructuring, an explicit (a.k.a. discrete) model for water should be used. If, on the other hand, one is only interested in energetics of solvation process, it may be sufficient to use an implicit water model.

2.6.1. Implicit Treatment of Protein-Water Systems

Implicit methods are those that do not delve into full details of the molecular structure of the system. Although proteins may be represented using all-atom approach, solvent is treated as a continuous medium, using equations of continuum electrostatics. Depending on details of models used to represent protein and water, further classification of implicit approaches may be performed.

Electrostatic Screening Methods

Methods of electrostatic screening assume that electrostatic interactions between two charges are screened by the solvent that occupies space between them (Orozco and Luque, 2000). On a microscopic level, water molecules can be represented as dipoles. Dipoles trapped in an electric field tend to orient in a direction that reduces the strength of the field. Since two charges on a small distance from each other will create a local electric field, any dipoles found between the charges will effectively reduce the intensity of the electrostatic interaction between them compared to the electrostatic field between these two charges in vacuum. This

effect is called screening and one way to express it is using the relative dielectric constant, ε_r, which represents the ratio of intensity of electrostatic interactions between two charges in vacuum and the same charges, at the same distance, in a dielectric medium.

Higher values of ε_r correspond to stronger reduction of electrostatic interaction by the screening effect. Coulomb's law, in the presence of dielectric environment, is expressed as

$$E_{\text{es}} = \frac{1}{4\pi\varepsilon_0\varepsilon_r}\frac{q_1 q_2}{r} \tag{2.24}$$

where E_{es} is the energy of electrostatic interaction between point charges q_1 and q_2 separated by distance r, and ε_0 represents the electric permittivity of vacuum. A more consistent way to write this equation is using the dielectric constant, ε, defined as: $\varepsilon = 4\pi\varepsilon_0\varepsilon_r$. Using dielectric constant, equation (2.24) transforms to

$$E_{\text{es}} = \frac{q_1 q_2}{\varepsilon r} \tag{2.25}$$

In specific cases, where the solvent is homogeneous and solution very diluted, ε may be treated as a constant (Orozco and Luque, 2000). However, the majority of protein-water systems relevant in biochemical studies do not satisfy this condition. A straightforward way to overcome this situation is to treat the dielectric constant as a simple function of distance between charges involved in electrostatic energy calculation. The simplest function that can accomplish this is a linear dependence between the dielectric constant and the distance between point charges (Blaney et al., 1982). However, more complex functions for representation of dielectric constant have been developed, most of which describe exponential change of ε with distance (Warshel et al., 1984; Mehler and Solmajer, 1991). Despite being very simple compared to other approaches for treatment of protein-water interactions, screening methods are still in use, especially in ligand-docking simulations (Morris et al., 1998). The major advantage of screening methods is speed, but, due to their oversimplified representation of the solvent, they are not able to capture the behaviour of local elements of the system, especially where there is significant heterogeneity (Orozco and Luque, 2000).

Group Solvation Methods

Group solvation methods assume that the total solvation energy of a solute may be expressed as a sum of contributions of all of its constituent groups (Orozco and Luque, 2000). Depending on the way this contribution is calculated, methods of this group can be further divided into two subclasses: techniques based on molecular topology, and the solvent accessible surface approach. Both of them use free energy of solvation, ΔG_s, for numerical representation of the solvation process. Free energy of solvation is commonly defined using the concept of Ben-Naim (Ben-Naim, 1978), according to which solvation is described as a process of transfer of solute molecule from its gas phase into a solution at constant temperature, pressure and solvent composition. Free energy of solvation is, hence, defined as the work spent in this process (Orozco and Luque, 2000).

Approaches based on molecular topology assume that the free energy of solvation can be calculated as a sum of intrinsic properties defined for isolated solute constituents – atoms or groups (Leo et al., 1971)

$$\Delta G_s = \sum_{i=1}^{N} g_s^i \tag{2.26}$$

where g_s^i is the contribution of atom or group i to the free energy of solvation, ΔG_s. In protein terminology, solvation free energy could be calculated as a sum of contributions of individual residues. However, a serious disadvantage of this approach, especially when it is applied to molecules as complex as proteins, is the neglect of conformational changes. Intrinsic properties of constituent groups remain constant despite possible structural changes, thus making this model inapplicable for studies that investigate structure-energy relationship.

An improvement that takes solute conformation into account has been enabled through the application of solvent accessible surface methods. This group of models is based on the assumption that solvation free energy contribution of individual constituents (atoms or groups) depends on the amount of surface area of these constituents that is exposed to solvent (Chothia, 1974; Orozco and Luque, 2000). Conformational dependence is expressed through the relationship between solute conformation and area of its surface that is exposed to the solvent. In order to obtain numerical value of solvation free energy, intrinsic contributions of constituent groups

or atoms have to be multiplied by their respective surface areas exposed to the solvent

$$\Delta G_{\mathrm{s}} = \sum_{i=1}^{N} \gamma_i A_i \tag{2.27}$$

where γ_i represents solvation free energy of constituent i per unit area, while A_i represents the surface area of the same constituent exposed to the solvent.

Despite further improvements, group solvation methods suffer from several disadvantages. First of them is computational inefficiency of calculation of exposed surface area. Another drawback of these methods is related to their accuracy. The intrinsic properties of individual groups are usually calculated from solvation characteristics of small molecules and may not have the same values when the constituent is found inside a large molecule (Orozco and Luque, 2000).

Continuum Electrostatics Methods

A common feature of continuum electrostatics methods is placement of the solute molecule into the interior of a cavity formed inside the solvent (Orozco and Luque, 2000). The solvent is treated as a polarisable continuous medium with the dielectric constant ε_s. The interior of the solute cavity is characterised with a different value of dielectric constant – ε_i. All models of this class are based on the Poisson equation

$$\nabla D(\mathbf{r}) = 4\pi\rho(\mathbf{r}) \tag{2.28}$$

where $D(\mathbf{r})$ is the electric displacement at position $\mathbf{r}$ where the charge density is $\rho(\mathbf{r})$. $D(\mathbf{r})$ is defined in terms of electrostatic potential at position $\mathbf{r}$, $\Phi(\mathbf{r})$, and dielectric constant $\varepsilon(\mathbf{r})$

$$D(\mathbf{r}) = -\varepsilon(\mathbf{r})\nabla\Phi(\mathbf{r}) \tag{2.29}$$

The dielectric constant, $\varepsilon(\mathbf{r})$, changes discontinuously from the interior of the solute cavity to the bulk solvent.

Equations (2.28) and (2.29) are solved for Φ, which is then used to obtain the electrostatic contribution to the free energy of solvation

$$\Delta G_{\mathrm{es}} = \frac{1}{2}\sum_{i=1}^{N} q_i \left(\Phi_i^s - \Phi_i^g\right) \tag{2.30}$$

where superscripts *s* and *g* stand for solution and gas phase, respectively, i.e. the Poisson equation should be solved for both phases in order to calculate the free energy of solvation.

Solutions of proteins and other biomolecules represent complex systems, in which Poisson equation can be solved only numerically. Based on the algorithm used to obtain the solutions, continuum electrostatics methods may be subdivided into several classes. Most often utilized among these are finite difference approach, the Born model and the boundary element method.

Finite difference method represents domain of interest as a cubical grid (Warwicker and Watson, 1982; Honig and Nicholls, 1995; Orozco and Luque, 2000). Solute is mapped onto the grid for which Poisson equation along with necessary derivatives is solved for all the nodes. Since electrostatic potential on a grid node depends on potential on all surrounding nodes, an iterative procedure must be applied to solve the system. Although the method may be very fast, problems occur if the initial values are incorrectly guessed, in which case a system may never converge (Orozco and Luque, 2000).

The Born model is derived from analytical solution of Poisson equation. Although it cannot be solved analytically for complex systems with irregularly shaped cavities, single atoms and spherical cavities represent much simpler case in which solution to Poisson equation is obtained as a series of spherical harmonics (Orozco and Luque, 2000). If higher harmonics are neglected, the solution is expressed using equation of Born (Born, 1920a)

$$\Delta G_{\text{es}} = -\frac{1}{8\pi\varepsilon_0}\left(1-\frac{1}{\varepsilon_r}\right)\frac{q^2}{r} \tag{2.31}$$

where ε_r is, again, relative dielectric constant, and r represents the radius of the cavity in whose center charge q resides. In terms of atoms, r can be related to atomic radius, or the distance from the center of atom to the surface that water molecules cannot penetrate. For complex systems, ΔG_{es} may be obtained as a sum of contributions of individual atoms, corrected by perturbing effect of surrounding atoms (Still et al., 1990; Orozco and Luque, 2000). One of the disadvantages of Born's model is its inaccuracy when applied to complex molecules. However, this problem may be alleviated with additional parametrisation for constituent atoms.

Following the philosophy of the finite difference approach, the boundary element method also performs decomposition of domains of interest into smaller elements. However, in this case, elements do not occupy the whole volume of the system. Instead, only the surface of the solute exposed to the solvent is partitioned, usually using triangular elements (Connolly, 1983). The underlying idea of this approach is that the reaction of the solvent to the presence of solute charges can be described using distribution of charged surface elements on the solvent accessible surface of the solute (Orozco and Luque, 2000). Charges of surface triangles are calculated in a self consistent manner, using positions and intensities of solute charges as well as other charged surface triangles (Zauhar and Morgan, 1988). This method can be very fast (Vorobjev and Scheraga, 1997), but problems may occur if the surface of the solute is so complex that triangulation methods fail to partition it properly.

Summary of Implicit Methods

A general conclusion that may be drawn for all implicit approaches is that they may be very fast, but could suffer from serious disadvantages in situations where micro-effects, such as hydrogen bonds and solvent structuring between surfaces, are important. Due to their simplified representation of solvent nature, they are unable to capture heterogeneity in such cases.

2.6.2. Explicit Methods for Description of Protein-Water Systems

The main feature of this group of methods is their representation of the whole system at a molecular level. Molecular representation of the solvent enables capturing some of the phenomena that are intractable for implicit models. A well known example is hydrogen bond, which may be established between two protein atoms connected over a bridge created by specifically positioned water molecules (Beglov and Roux, 1995). Since implicit solvent methods do not recognize individual solvent molecules, they are unable to capture specific molecular orientations that may involve significant changes in free energy. Capturing of these local effects is the main advantage of explicit methods over implicit solvent representation. Most of the explicit approaches belong to one of two groups: molecular mechanics and quantum mechanics methods.

Molecular Mechanics Methods

Molecular mechanics techniques use laws of classical physics to describe interactions between the molecules. Solvent restructuring and interactions with the solute are obtained through a procedure of ensemble sampling. Two molecular simulation techniques that are predominantly exploited in the context of water configuration sampling are Monte Carlo and molecular dynamics (Allen and Tildesley, 1989; Frenkel and Smit, 1996).

Water molecules in molecular mechanics methods can be represented using models of various degree of complexity. Simple models are those that do not include effects of molecule polarisation. The nonpolarisable water models can be further divided based on the number of active sites in the molecule. The three site models, such as the simple point charge or SPC (Berendsen et al., 1981) and 3-point transferable intermolecular potential (TIP3P) model (Jorgensen et al., 1983) assign a point charge to each atom in the water molecule. These models have gained significant popularity due to the simplicity of their implementation and associated low computational costs. However, their disadvantage is low accuracy in prediction of physical properties of water. It was found that somewhat better results can be achieved with the 4-site molecular models, in which the negative charge is shifted from the center of the oxygen atom towards the hydrogen atoms (Jorgensen et al., 1983). Popular models from this group include the TIPS2 (Jorgensen, 1982) the TIP4P (Jorgensen et al., 1983). Among the most sophisticated nonpolarisable water representations are those which use 5 active centers to characterise the water molecule. These models place two negative point charges into the vertices of tetrahedron which has the oxygen atom in its center and the two hydrogen atoms in its remaining vertices. Examples of 5-site models include the Bernal-Fowler or BF model (Bernal and Fowler, 1933), Stillinger's ST2 (Stillinger and Rahman, 1974) and TIP5P model (Mahoney and Jorgensen, 2000).

Polarisable water models are able to capture molecular polarisation using approaches of dipole polarisability and fluctuating charges (Stern et al., 2001). A common practice for development of these models is to use one of the nonpolarisable models as a basis and extended it by allowing deformations of bond lengths and angles. This enables displacement of charges inside the molecule during the

simulation. Such an approach is applied in the development of the SPC/Fw model (Wu et al., 2006), which, as its name suggests, is a flexible extension of the rigid SPC model (Berendsen et al., 1981).

Quantum Mechanics Methods

These are the most rigorous of all the techniques that are used in molecular simulations. All calculations are based on equations of quantum mechanics. In general, field of application of these methods is theoretical analysis of small systems that are transformed in chemical reactions. Classical molecular mechanics methods cannot deal with chemical changes. Since systems of interest in biochemical studies are usually of significant size, pure quantum mechanical approach did not find broad application in simulations of biomolecules. However, a combination of quantum and molecular mechanics proved to be very useful in studies where interest is concentrated on a particular subdomain of the system, while the rest of it may be represented using some less accurate method. A field in which this is particularly important is the study of mechanism of enzymatic reactions (Warshel and Levitt, 1976).

2.6.3. Bridging the Gap between Implicit and Explicit Solvent Methods

Both molecular and quantum mechanics share some common characteristics that distinguish them from implicit solvent approach. They are far superior in analysis of the system structure, especially the structure of the solvent in the vicinity of the solute. Both of the methods are, however, very slow compared to implicit approach. This makes them undesirable in simulations that involve big systems and many execution steps. A good compromise in these situations is to use a method that can still provide an insight into the behaviour of solvent on a structural level, but not in such a detailed way as quantum and molecular mechanics. Langevin dipole (LD) model, a method tailored to achieve this goal, has been developed by Warshel and co-workers (Warshel and Levitt, 1976; Warshel and Russell, 1984; Florián and Warshel, 1997).

In the LD method, water molecules are modeled by dipoles fixed on a regular lattice, where strength and orientation of the dipoles are determined in a self-consistent manner under the influence of the charges on the solute atoms. It also includes all non-polar aspects of the solute-solvent interaction and hydration entropy.

Whilst the reduced water molecule representation means this method is in principle less accurate than the traditional explicit approaches, it is still able to capture heterogeneity at the molecular level but with much less computational resource. At the same time, it allows a considerably higher level of insight into structural details of the solvent than implicit methods are able to provide.

Being the method of choice in our study of proteins in water solutions and at solid-liquid interfaces, the Langevin dipole model will be explained in more details in Chapters 6 and 7.

2.7. Previous Studies of Met-enkephalin 3D Structure

Since its discovery more than 30 years ago (Hughes et al., 1975), met-enkephalin drew considerable attention of both experimentalists and molecular modelling community alike. Despite being among the smallest biologically relevant peptides (only 5 amino acid residues long), its structure in water solutions, as determined by experimental methods, has proven to be elusive (van der Spoel and Berendsen, 1997). High conformational flexibility has also shown to be a stumbling block for a range of molecular simulation methods.

2.7.1. Experimental Studies

Met-enkephalin is represented with two entries in Protein Data Bank (Berman et al., 2000): 1PLW and 1PLX (Marcotte et al., 2004). The study in which these two structures were obtained was oriented towards representing met-enkephalin molecule in an environment similar to the one in which it expresses its biological function. The primary biological role of met-enkephalin is as neurotransmitter that binds to cell membrane based opiate receptors (Hughes et al., 1975). Thus, the study by Marcotte et al. was conducted in bicelles – a model of cell membranes. This environment is, however, different than water solution and studies of other proteins show that similar nonpolar environments may promote conformational changes resulting in a 3D structure different than the one observed in aqueous solution (Losonczi et al., 2000). As we are primarily interested in proteins in solutions and at solid-fluid interfaces, this work is of less relevance here.

First experimental works devoted to elucidation of met-enkephalin 3D structure in water have emerged in the first years after the discovery of the molecule.

Roques and coauthors (Roques et al., 1976) have used proton magnetic resonance (PMR) to obtain insight into 3D structure of met-enkephalin in water and dimethyl sulfoxide (DMSO) solution. Although the results could not be interpreted with complete distinction, the authors conclude that the most probable conformation in both solutions is characterised with hydrogen bond between CO of the first glycine residue and NH-group of methionine, with high mobility of N-term tyrosine residue. Using the same methodology, Jones and his group have reached analogous conclusions (Jones et al., 1976). However, PMR and ^{13}C NMR (CMR) studies conducted in very similar environmental conditions (Bleich et al., 1976) have provided results that can be interpreted through the lack of intramolecular hydrogen bonding and possible interactions between tyrosine side chain and solvent molecules.

Khaled and co-workers (Khaled et al., 1977) have conducted an extensive study of both met- and leu-enkephalin in a range of solvents using several different experimental methods: PMR, CMR, ultraviolet (UV) and circular dichroism (CD) spectroscopies. In an effort to shed some light on discrepancies in met-enkephalin conformation encountered by previous research groups, they have discovered effects of temperature and enkephalin concentration on its 3D structure. The proposed conformation of met-enkephalin in diluted solutions, i.e. in its monomeric form, is similar to the structure derived by groups of Roques and Jones, with β-turn between second glycine and phenylalanine residues, and hydrogen bond between first glycine and methionine. An additional H-bond is also speculated between OH-group in tyrosine side chain and NH of the second glycine residue.

In another study, Jones and co-workers (Jones et al., 1977) have attributed previously observed discrepancies in PMR spectra of met-enkephalin to existence of two forms – cationic and zwitterionic. The conformation previously described by the same group (Jones et al., 1976) has been assigned to zwitterionic form, while structures similar to that produced by Bleich and co-workers (Bleich et al., 1976) correspond to met-enkephalin cation. However, an even more important discovery is the very high conformational flexibility of met-enkephalin, resulting in a group of structures with similar probability of occurrence. A similar conclusion with regards to structural flexibility has been reached in an independent CD study conducted in a wide range of temperatures and pH values leading to different ionised forms (Spirtes

et al., 1978). The notion of conformational flexibility is further supported by investigation conducted by Graham and coauthors (Graham et al., 1992). While rigidity of met-enkephalin was increased in the presence of sodium dodecyl sulphate (SDS) micelles, the absence of micelles in aqueous solutions increases the number of detected structures to as many as 20 different conformations obtained by combining experimental data with molecular modelling. Structural flexibility of met-enkephalin is understood to be a consequence of its low molecular mass and the presence of two consecutive glycine residues (Spirtes et al., 1978), which are known to be very mobile due to lack of side chain groups (Rappé and Casewit, 1997).

2.7.2. Molecular Simulations

The inability of experimental techniques to discern a single stable met-enkephalin 3D structure in aqueous solutions has motivated many researchers to approach the problem using molecular models. The first attempt to utilise molecular simulations in elucidating met-enkephalin conformation (Isogai et al., 1977) occurred soon after the molecule's discovery and was based on application of the ECEPP potential energy model (Momany et al., 1975). This study is, however, important only for historical reasons as computational resources of the time were prohibitive for solvent representation and application of the method was limited to met-enkephalin in vacuum. One of the first molecular simulations of met-enkephalin in solutions was based on the application of the ECEPP/2 potential energy model (Nemethy et al., 1983; Sippl et al., 1984) for solute atomic parameters, while implicit representation has been used for water (Li and Scheraga, 1987). Results of Monte Carlo energy minimisation of the system were in accordance with experimental findings and suggest existence of an ensemble of stable unfolded conformations in water, contrasting a single dominant structure in which simulations in the absence of water were resulting. Another study involving continuous solvent with a similar outcome (Koča and Carlsen, 1995) further confirmed the notion of met-enkephalin flexibility in aqueous solution by producing over 500 different structures within a span of only 4 kcal/mol.

Advances in computational power achieved in recent years have allowed utilisation of finer grained solvent representations. Studies in which reference interaction site model (RISM) theory (Chandler and Andersen, 1972) have been

applied for determining the solvated structure of met-enkephalin have yielded results that were in general accordance with experimental studies and implicit solvent simulations and which favour conformational flexibility and unfolding of the solute molecule (Kinoshita et al., 1997, 1998). The latter work, through the combination of Monte Carlo energy minimisation and RISM based calculation of interactions with solvent, results in a set of almost fully extended structures of similar energies, characterised with large fluctuations in side chain conformations. Although the study used an un-ionised molecule for all energy calculations, the authors stress that their first study has shown remarkable similarities in solvation behaviour of the un-ionised and zwitterionic met-enkephalin form. Similar results, with low energy barriers for transition between various conformations, have been confirmed in molecular dynamics studies with explicit representation of water (Sanbonmatsu and García, 2002).

2.7.3. Summary of Met-enkephalin Structure Determination Studies

A general conclusion derived from molecular simulation studies, as well as from experimental approach in investigation of met-enkephalin structure, is that its conformation in aqueous solution is very flexible, both in unionised and in zwitterionic form. Whilst there is no general agreement with respect to the most stable conformation of the molecule in the solution, most of the studies, both experimental and modelling, have observed a tendency of met-enkephalin to unfold and extend its backbone in the presence of water molecules, thus exposing its atoms to the solvent.

Chapter 3. Methods

3.1. Introduction

A generally accepted dogma in the field of protein 3D structure studies is the Anfinsen's thermodynamic hypothesis, according to which the native 3D structure of a protein is the one in which the free energy of the protein and its surrounding environment (e.g. water solution) is at the minimum (Anfinsen, 1973). The free energy of a protein is a sum of its potential energy (PE) and the entropic contribution. The entropic contribution is, however, often neglected, thus approximating the free energy of the protein with its intramolecular potential energy.

Further simplification commonly adopted in protein folding studies involves fixing the bond lengths and angles between chemical bonds (details of the protein structure are provided in Appendix A) at values that provide minimal potential energy for individual amino acids. Such an approach has been used in all versions of the ECEPP PE model (Momany et al., 1975; Nemethy et al., 1983; Nemethy et al., 1992; Arnautova et al., 2006). This assumption has also been used throughout this work.

With the bond lengths and angles fixed, the only way to change the 3D structure of a molecule is by performing torsions around chemical bonds (explained in greater detail in Appendix A). Accordingly, the potential energy of a molecule can be expressed as a function of all the dihedral angles. This multidimensional function is usually called the PE surface, or, due to approximation of the free energy of a protein with its potential energy, free energy (FE) surface. Since peptide bonds are assumed to be perfectly planar and fixed in *trans*-conformation (Mizushima et al., 1950; Kitano et al., 1973; Kitano and Kuchitsu, 1973), potential energy of a protein is now a function of its ϕ and ψ backbone dihedral angles and χ_1 to χ_N side chain

dihedrals. The number of side chain dihedral angles, *N*, depends on the amino acid residue and ranges from 0 for glycine (since it has no side chain whatsoever) to 7 for arginine.

The potential energy of a protein associated with each set of dihedral angles is usually calculated using one of the empirical PE models (also known as *force fields*, Appendix B). The free energy surface will, therefore, be a function of a protein (i.e. its primary structure) and the PE model used to calculate its potential energy. Applying Anfinsen's hypothesis, in order to determine the 3D structure of a protein, all that is needed is the protein's primary structure, the PE model and the numerical method for minimisation of the FE surface. Methods that are able to obtain the 3D structure of a protein based only on its primary structure and the PE model are known as *ab initio* methods for protein 3D structure prediction. Our goal is the development of a novel *ab initio* method for prediction of structure of proteins in native and non-native environments.

3.2. Free Energy Surface Exploration

Free energy surfaces of proteins are, generally, very complex, multidimensional and multimodal functions. An example of a free energy surface, given as a function of a single pair of backbone dihedral angles is illustrated in Figure 3.1. In reality, however, this function will be even more complex as it will depend on more dihedral angles.

Two techniques that have found the most widespread use in exploration of the free energy landscape are Monte Carlo (MC) and molecular dynamics (MD) (Allen and Tildesley, 1989; Frenkel and Smit, 1996), discussed in detail in Chapte 2. Although algorithmic improvements to the basic MC and MD implementations have allowed them to operate successfully with rough free energy landscapes, both methods tend to spend significant amount of time in configurational sampling. MD simulations, for example, usually simulate molecules in the time span of nanoseconds or even longer, whilst the time step is measured in femtoseconds (Bolhuis, 2003; Karplus and Kuriyan, 2005). It is, therefore, often necessary to performs millions of energy calculations in order to obtain reliable results through MC and MD approaches. This has led many researchers to apply other free energy surface exploration methods in prediction of protein 3D structure. A method that

gained popularity in protein folding field is genetic algorithm (Goldberg, 1989; Mitchell, 1996). Genetic algorithm itself belongs to a broader group of methods based on principles of natural selection and survival of the fittest – evolutionary algorithms (EA).

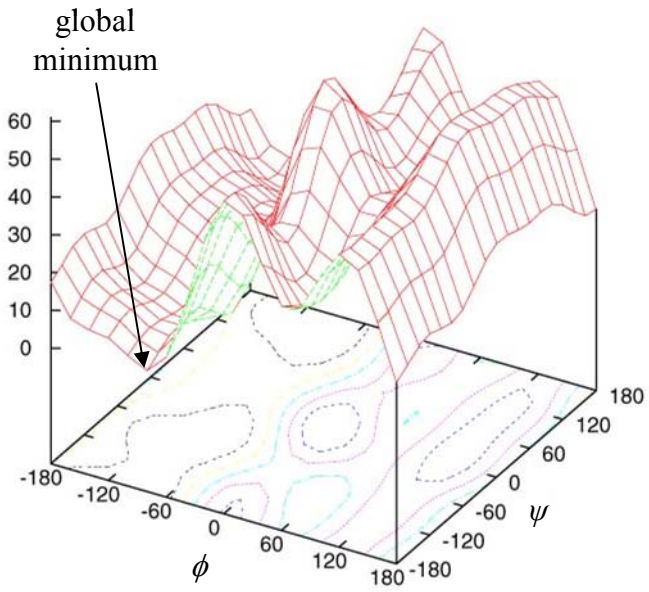


Figure 3.1 An example of the free energy surface of a peptide as a 2D function of ϕ and ψ backbone dihedral angles in a single residue. All the other dihedral angles are kept fixed.

3.3. Evolutionary Algorithms in Protein Folding Prediction

The main advantage of evolutionary algorithms in prediction of 3D structure of proteins is their superiority in handling barriers in the free energy landscape. Unlike MC and MD methods, which rely on incremental exploration of the free energy function, EA based approach is capable of performing jumps out of local minima, irrespective of the size of surrounding energy barriers. This capacity of EA based methods is embedded in the design of the algorithm itself.

Evolutionary algorithms are global optimisation methods based on the mechanisms of natural genetics and natural selection (Goldberg, 1989). Natural evolution is a constant process in which a species' survival capabilities are perfected from generation to generation. The main driving force for this improvement is the *survival of the fittest*. When two individuals mate and form offspring, the chances of the offspring survival are increased if its genetic material equips it with greater fitness. For predatory species, it could be better eye sight or greater speed. Prey

species would, for instance, benefit from better sense of hearing. On the other hand, "bad" genetic material will deteriorate the offspring chances of survival and reaching reproductive period, thus diminishing the probability of passing unfit genes into the new generation. Consequently, as generations progress, the species constantly evolves and improves its average fitness.

Evolutionary algorithms are driven by the same principle – survival of the fittest members and disposing of those less fit. Function whose global optimum is searched for is used as the fitness in EA methods. In protein structure prediction, this is the free energy. Since, by Anfinsen's thermodynamic hypothesis (Anfinsen, 1973), native structure of a protein is the one with the minimal free energy, the structures that have lower free energy will have higher fitness and vice versa.

The essential details of the design of an EA based method used in this thesis are adopted from the study of proteins in their native conformation by Djurdjevic (Djurdjević, 2006). A simplified flow diagram of the algorithm is shown in Figure 3.2. Basic elements of the method are described in greater details below.

3.3.1. Population

The EA based determination of protein 3D structure operates on a set of protein conformations that form a population. Each individual conformation is called a member of the population. All genetic operators (such as crossovers and mutations, described below) are performed with individual members. The most important numerical characteristic of a population is its size, N_P. All our studies have been performed with populations of between 100 and 600 members.

As the evolutionary algorithm progresses, its population changes, thus increasing the average fitness of its members. The evolution of the population is governed by the principles of natural evolution. However, the creation of the first set of members is always performed randomly in protein *ab initio* structure prediction.

3.3.2. Genes, Chromosomes and Population Members

Ability of a biological organism to survive depends on its fitness, which is, in turn, encoded in its genetic material – chromosomes and genes. From the perspective of natural evolution, genetic material is the only thing that defines an individual

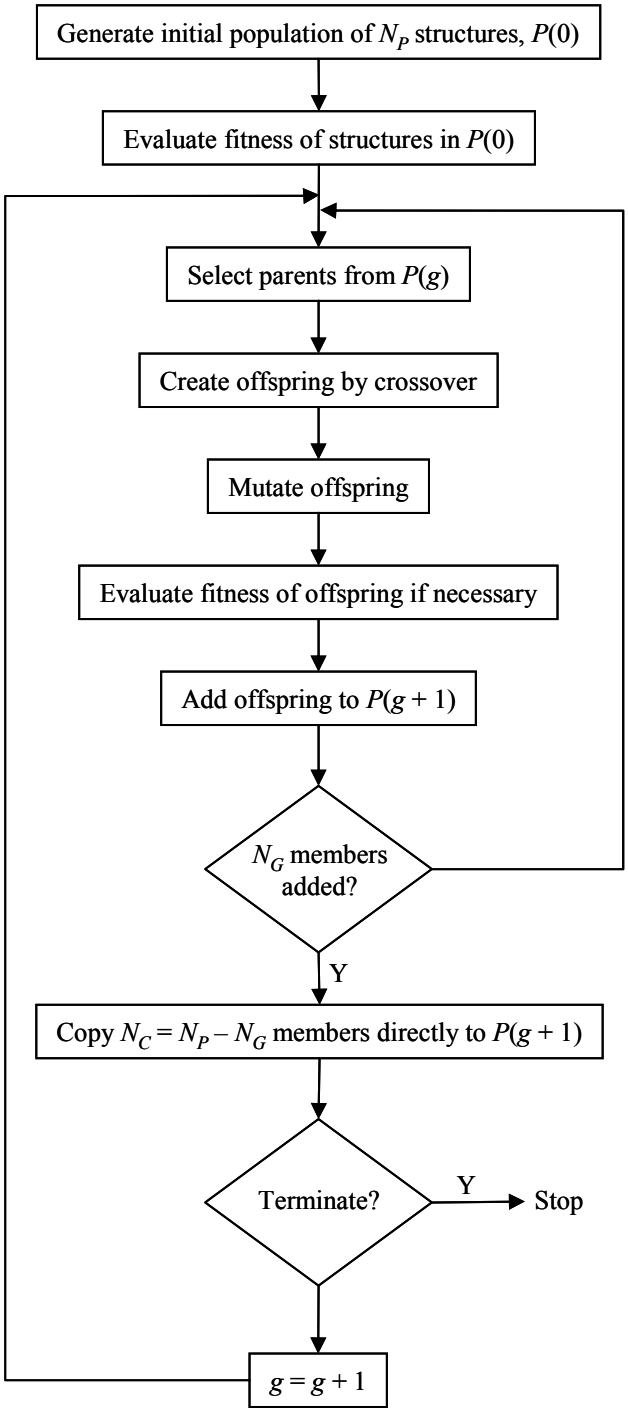


Figure 3.2 Simplified flow diagram of the evolutionary algorithm implementation used throughout the study.

organism. Analogous reasoning can be applied in evolutionary algorithms, i.e. a population member is completely described by its associated "chromosome".

Natural evolution is conducted through genetic manipulation of chromosomes, which can be thought of as strings of genes. In an analogous way, evolutionary algorithm methods create chromosomes as strings of variables that the fitness function depends on. In the protein folding studies, a gene corresponds to a value of a single dihedral angle, whilst chromosome is expressed as a set of all dihedral angles that determine protein 3D structure. All genetic operations are performed on this string of dihedral angles.

3.3.3. Gene Encoding

Mapping between the set of dihedral angles and the chromosome string in EA methods is not always straightforward. The majority of the EA based methods perform this mapping using binary encoding (Goldberg, 1989). Binary encoding uses binary numbers to store values of dihedral angles in a chromosome string. The precision of binary numbers is, however, usually low (including less than 10 binary digits), which has significant repercussions on the EA performance if the fitness function is a function of variables whose values belong to the set of real numbers. Such is the case with protein 3D structure prediction. Dihedral angles can take any real value between 0 and 360^{o}. It has been shown that the EA performance in protein 3D structure prediction in a gas phase is much higher if the real encoding is used instead of binary (Djurdjevic and Biggs, 2006). The real encoding will, hence, be applied in this study.

3.3.4. Member Fitness

Calculation of the fitness (the second step of the algorithm in Figure 3.2), as discussed above, is performed using the PE model chosen for the study. It is conducted for each population member in the first generation and for every newly created member in latter generations.

The functional relationship between a chromosome and its associated fitness is resolved by first decoding genes into a set of dihedral angles. The dihedral angles are then used to obtain Cartesian coordinates of all the atoms in the protein. Atomic

coordinates can then be used to calculate all the terms in the intramolecular energy sum.

It should be noted that the fitness calculation will be discussed in more detail in later chapters. Our work substantially relies on development of new methods, for which the fitness function is not only the intramolecular PE of a protein, but includes other terms, such as free energy of solvation, or interaction between the protein and a solid surface.

The inner steps of the algorithm shown in Figure 3.2 are completely governed by the set of evolutionary operators: *selection for reproduction*, *crossover*, and *mutation* (Goldberg, 1989). When applied on a population of structures of an arbitrary size, these operators direct the population towards the member with the highest fitness, i.e. the 3D structure with the lowest free energy or native conformation.

3.3.5. Selection for Reproduction in Evolutionary Algorithms

Reproduction is the main process for transfer of fit genes from the old to the new generation. Since its purpose is to propagate good genetic material, it is tightly related to the process of selection of fit members of a population. Selection of a member whose genetic material will be passed to the new generation is, hence, based on its fitness – probability to be selected increases with the fitness. The relationship between fitness of a member and probability of transfer of its genes to the new generation depends on specific algorithmic implementation. A method that is often used in evolutionary algorithms is the roulette wheel selection (Goldberg, 1989). It has been shown, however, that in protein 3D structure prediction, tournament (Goldberg et al., 1989) and uniform selection (Schwefel, 1981; Bäck and Hoffmeister, 1991) are both far superior (Djurdjević, 2006). These two approaches have been used throughout this work.

Tournament selection is conducted by randomly choosing several members of a population and forming a subpopulation of these members. The member with the highest fitness in the subpopulation is then used in the following crossover step of the algorithm. Since the crossover is performed between two members, the other member of the population is selected using the same tournament procedure. It is clear

that the higher the fitness of members, they will have more chances of "winning the tournament" within the subpopulation.

Whilst tournament selection creates subpopulation by random choice and then performs fitness ordering within the subpopulation, uniform selection takes the opposite approach – it first orders all the members of the population according to their fitness and then chooses randomly a member to be passed to the new generation from the fraction of the population formed by the fittest members. The fraction of the population considered for the selection is designated as the *truncation selection parameter*. Higher values of this parameter correspond to low selection pressures, while very low values indicate that the selection pressure is very high as only the fittest members are allowed to pass to the new generation (Djurdjević, 2006).

3.3.6. Crossover Operator

Reproduction is responsible for passing good genetic material to new generations. However, if the genes were passed without any modifications, the new generations would quickly become saturated with the optimal member from the first generation, without any possibility of further improvement in fitness. Crossover is one of the methods for creating qualitatively new population members. As such, it is a means of fitness landscape exploration. Crossover creates new population members, with new sets of strings, which correspond to unexplored points on the fitness function surface.

Crossover operator couples two "parent" chromosomes (i.e. chromosomes from the previous generation) and mixes their genes in order to create two "child" chromosomes. Depending on the type of gene mixing, there are three basic crossover implementations: *single point*, *multipoint* and *uniform crossover*. (Haupt and Haupt, 1998).

In a single point crossover, a location between the two consecutive genes is chosen randomly and parents swap the parts of the chromosome after the chosen location. Thus, one child will have the starting sequence of genes from the first parent, whilst the ending sequence will be identical to the ending sequence of the second parent. The other child will have a complementary distribution of genes – its starting sequence will be identical to the starting sequence of the second parent, whilst its ending genes will match that of the first parent.

Multipoint crossover is merely a generalisation of a crossover with single point. Whilst a single point divides chromosome in two segments, crossover with N_X points will create $N_X + 1$ segments on both parent chromosomes. Parents then swap genes of every even-numbered segment. Thus, the first child will have the first segment of genes identical to the first parent, but its second segment will come from the second parent. The third segment will, again, correspond to the first parent and so on. The other child will have a complementary set of genes, as for the single point crossover.

Unlike single and multipoint crossover strategies that operate on whole segments of chromosomes, uniform crossover swaps individual genes of two parents (Haupt and Haupt, 1998). Chromosomes of both parents are scanned and each gene of each parent is randomly copied to the chromosome of the first or the second child.

The premise of crossover is the idea that fitness is carried by individual genes and that, in order to achieve maximal fitness, the optimal combination of genes should be established in a single individual. Individuals with high fitness, hence, carry some of the good genes, but not necessarily their optimal combination. In the early generations of evolutionary algorithm, good genes are often mixed with bad ones in individual members. Thus, choosing two fit members from an early generation will increase the overall number of good genes as each of the chosen individuals carries their own set of them. Crossing the chromosomes over creates a child that will potentially include both of the sets of good genes, and, therefore, be even more fit than its parents. Obviously, due to complementary set of genes between the children, the other child produced from the same crossover operation will have lower performance, but it does not deteriorate the overall performance of the EA method as the unfit offspring will quickly be replaced with more fit members of the population.

Previous work has shown that multipoint crossover shows advantageous performance in protein 3D structure prediction (Djurdjevic and Biggs, 2006) and is, thus, used in this work.

3.3.7. Genetic Mutation

In addition to crossover, mutation is another operator that facilitates exploration of fitness landscape and increases population diversity. However, whilst crossover produces new population members based on genes inherited from previous generations, mutation may introduce completely novel genetic information. It is performed by randomly changing individual genes to a new value that can be any number from the allowed range. For protein dihedral angles, this is the range of values between 0 and 360^o. Mutation is implemented by scanning all the genes in a chromosome and changing each of them with the prespecified probability of mutation, P_M.

As a means for increasing diversity, mutation is used as a measure for prevention of premature convergence of a population to a point of locally maximal fitness. Although crossover can also be used to increase diversity of the population, it has the highest potential to do so in the early stages of the algorithm, while individual members are to a significant degree genetically different from each other. Progress of the algorithm, however, enriches new generations with earlier fit members. After a number of generations, there is a possibility that the population is saturated with a single chromosome, that may not correspond to globally maximal fitness. In such cases, crossover will operate on two members whose chromosomes are genetically identical. Thus, both children will be the exact genetic replicas of their parents. Only mutation can introduce new genes into such a population and enable the algorithm to jump out of the local fitness maximum.

3.3.8. Steady-State EA

The flow diagram in Figure 3.2 shows that the genetic mutation of the offspring is followed by evaluation of their fitness (described above) and adding the newly created members to the new generation. The number of the new members in a new generation defines the type of an evolutionary algorithm. The type used in all our studies is the *steady-state evolutionary algorithm* (De Jong, 1975; Holland, 1975; Mitchell, 1996).

Steady-state evolutionary algorithm is characterised with specific strategy for replacement of old population members with the offspring. Whilst traditional genetic algorithm implementation (also known as *generational genetic algorithm*) performs

replacement of the whole population with the offspring, steady-state algorithm is based on replacement of limited fraction of the population with the new members (Mitchell, 1996). The number of old members that are being replaced is called *generation gap*, N_G (De Jong, 1975).

Thus, in the algorithm used in our studies (Figure 3.2), only N_G members of the offspring are added to the new generation. The rest of the generation, its $N_C = N_P - N_G$ members, is filled by the part of the population copied from the previous generation.

Copying only N_C members from the previous generation to the new one means that N_G members of the previous generation have been discarded, or replaced with the new members. Replacement of the old population members in our studies has been conducted using *the exponential replacement strategy* (Syswerda, 1991). Exponential replacement is performed by first ranking all of the population members according to their fitness and then, starting from the least fit and moving upwards, testing if the member should be replaced by generating a random number whose value should be lower than prespecified replacement probability in order for replacement to occur. This strategy leaves a possibility for the least fit members to survive for several generations, which is desirable as, although they are overall unfit, they may possess some unexpressed genetic quality that would be lost with their replacement.

3.3.9. Convergence Criterion

After completion of each new generation, the algorithm performs a check whether the population has converged to the optimum. It is assumed that convergence is achieved if the fittest member of the population does not change for 5000 generations. If a single member is the fittest for such a long period, there is a high probability that it has copied its genes throughout the whole population (i.e. the whole population has the same or very similar chromosomes), thus reducing its genetic diversity. The reduction of genetic diversity, even if achieved in a local optimum, leaves poor chances of ever finding better solution and the algorithm is, hence, terminated. If the convergence has not been achieved, the algorithm continues

execution by selecting parents for the offspring of the new generation, as shown in the outer loop of Figure 3.2.

3.4. Other Numerical Elements Used in the Study

3.4.1. Local Minimisation of the Fittest Member

Whilst being praised for its robustness, evolutionary algorithm is known to suffer from lower level of accuracy. In the context of protein 3D structure prediction, it can predict a structure that is close to the global minimum of free energy, but never at the exact point of global minimum. It is, therefore, necessary to couple EA with a local minimisation method, which, when the conformation is in the right region (close to the FE minimum), performs better than EA in pinpointing the exact minimum position. The local minimisation method used in our studies is the same as the one used by Djurdjevic (2006) – Broyden-Fletcher-Goldfarb-Shanno algorithm or BFGS (Press et al., 1992). BFGS is a gradient minimisation method and, although expensive, it is applied on a single optimal conformation from the EA run, thus significantly improving the accuracy of the method for only a fraction of the overall computational time.

3.4.2. Evaluation of the Quality of Structure Prediction

Where applicable, the 3D structures predicted by an EA based approach were compared to the already known structures that correspond to the global minimum on the FE surface. As in the study of Djurdjevic (2006) the comparison is performed using root mean square difference (RMSD) between the recovered and the expected structure. RMSD between the two conformations ("1" and "2") of the same molecule is calculated using the following equation

$$RMSD = \sqrt{\frac{\sum_{i=1}^{N}\left(x_{1,i}-x_{2,i}\right)^2+\left(y_{1,i}-y_{2,i}\right)^2+\left(z_{1,i}-z_{2,i}\right)^2}{N}} \tag{3.1}$$

where x, y, and z are atomic coordinates and N is the number of atoms in the molecule used for the comparison. In many instances, not all of the atoms of a protein were used in RMSD calculation. It is a common practice in protein folding studies to use only positions of α-carbons in this calculation (Djurdjević, 2006).

Chapter 4. EA Performance for Common Potential Energy Models

4.1. Introduction

According to Anfinsen's hypothesis (Anfinsen, 1973), the native conformation of a protein corresponds to the minimum of free energy of the protein and its surrounding environment (such as water solution). For proteins in a gas phase or in vacuum, however, physical properties of the environment are not considered, while the entropic contribution to the free energy of the protein may be neglected. The determination of the most stable 3D structure of a protein in a gas phase is, therefore, equivalent to search for the global minimum of its potential energy. Whilst potential energy (PE) surface of various proteins in a gas phase has been explored by various methods, including evolutionary algorithms (Shulze-Kremer, 1992; Le Grand and Merz, 1993, 1994; Djurdjević, 2006), there is still little understanding of the details that influence EA performance and its functional relationship with the EA design and control parameters. In particular, no study has previously addressed the effect of the choice of the EA fitness function (i.e. the PE model) to the performance and optimisation of evolutionary algorithms. This chapter describes our efforts to address this issue.

The influence of the fitness function choice has been analysed using met-enkephalin molecule (Hughes et al., 1975) in the gas phase. Four common PE models have been used to describe the fitness function. The chapter first describes the details of the system used in the study. System description is followed by representation of the major findings and analysis of the results. Major findings are summarised in the conclusion.

4.1. Study Details

4.1.1. Overview of the Study

The primary aim of the study was to elucidate how the performance of an EA in the *ab initio* protein fold prediction context and the optimal control parameters are influenced by the potential energy (PE) model used. This was achieved by determining how the performance of an EA varied with the control parameters when applied to a small peptide using four different PE models. The performance characteristics were determined through a full sweep of the control parameter space.

A number of criteria were used to select the PE models considered. We felt it was important to consider at least some of the more popular PE models. This requirement was, however, tempered against the desire to study PE models with different functional forms as well as models that primarily differed in their parameters to determine if this would substantially influence performance.

A secondary aim was to elucidate how EA performance characteristics are influenced by the degree of accuracy demanded for the fold prediction. This was done by considering how the performance of the EA varied with the fold accuracy when using the Amber PE model. It should be stressed that it was *not* the intention of this study to identify optimal control parameters for the *ab initio* protein fold prediction problems – as the previous work in our group hinted at (Djurdjevic and Biggs, 2006) and this study confirms, these parameters are a strong function of the nature of the search space, which is affected by not only the PE model, but also the peptide and its representation.

4.1.2. Evolutionary Algorithm

The EA used throughout the work reported here was based on the SRM design described earlier (Djurdjevic and Biggs, 2006) with one exception: tournament selection was replaced by truncation selection (Schwefel, 1981; Bäck and Hoffmeister, 1991) with exponential ranking (Hancock, 1994). We have found this design, which is based on steady-state replacement with elitism, real encoding and multipoint crossover, to be generally superior to other designs we have considered in the *ab initio* protein fold prediction context.

4.1.3. Representation and Encoding of the Peptide

Met-enkephalin, a natural endogenous opioid (Clement-Jones et al., 1982; Spadaccini and Temussi, 2001) consisting of 5 residues as illustrated in Figure 4.1, was considered in the gas phase with the N- and C-terms capped by acetyl and methyl-amide groups respectively. Although this peptide is relatively small compared to many natural proteins, it is ideal here as its heterogeneity and flexibility (which arises from the presence of the two glycine residues in the middle) makes it non-trivial to determine its fold, yet its size allows adequate statistics to be obtained in reasonable (although still considerable) computational resource. It is for these reasons that the peptide has been widely studied in the protein fold prediction context (Kawai et al., 1989; Ripoll and Scheraga, 1989; Nayeem et al., 1991; Olszewski et al., 1992; Le Grand and Merz, 1993; Androulakis et al., 1997; Lee et al., 1997; Jin et al., 1999; Klepeis and Floudas, 1999; Vengadesan and Gautham, 2004). Whilst it may be argued that, as a flexible molecule, met-enkephalin is not suitable for the application of Anfinsen's hypothesis (Anfinsen, 1973), we stress that our goal here is not to retrieve all experimentally determined 3D structures, but to test the EA performance, for which reason a highly flexible molecule with rough free energy landscape is very desirable.

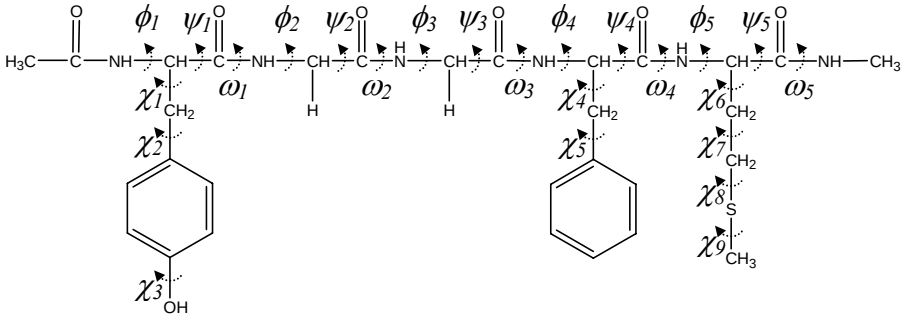


Figure 4.1 Structural formula of met-enkephalin molecule analysed in this study.

The peptide was modeled entirely at an atomistic level. In common with many of the previous *ab initio* studies, only the main backbone, ϕ_i and ψ_i, and side chain dihedral angles, χ_i, were varied during the course of the simulation. These 19 dihedral angles were encoded as real numbers in a linear chromosome. Initial values of the angles were sampled from a uniform distribution spanning the range [0°, 360°), and they were allowed to vary over this range during a simulation. As per

standard practice in the field, the dihedral angles about the peptide bond were fixed at $\omega_i = 180^\circ$ (Pauling, 1940; Mizushima et al., 1950; Kitano et al., 1973) throughout the simulations, whilst all the bond lengths and angles were similarly fixed at values determined by an exhaustive search for the global minimum in the PE surface of each of the residues when capped at the N- and C-terms by acetyl and methyl-amide groups respectively.

4.1.4. Potential Energy Models

Potential energy of a molecule and its relationship with the molecule's 3D structure is captured by the potential energy (PE) models (discussed in greater detail in Appendix B). Potential energy models are sums of various terms that define contribution to the total energy of the molecule associated with specific intramolecular interactions. The functional form of these terms is one of the primary sources of difference between the various PE models available in the open literature. The second major source of variation between PE models is the set of model parameters.

Four PE models were considered in detail here: Amber94 (Cornell et al., 1995), OPLS (Jorgensen et al., 1996), CVFF (Dauber-Osguthorpe et al., 1988) and ECEPP/3 (Nemethy et al., 1992). The details of these PE models are summarised in Table B.1 in Appendix B. The Amber model is typical of many of the biomolecular PE models in that it seeks to capture the major sources of PE variation without excessive complexity. For example, it adopts the most basic forms of the bonded interactions, and omits explicit mention of hydrogen bonds, which are instead accounted for implicitly. The OPLS model, which is more modern, is largely based on Amber but has different parameter values (Jorgensen et al., 1996). The ECEPP model, which is perhaps one of the most popular PE models, is even simpler than Amber or OPLS in that it includes no bond length or angle terms, although it does include hydrogen bonds explicitly.

The parameter sets used for each PE model were taken from original sources. Our implementations of Amber and OPLS models were assessed by comparing energies produced by our code against those obtained from TINKER (Ponder, 2004). CVFF implementation was tested against original force field source, whilst ECEPP

results have been compared with those produced by ECEPPAK package (Ripoll et al., 1995).

4.1.5. Parameter Ranges

The EA used here is controlled by a total of nine parameters: the mutation probability, P_M, crossover probability, P_X, number of crossover points, N_X, population size, N_P, number of the best-to-worst rank ordered members of the population used in selection, $\lambda = \alpha N_P$ for $\alpha \in (0,1]$, the number of population members retained per generation, N_C, exponential replacement factor, s, the relative PE change of the best fold in a generation below which no change is considered to have occurred, ε_{PE}, and the number of generations of no change in the PE of the best fold required to trigger termination of a simulation, N_T; the values of these parameters are shown in Table 4.1.

Table 4.1 Control parameters considered in the study and their numerical values

Parameter	Values considered
P_M	0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0
P_X	0.0, 0.1, 0.3, 0.5, 0.7, 0.9, 1.0
N_X	2, 4, 8
N_P	100, 200, 300, 400, 500, 600
α	0.1, 0.5, 0.9
N_C	$N_P - 1$
s	0.1
ε_{PE}	10^{-4}
N_T	5000

Previous work in our group (Djurdjevic et al.; Djurdjevic and Biggs, 2006) suggests that the mutation probability has the most profound influence on performance, whilst the crossover probability, number of crossover points, truncation fraction, α, and population size also influence performance, albeit to a lesser degree. We, therefore, considered a range of values for each of these parameters. The remaining parameters were fixed at values that our experience suggests are satisfactory.

4.1.6. Performance Measures

When presenting the performance of an EA, many use the average performance of N_R separate simulations under identical conditions (i.e. N_R realisations)

$$\bar{F} = \sum_{i=1}^{N_R} \frac{F_i}{N_R} \tag{4.1}$$

where F_i is a performance measure for a simulation such as, for example, the number of PE function evaluations required. The average performance is not, however, always a good measure of the performance. For example, some control parameter settings led to premature termination (i.e. before the global minimum is reached) in a small number of function evaluations because of insufficient chromosomal disruption. Clearly the average number of function evaluations is not a good indicator of performance in such cases. Success rates approaching 100% are also possible, but often at the expense of long simulations. It is clear that there must be some balance between both success and number of function evaluations when assessing performance. We have, therefore, used here the number of potential function evaluations required to be 99% sure that the fold is correct (Djurdjevic and Biggs, 2006)

$$F^{(99)} = \frac{-2\bar{F}}{\log(1-S)} \tag{4.2}$$

where S is the fraction of the N_R realisations that are deemed successful. A realisation is judged successful if the root mean square difference (RMSD) between the best fold obtained from the realisation and the "correct" fold is less than some threshold, ε_{RMSD}. For the vast majority of the work reported here, $\varepsilon_{RMSD} = 1\text{Å}$ is used which is less than half the average RMSD obtained if the met-enkephalin structures were randomly generated (Djurdjevic et al.) and a value considered satisfactory for most biologically relevant work (Baker and Sali, 2001). The last part of the work reported here does, however, consider the effect that this parameter has on the performance characteristics of the Amber PE model.

Table 4.2 The number of observed "correct" folds for each of the PE models studied[a]

"Correct" fold	Number observed	Fraction of total
… for Amber	1288	0.114 %
… for OPLS	2151	0.231 %
… for CVFF	676	0.423 %
… for ECEPP	6	0.00037%

a. Folds were considered to be the same as the 'correct' fold provided their PE was within 0.5 kcal/mol of the 'correct' fold and the RMSD between the two folds was no more than 0.1 Å.

We have assumed here that the "correct" fold for a PE model corresponds to the lowest energy fold obtained from all the simulations done for the PE model. As Table 4.2 shows that these folds were observed multiple times, it is reasonable to assume they are the fold associated with the global PE minimum, although this is not essential for our purposes here.

The RMSD in Table 4.3 indicates that the backbone of the "correct" folds for the various PE models differ from each other. The PE energy of the various "correct" folds obtained using the other PE models are, however, always higher (see Table 4.3), suggesting that the different folds have some basis in fact. Indeed, such differences are not unsurprising given the disparities in the functional forms of the PE models and their parameter values.

Table 4.3 The RMSD and the PE of the "correct" fold for the PE models relative to the "correct" fold of the other PE models

PE model		"Correct" fold values for PE model below relative to values of "correct" fold for PE models left.			
		Amber	OPLS	CVFF	ECEPP
Amber	RMSD (Å)	0.0	-	-	-
	ΔU_t (kcal/mol)[a]	0.0	11.44	5.97	24.96
OPLS	RMSD (Å)	1.29	0.0	-	-
	ΔU_t (kcal/mol)[a]	1.36	0.0	8.20	35.39
CVFF	RMSD (Å)	0.78	0.69	0.0	-
	ΔU_t (kcal/mol)[a]	44.04	43.47	0.0	107.11
ECEPP	RMSD (Å)	2.38	2.14	2.42	0.0
	ΔU_t (kcal/mol)[a]	14.03	13.35	96.17	0.0

a. $\Delta U_t = U_t$ ("correct" fold in PE model 1) – U_t ("correct" fold in PE model 2)

Despite these quantitative differences, Figure 4.2 shows that the "correct" folds for the PE models are not dissimilar in some respects – all contain bends with hydrogen bonds on either side (i.e. they could be described as β-bends). The predicted folds are, therefore, broadly consistent with those obtained by others (Ripoll and Scheraga, 1989; Androulakis et al., 1997), although direct comparison is impossible because of differences in the caps and PE models.

4.2. Results and Discussion

4.2.1. Influence of N_R on Accuracy of Performance Measure

The uncertainty of the $F^{(99)}$ data will clearly tend to decrease as the number of realizations, N_R, increases. Figure 4.3 indicates that this is indeed the case, with the $F^{(99)}$ value in a better performing part of the control parameter space tending to remain roughly constant beyond $N_R \approx 500$ for all the PE models, whilst the standard deviation tends to stabilize beyond $N_R \approx 900$.

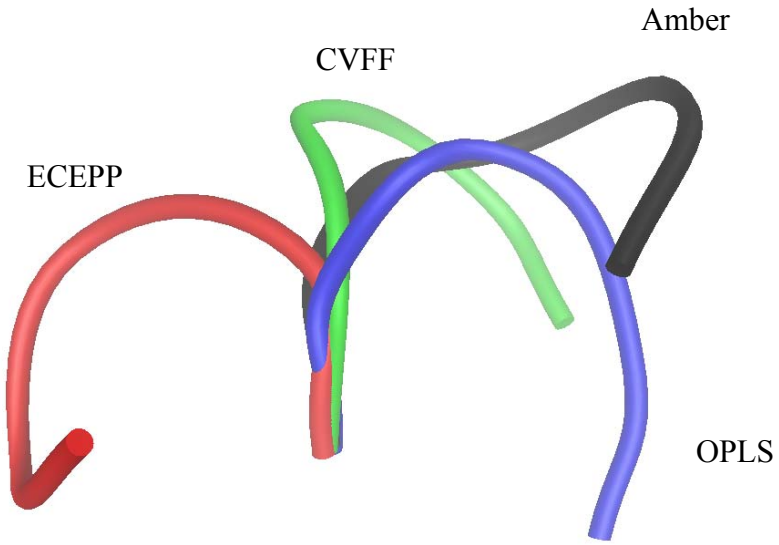


Figure 4.2 "Correct" folds for the PE models considered in the study superimposed on N-terms.

Unfortunately, carrying out even 500 realizations per control parameter combination for all the PE models would be computationally prohibitive[2]. We have, instead chosen to use $N_R = 300$ realizations for the vast majority of the work reported here, and then subdivide the performance-control parameter space into regions of $3\sigma_{F^{(99)}}$ or $6\sigma_{F^{(99)}}$ depending on the spread of the performance, where $\sigma_{F^{(99)}}$ is the standard deviation associated with the $F^{(99)}$ data as shown in Table 4.4. Figure 4.4, which compares $F^{(99)}$ evaluated for the OPLS PE model using $N_R = 300$

[2] A very large amount of resource – approximately 47 CPU core years on a cluster of 200+ Intel Xeon 5160 dual core CPUs running at 3.0GHz – was used in generating the data for this study.

and $N_R = 2000$ for a part of the control parameter space, suggests that the use of the lower resolution data is satisfactory.

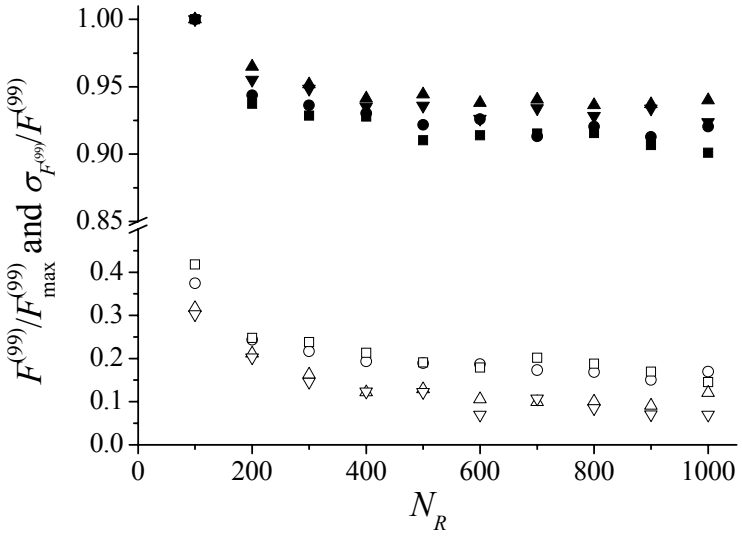


Figure 4.3 Variation of the performance measure, $F^{(99)}$, (closed symbols) and its standard deviation, $\sigma_{F^{(99)}}$, (open symbols) with the number of realizations, N_R, for the Amber (circles), OPLS (squares), CVFF (upward triangles) and ECEPP (downward triangles) PE models; the performance is shown relative to the maximum and the standard deviation as a fraction of the performance.

Table 4.4 Best and associated standard deviation for the PE models considered

PE model	$F^{(99)}_{best}$	$\sigma_{F^{(99)}}$
Amber	403419	119763
OPLS	299449	122250
CVFF	284976	73298
ECEPP	752592	132776

4.2.2. Influence of Potential Energy Model on Performance

Table 4.4 indicates that the best performances for the Amber, OPLS and CVFF PE models are, within statistical uncertainty, the same. The number of potential function evaluations required for the ECEPP model is, on the other hand, clearly much greater than the other two models – this difference may in part explain why Jin and co-workers (Jin et al., 1999) had far less success than Le Grand and Merz (Le Grand and Merz, 1993, 1994) in the earliest attempts to use EAs in the *ab initio* fold

prediction context. It is clear that the best performance that can be achieved for a given EA design is at the very least PE model dependent.

Figures 4.5 to 4.8 show the variation of EA performance with the various control parameters for all PE models considered in the study. All of the results represented here were collected at $P_X = 0.5$, but since crossover probability does not

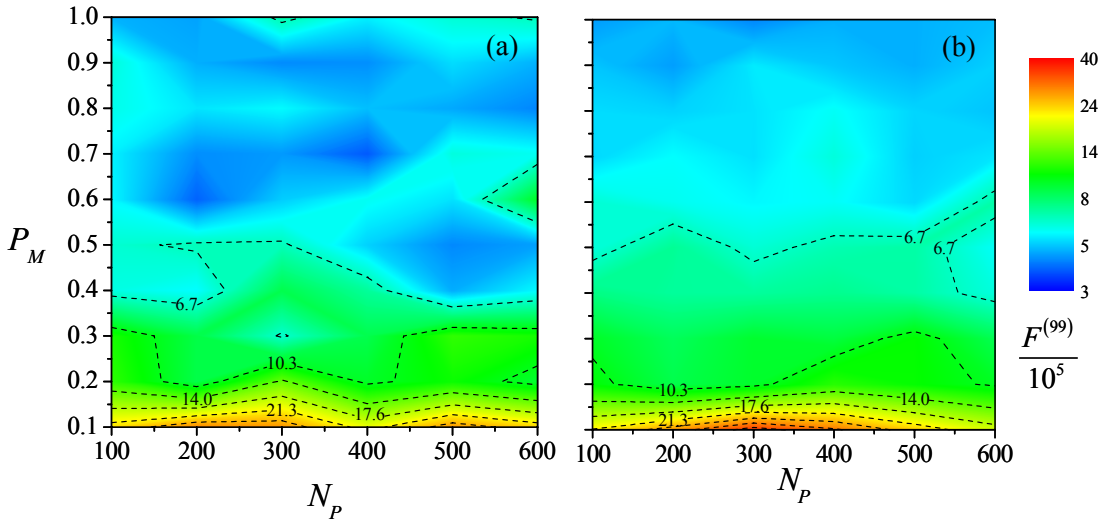


Figure 4.4 Sample variation of EA performance measure, $F^{(99)}$, with mutation probability, P_M, and population size, N_P, for the OPLS potential energy model evaluated using: (a) $N_R = 300$, and (b) $N_R = 2000$ realizations; other control parameter values are $P_X = 0.5$, $N_X = 4$, and $\alpha = 0.5$. The color scale is $\log_{10}$-based to enable the wide range of performances to be seen on a single set of plots. The lines are located at intervals of $3\sigma_{F^{(99)}}$ starting from $F_{best}^{(99)} + 3\sigma_{F^{(99)}}$, where the best performance, $F_{best}^{(99)}$, and standard deviation, $\sigma_{F^{(99)}}$, are given in Table 4.4.

have a strong effect on performance, results obtained for other P_X values show identical trends and are, hence, excluded from the analysis. For the sake of completeness, variation of EA performance with mutation and crossover probability is shown in Figure 4.9 for a single combination of N_P, N_X and α. It is clear that, although small changes in $F^{(99)}$ values with P_X are present, the variation is not high enough to be statistically significant and is well below the level of variation with P_M.

The level of fluctuations in the data, which is indicated by the standard deviation data in Table 4.4, means it is not possible to determine definitively the best optimal control parameter settings. However, the figures indicate that the performance is most influenced by the mutation probability, with the best

performance for Amber, OPLS and CVFF PE models occurring at higher mutation probabilities, whilst optimal performance of ECEPP PE model is achieved at mutation probabilities between 0.2 and 0.3.

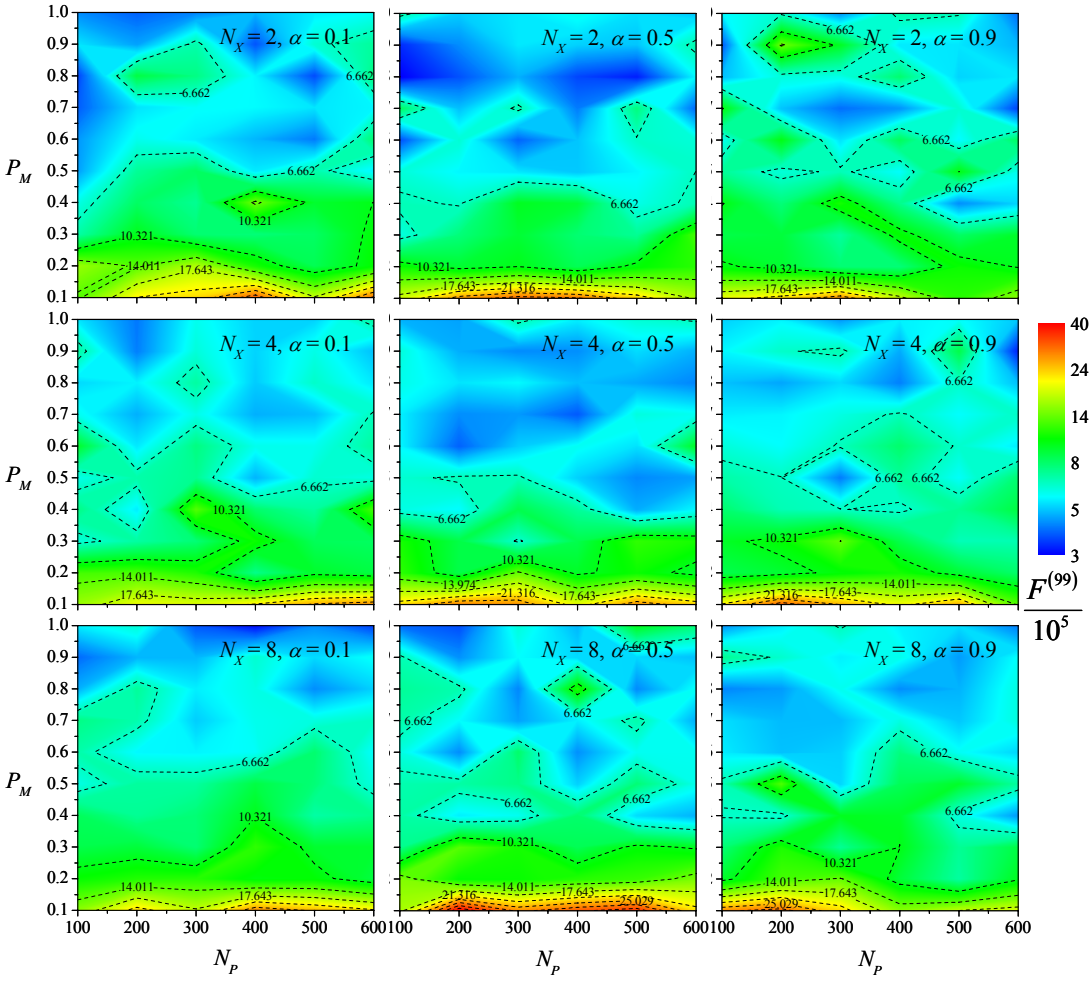


Figure 4.5 Variation of EA performance measure, $F^{(99)}$, with mutation probability, P_M, population size, N_P, number of crossover points, N_X, and selection parameter, α, for the OPLS PE model evaluated using $P_X = 0.5$ and $N_R = 300$.

For all PE models considered in the study, figures show that population size, N_P, has little effect on performance, provided the mutation probability is in a suitable range (high values of P_M for Amber, OPLS and CVFF and lower values for ECEPP PE models). Lower values of P_M for Amber and CVFF PE models, however, tend to require larger population sizes for better performance. To some extent, this effect can also be observed with OPLS PE model, especially for high α. This behaviour is explained through balance of diversity. Whilst decrease of P_M causes reduction in diversity, increase in population size has an opposite effect which compensates the

diversity loss. Amber and CVFF PE models are also characterised with an observable effect of truncation selection. Weaker selection pressures (i.e. $\alpha \to 1$) should be adopted at lower mutation probabilities in order to achieve better performance for these two PE models. Effect of α in ECEPP PE model is expressed through shift in optimal mutation probabilities towards lower values with the increase of α. Similar to crossover probability, number of crossover points, N_X, appears to play very small role in behaviour of any of the PE models.

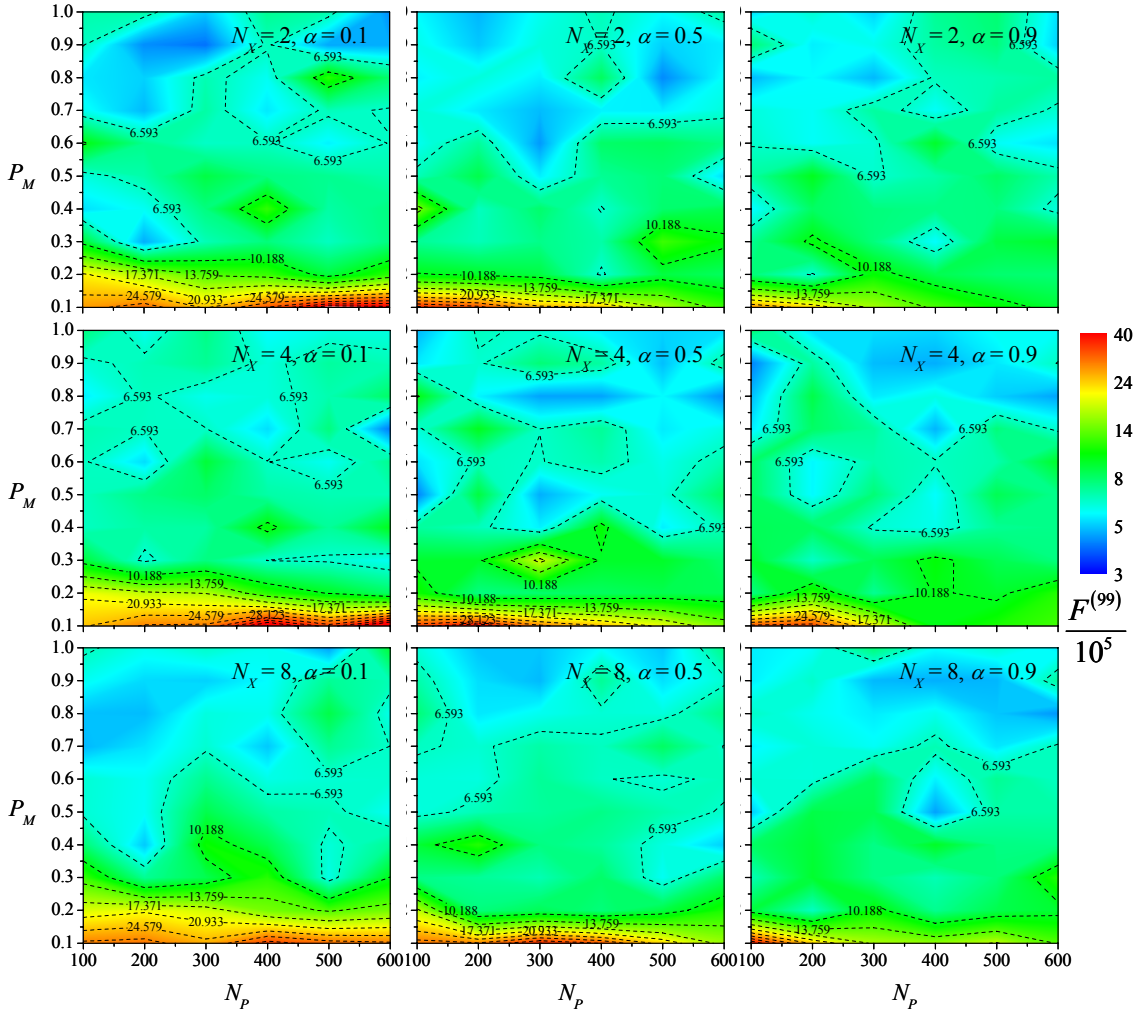


Figure 4.6 Variation of EA performance measure, $F^{(99)}$, with mutation probability, P_M, population size, N_P, number of crossover points, N_X, and selection parameter, α, for the Amber potential energy model evaluated using $P_X = 0.5$ and $N_R = 300$. Figure 4.4 provides additional explanation of color scale and positions of lines.

Within statistical uncertainty, Table 4.4 suggests that the switch to OPLS or CVFF PE models does not bring major changes in the optimal performance of the EA compared to Amber. Comparison of Figure 4.6 with corresponding plots for

OPLS and CVFF PE models (Figure 4.5 and Figure 4.7, respectively) suggests that the variation of performance with the control parameters is also little affected by the switch from Amber to the other two models. These results suggest that the differences in the functional forms of these PE models and their different parameter values have little effect on the performance characteristics of the EA

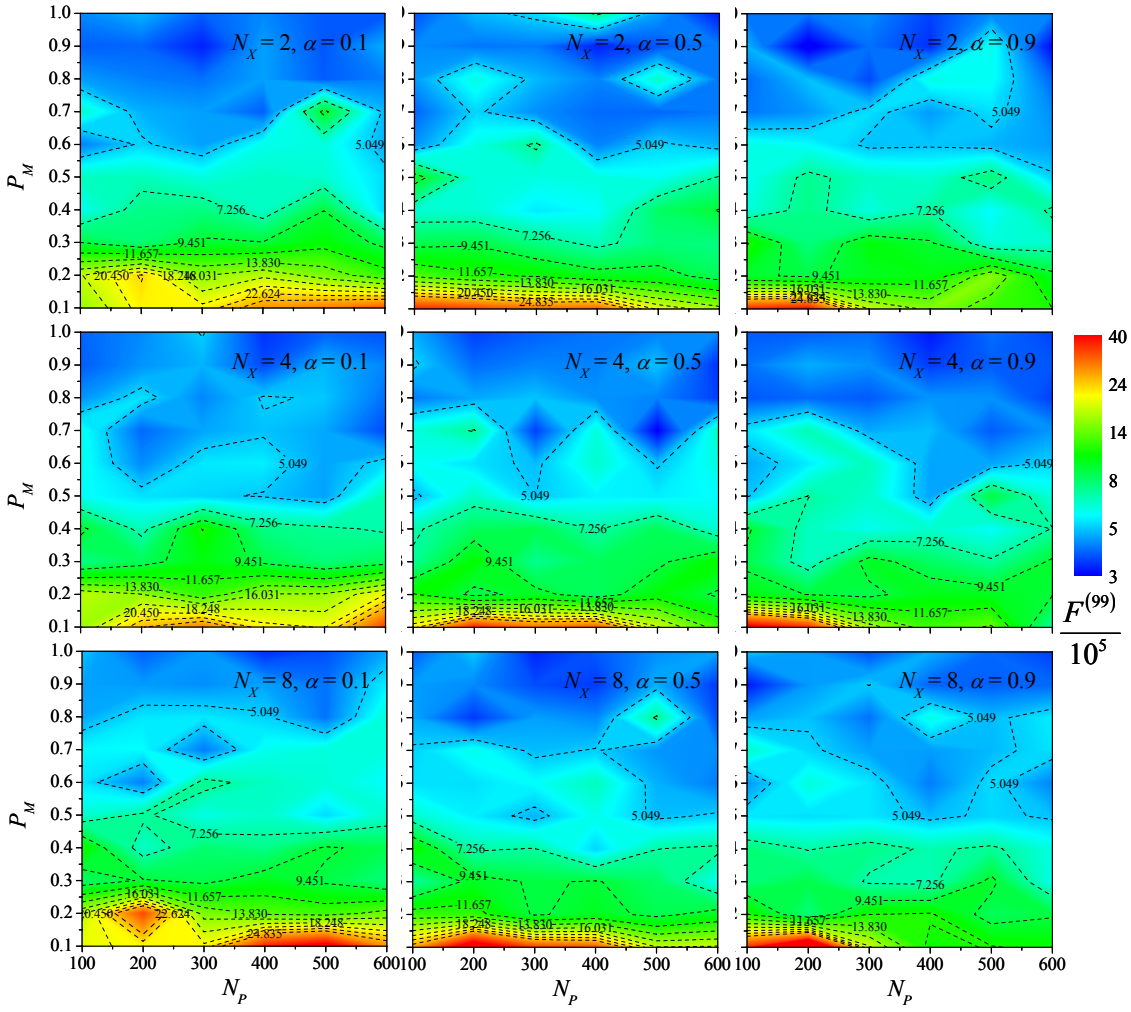


Figure 4.7 Variation of EA performance measure, $F^{(99)}$, with mutation probability, P_M, population size, N_P, number of crossover points, N_X, and selection parameter, α, for the CVFF PE model evaluated using $P_X = 0.5$ and $N_R = 300$.

On the other hand, analysis of Figure 4.8 shows that variation of EA performance in ECEPP PE model is substantially different from the results obtained for the other three PE models considered. The most notable difference is in the location of optimal performance region, which in ECEPP occurs at the lower end of the P_M range investigated in this study. This phenomenon, opposite to all three other PE models, is also accompanied by an increase in optimal $F^{(99)}$ value, which, as

Table 4.4 suggests, is no longer within statistical uncertainty from the best $F^{(99)}$ values of the other force fields.

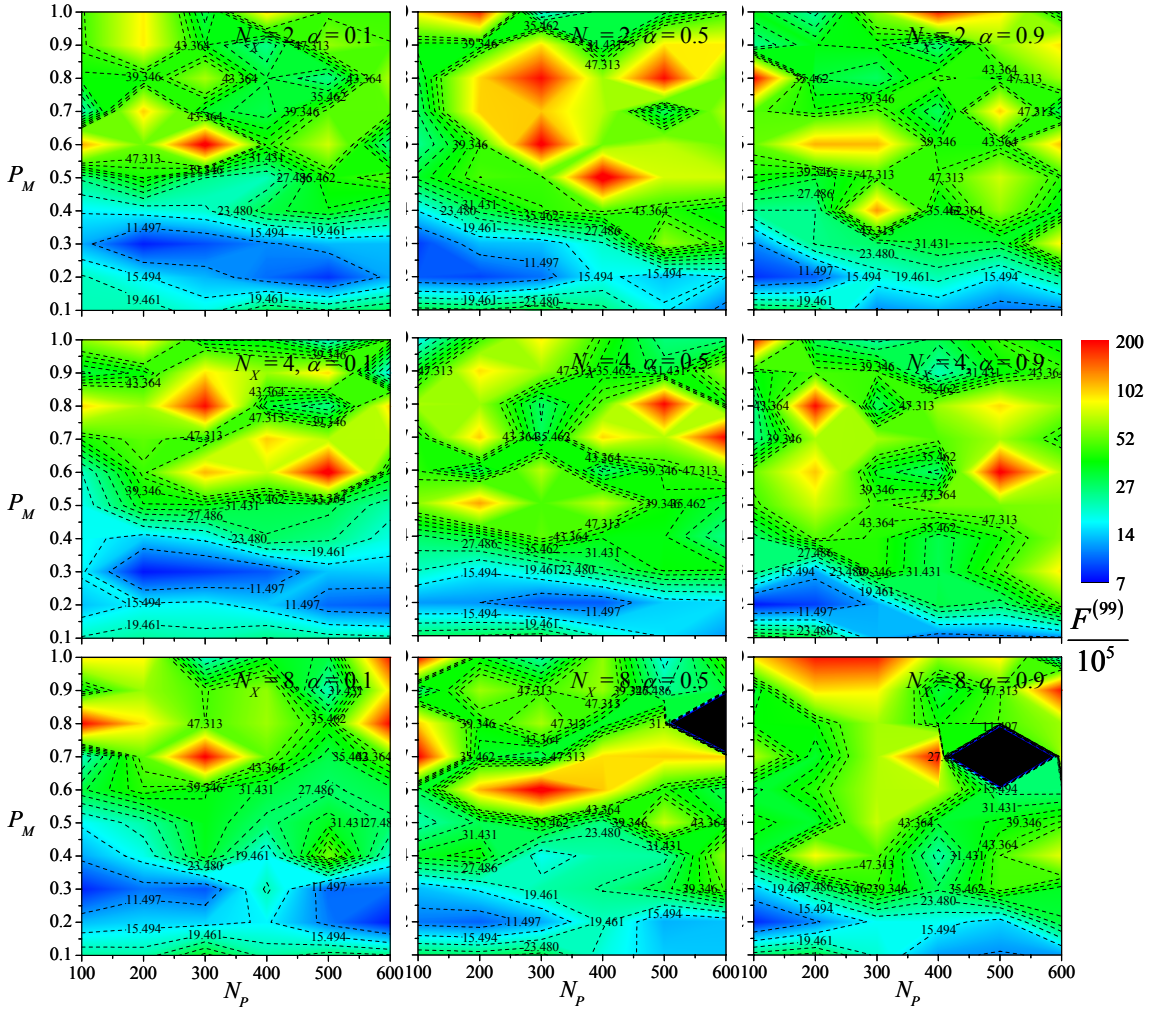


Figure 4.8 Variation of EA performance measure, $F^{(99)}$, with mutation probability, P_M, population size, N_P, number of crossover points, N_X, and selection parameter, α, for the ECEPP/3 PE model evaluated using $P_X = 0.5$ and $N_R = 300$.

The results obtained here indicate that EA performance is very much dependent on the nature of the PE model. However, not all differences are equal – using Amber as the benchmark, the hydrogen bond term in ECEPP has a disproportionate effect on performance compared to the higher-order torsional terms in the OPLS model. The results obtained here also suggest that EA performance is less dependent on the PE parameter values, although this presumably must be caveated by the need for their differences to not affect the fundamental character of the PE model (e.g. by switching the dominance of one or more terms). Bearing in mind that the character of the underlying fitness landscape can also be affected by

the protein and degrees of freedom being considered, the above results all suggest that an adaptive mutation probability should be used in the *ab initio* protein fold prediction context.

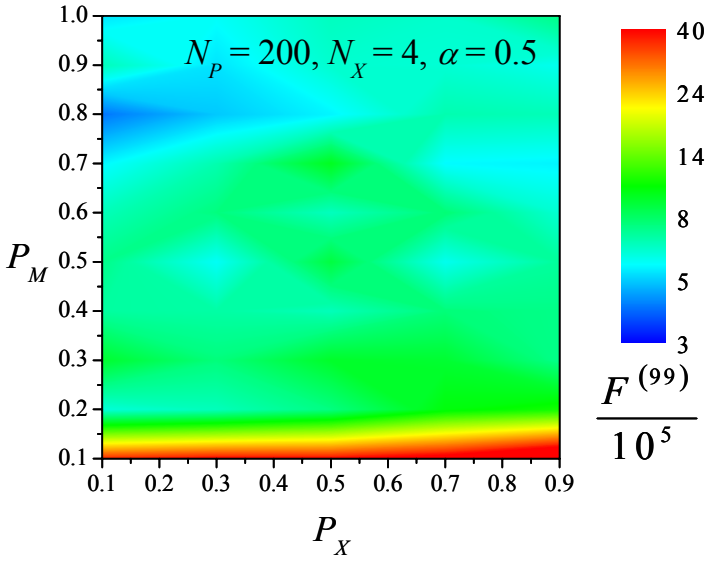


Figure 4.9 Variation of EA performance measure, $F^{(99)}$, with mutation probability, P_M, and crossover probability, P_X, for the Amber PE model evaluated using $N_P = 200$, $N_X = 4$, $\alpha = 0.5$ and $N_R = 300$

4.2.3. Influence of the Desired Level of Accuracy on Performance

An RMSD of 1 Å would be considered adequate for many purposes (Baker and Sali, 2001). However, higher levels of accuracy may be desirable under certain circumstances. Figure 4.10 shows how the performance characteristics for the Amber PE model are influenced by the level of accuracy demanded. It is clear that the mutation rates that lead to the best performance switch from higher to lower rates as the level of accuracy demanded increases, with the switch occurring at $\varepsilon_{RMSD} \approx 0.4$.

As expected, Figure 4.10 shows that the number of computations required to be 99% sure that the correct fold is obtained increases by a factor of 30 as the level of accuracy demanded changes from $\varepsilon_{RMSD} = 1$ to $\varepsilon_{RMSD} = 0.1$ Å. This level of computation would, of course, be unacceptable in general. However, the results of this and the previous section suggest that a much lower number of computations could be achieved whilst still achieving ultra-high accuracy by adapting the mutation probabilities during the course of the simulation, with higher rates early in a simulation and lower rates later.

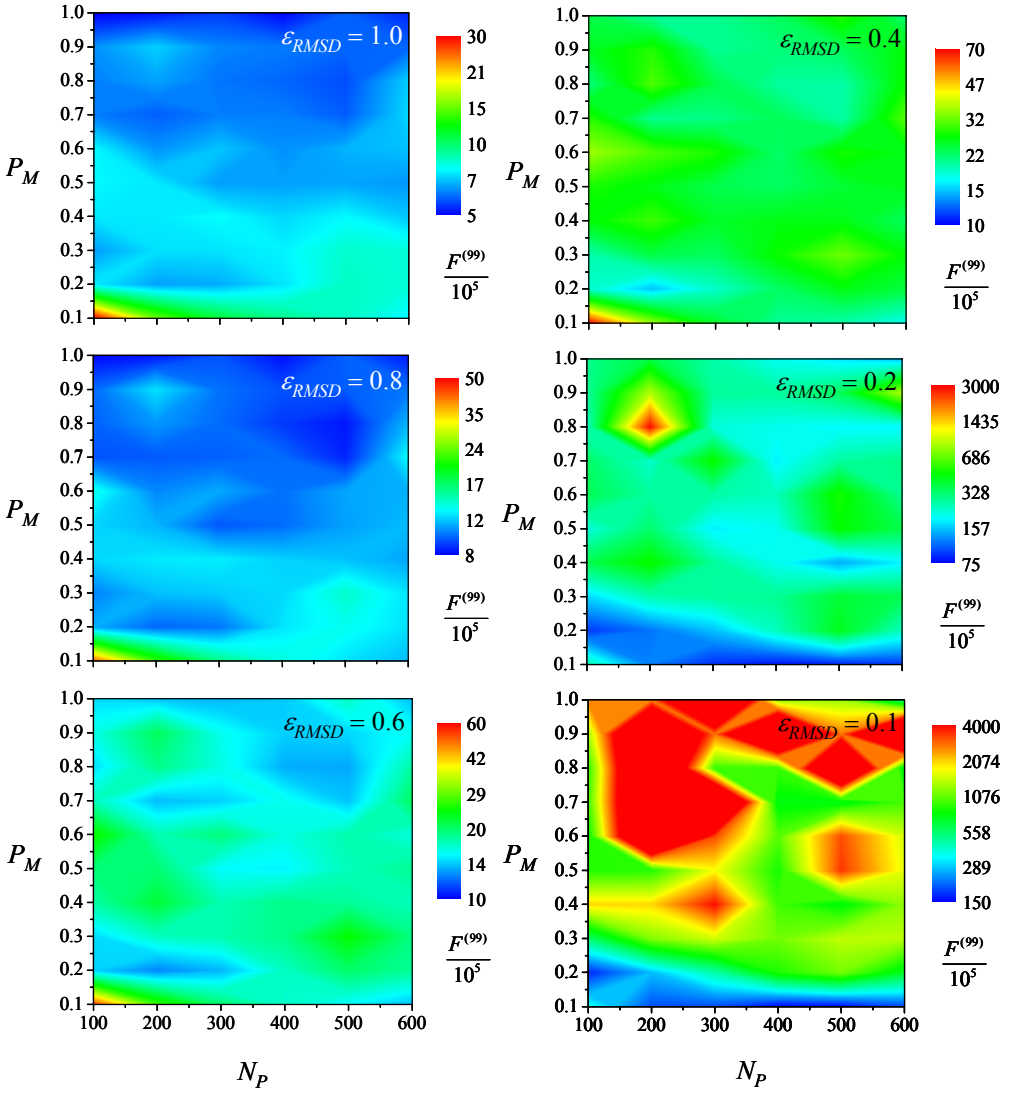


Figure 4.10 Variation of EA performance measure, $F^{(99)}$, with mutation probability, P_M, population size, N_P, and level of accuracy demanded from EA, ε_{RMSD}, for the Amber potential energy model evaluated using $P_X = 0.0$, $\alpha = 0.9$, and $N_R = 5000$.

4.3. Conclusion

It has been shown that the performance characteristics of an EA can be profoundly influenced by the potential energy (PE) model used in *ab initio* protein fold prediction. The minimum number of PE function evaluations required for the ECEPP PE model was approximately double that required for the Amber, OPLS and CVFF PE models. The range of optimal EA control parameters also differed significantly, with lower mutation rates being preferred by the ECEPP model and higher for the other PE models considered here.

It has also been shown that the EA performance characteristics are profoundly influenced by the level of accuracy demanded in a simulation – the amount of resource was found to increase 30 fold as the level of accuracy demanded increased by an order of magnitude, and the preferred range of mutation rates changed from high to low values.

The results here, as well as those in a previous study in our group (Djurdjevic and Biggs, 2006), all suggest that adaptive mutation probabilities are highly desirable when applying EAs in the *ab initio* context. The work here also suggests, on the other hand, that there is less of a need to implement adaptivity in other control parameters such as population size, crossover probability, selection pressure and number of crossover points.

Strong influence of the PE model choice to the EA performance is an indicator that the EA implementation used in this study is not very robust as the optimal set of parameters would have to be adjusted every time a new PE model is used. It should be noted, however, that the application of static EA parameters is inherent only to the basic EA implementation. It is expected that algorithmic improvements, such as adaptable mutation rate, will boost the performance as well as the robustness of the algorithm.

Chapter 5. EA Based Study of Polyalanine at a Gas-Solid Interface

5.1. Introduction

All of the studies conducted in our group so far (including Chapter 4 of this thesis) have been based on application of evolutionary algorithms (EA) in prediction of the 3D structure of proteins in vacuum or dilute gas phase (Djurdjević, 2006). Vacuum-based simulations were, however, used only as a testbed for developing a robust EA approach for prediction of protein conformation in an arbitrary environment. Of particular interest here are proteins at a solid-fluid interface.

Although majority of the applications of proteins at solid interfaces described in the Introduction occur on a liquid-solid interface, analysis of protein adsorption on solid surface from gas phase is critical for understanding of the integral adsorption phenomenon. Simplification of the observed system by decoupling of protein-solid from protein-solvent and solid-liquid interactions allows better understanding of the mechanism of protein conformational changes induced by adsorption. A further generalisation was achieved by replacing met-enkephalin studied in Chapter 4 with a simpler molecule. Polyalanine was chosen here due to the small size of its side chain and well defined conformation in gas phase (Djurdjević, 2006). Finally, in order to reduce complexity of protein-surface interactions, smooth solid surface was used, thus allowing only van der Waals interactions between the two to be considered.

The description of the studied system is given first, including the definition of the molecules, interaction potentials and energy minimisation procedure. This is followed by analysis of conformational changes of polyalanine molecules and discussion of the relationship between these changes and surface energies. Major findings are summarised in the conclusion.

5.2. Study Details

5.2.1. Peptide, Solid Surface and Potential Energy Models

A fully atomistic off-lattice representation of polyalanine capped by acetyl (Ac) and amino-methyl (NHMe) groups at the N- and C-termini, respectively, as shown in Figure 5.1, was used. The intra-molecular potential energy (PE) for

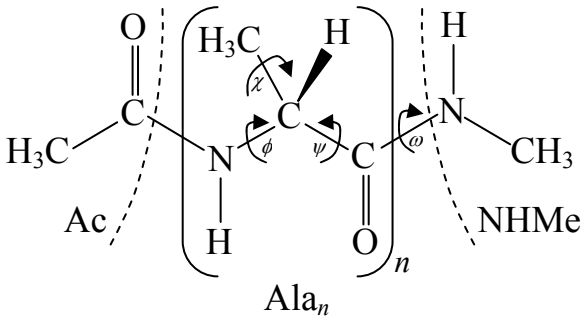


Figure 5.1 Schematic of polyalanine molecules considered here showing the acetyl (Ac) and amino-methyl (NHMe) caps at the N- and C-termini respectively, and the backbone, ϕ, ψ and ω, and sidechain, χ, dihedral angles.

polyalanine was modeled by the Amber potential (Cornell et al., 1995).

A rigid, uncharged smooth solid surface composed of L layers of solid atoms was considered. The PE arising from the interaction between this solid surface and an atom-j of the protein was modeled by the Steele potential (Steele, 1974)

$$E_{js} = 2\pi\rho\sum_{l=0}^{L-1}\varepsilon_{js}\sigma_{js}^{2}\left[\frac{2}{5}\left(\frac{\sigma_{js}}{z+l\Delta}\right)^{10}-\left(\frac{\sigma_{js}}{z+l\Delta}\right)^{4}\right] \tag{5.1}$$

where Δ is the distance between the solid layers, ρ is the density of the solid atoms within the layers, σ_{js} and ε_{js} are the Lennard-Jones length and energy parameters, respectively, for the interaction between a protein atom-j and a solid atom, and z is the normal distance between the protein atom and the solid surface. Given the simplicity of the side chain of polyalanine and the overall charge neutrality of the molecule, the use of a smooth surface should be a satisfactory model for various metal surfaces (Braun et al., 2002) and other materials such as the basal plane of graphite (Cracknell et al., 1995; Nicholson, 1996; Bandosz et al., 2003).

The entropy contribution has not been analysed numerically in this study, which, in the first instance, limits some of our findings to very low temperatures.

This issue will, however, be of concern in a limited range of investigated parameters and should not affect general conclusions. Possible effects of the entropy inclusion will be discussed further below.

5.2.2. Methodology

Stable polyalanine conformations were identified with the global potential energy (PE) minimum found by an evolutionary algorithm (EA) acting on the distance of the first nitrogen atom from the solid surface, the angle between the solid surface normal and the N-C_α bond in the first residue, the angle of the second C atom about the N-C_α bond of the first residue, and all dihedral angles except those about the peptides bonds, which were fixed at $\omega = 180^\circ$. All bond lengths and bond angles were fixed at the values obtained by locally relaxing a single Ac/NHMe-capped alanine residue initialized with the equilibrium PE model parameters.

The EA was based on the SRM (steady state, real encoding and multipoint crossover) design described in Djurdjevic and Biggs (Djurdjevic and Biggs, 2006) except truncation (Schwefel, 1981; Bäck and Hoffmeister, 1991) rather than tournament selection was used, as in Chapter 4. For each polyalanine/solid surface combination, the EA was initially run 10^4 times with different random number seeds. If the lowest energy conformation had not been identified more than once, further runs were undertaken until this occurred; the number of runs required typically varied from around 10^4 for the smallest molecules in the less challenging regions of the conformational space, to 2-7 times this number in the more challenging parts of the space such as near the switching points where the small energy differentials between the conformations meant the minimum with the wider funnel mouth tended to be preferentially identified even when it was not the global minimum. The lowest energy structure identified was always compared against the other main possibilities, as described below.

As the number of runs required increased substantially with the number of residues, *n*, it was necessary to adopt a three stage strategy in some of the more challenging parts of the conformational space for the larger molecules. The first stage involved application of the EA to all degrees of freedom as usual. In the second stage, the initial population was seeded with the best conformation from the first stage and only the distance and orientation of the polyalanine molecule from the

solid surface were varied (i.e. all the dihedral angles were fixed). The final stage was the same as the first except the initial population was seeded with the best conformation from the second stage.

As the initial results obtained from the single and three-stage strategies for the larger peptides were inline with those of the smaller peptides, many of the results for the former were obtained by constraining the dihedral angles within ±15º of the angles of the expected conformations (described in more details below) and then selecting that with the lowest energy.

5.2.3. Study Details

Polyalanine molecules of $n = 6$, 8, 10 and 12 residues were considered in detail for surface energies in the range of $E_s = 0.0 - 4.0E_g$ at intervals of $0.1E_g$, where E_g is the energy arising from the interaction between the protein in the given configuration and the [111] gold surface characterized by the parameters given in Table 5.1 (Mahaffy et al., 1997). Polyalanine molecules of $n = 7$, 9, 11, 13 and 14 residues were also considered at surface energies around the conformational switching points for these peptides.

Table 5.1 Gold surface potential energy interaction parameters

Parameter	**Value**[a]
P	0.13886 atoms/Å^2
ε_g	0.0905 kcal/mol
σ_g	3.359 Å
Δ	2.3545 Å
L	2

a. The parameters are based on those of the [111] surface of gold (Mahaffy et al., 1997).

Although the change in surface energy may be interpreted physically in a number of ways – variation of the solid density (via ρ or Δ) or solid atom Lennard-Jones parameters – it was achieved here through the expedient of multiplying the energy evaluated for each protein atom-gold surface interaction by the requisite factor (i.e. for $E_s = KE_g$, the PE obtained from equation (5.1) using the parameters of Table 5.1 was multiplied by K). The Lennard-Jones (LJ) parameters for the protein-gold interaction, σ_{js} and ε_{js}, were obtained by combining the protein atom

parameters from the *ff*94 Amber parameter set (Cornell et al., 1995), σ_j and ε_j, with those of the gold atoms using the Lorentz-Berthelot rules (Allen and Tildesley, 1989)

$$\sigma_{js} = \frac{\sigma_j + \sigma_g}{2} \tag{5.2}$$

$$\varepsilon_{js} = \sqrt{\varepsilon_j \varepsilon_g} \tag{5.3}$$

The evolutionary algorithm (EA) required a number of control parameters to be set including the population size, N_P, mutation probability, P_M, crossover probability, P_X, number of crossover points, N_X, and the fraction of the rank-ordered population used in uniform selection, α (Djurdjević, 2006; Djurdjevic and Biggs, 2006). Previous work by Biggs and co-workers (Djurdjević, 2006; Djurdjevic and Biggs, 2006; Mijajlovic and Biggs, 2007a) and results presented in Chapter 4 have shown that EA performance is sensitive to the control parameter values, and that optimal values vary with the peptide details and potential energy model. As this study was focused on the phenomenology rather than the computational issues, limited effort was expended on determining the optimal parameter values. Instead, reasonable estimates of the optimal mutation and crossover probabilities were obtained by varying them as indicated in Table 5.2 whilst keeping the remaining parameters, which we have found to generally have a secondary effect on EA performance, fixed at the values also shown in this table. Although insufficient results were obtained to make definitive statements on the most appropriate values for crossover and mutation probabilities, good performances were in general obtained when using P_X and P_M values from the middle and bottom end of the ranges given in Table 5.2, respectively.

The principal results obtained from the EA were the conformation of the peptide in the form of its distance from and orientation to the solid surface and its dihedral angles, the intramolecular potential energy (PE) for the peptide broken down into its non-frozen components recognized by Amber PE model(Cornell et al., 1995) (i.e. torsional, electrostatic, dispersion, electron cloud overlap), and the peptide-surface PE. These data were used to generate a number of additional results as follows:

- The root mean square deviation (RMSD) of the C_α atoms relative to the gas phase conformation determined by the EA, which is an α-helix as expected (Ripoll and Scheraga, 1988; Park and Goddard, 2000).

- The number of residues per turn, S, as per Quine (Quine, 1999) and Otero-Cruz and co-workers (Otero-Cruz et al., 2007). Both methods give very similar results, with the average differences being less than the uncertainty associated with the number of residues per turn.
- The normal distance of the peptide centroid from the solid surface, d_c. The coordinates of the centroid were determined by $\overline{x}_i = \sum x_{ij} / N$ for $i = 1$, 2 and 3, where x_{ij} is the i^{th} coordinate of peptide atom-j in three-dimensional space, and the summation is over the N atoms of the peptide.
- The angle between the peptide axis and the solid surface, θ, which is termed henceforth the angle of tilt. The peptide axis was determined by minimizing the function $\sum d_j^2$, where d_j is the normal distance from the axis to atom-j of the peptide, and the summation is over all the atoms of the peptide except those associated with the caps. This definition is similar to that of Martin and co-workers (Martin et al., 2005), except they sum over the C_α atoms only.
- The strain along the peptide axis, $(l - l_0)/l_0$, where l is the distance between the N atom of the first residue and the C atom of the last residue, and l_0 is the length of the α-helix in the gas phase.
- The energy associated with the hydrogen bonds as per the DSSP (Kabsch and Sander, 1983).

Table 5.2 Evolutionary algorithm control parameter values used.[a] Meaning of the symbols is explained in the text.

Parameter		**Value**s
N_P		400
α		0.9
N_X	Peptide position/orientation	0
	Dihedral angles	4
P_X	Peptide position/orientation	0.0
	Dihedral angles	0.1, 0.3, 0.5, 0.7, 0.9
P_M	Peptide position/orientation	0.17
	Dihedral angles	0.03, 0.05, 0.1, 0.3, 0.5, 0.8, 0.85, 0.9, 0.95

a. Although stored in a single chromosome, different EA parameters were applied to the degrees of freedom defining the position and orientation of the peptide to the solid surface, and the dihedral angles.

5.3. Results and Discussion

5.3.1. Conformational Change with Surface Energy for 6-alanine

Figure 5.2 which shows the variation of the RMSD of 6-alanine with the surface energy, clearly indicates that the peptide conformation undergoes step changes at $E_s = 0.878E_g$ and $E_s = 2.158E_g$. Examples of the conformations associated with the three distinct RMSD ranges and the associated dihedral angle distributions in Ramachandran space are shown in Figure 5.3 and Figure 5.4, respectively.

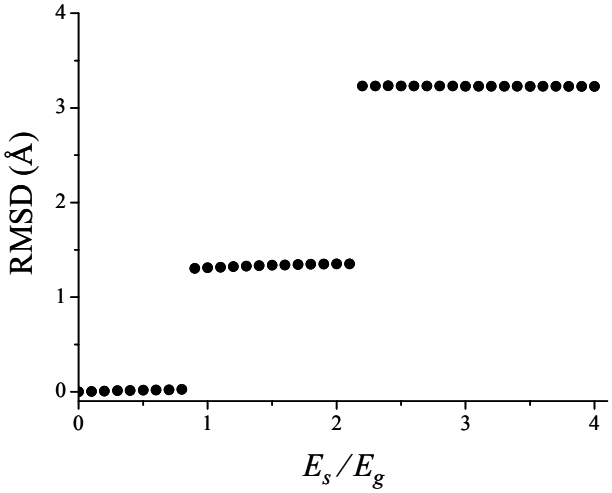


Figure 5.2 Variation of RMSD of 6-alanine with surface energy.

Analysis of the conformations associated with the lower of the three surface energy ranges, an example of which is shown in Figure 5.3(a), reveals all are characterized by an $i+4 \rightarrow i$ hydrogen bonding pattern, in which the weakest and strongest bonds have energies of -1.622 and -2.248 kcal/mol, respectively. All rings formed by hydrogen bond consist of 13 atoms. Although the dihedral angles of these conformations do deviate slightly from that of the gas phase conformation, the number of residues per turn is essentially the same for all the conformations at $S = 3.6$. The conformations up to $E_s = 0.878E_g$ may all, therefore, be properly termed 3.6_{13} (i.e. α) helices.

Analysis of the conformations determined within the intermediate surface energy range such as that shown in Figure 5.3(b) reveals in every case an $i+3 \rightarrow i$ hydrogen bonding pattern with 10 atoms per hydrogen bonded ring and with energies

of the weakest and the strongest bonds equal to -1.894 and -2.151 kcal/mol, respectively. Further analysis shows that whilst once again the dihedral angles of these conformations do change slightly over the surface energy range, the number of residues per turn essentially remains unchanged at $S = 3.1$. The conformations in the intermediate surface energy range may, therefore, be most correctly referred to as 3.1_{10}- helices.

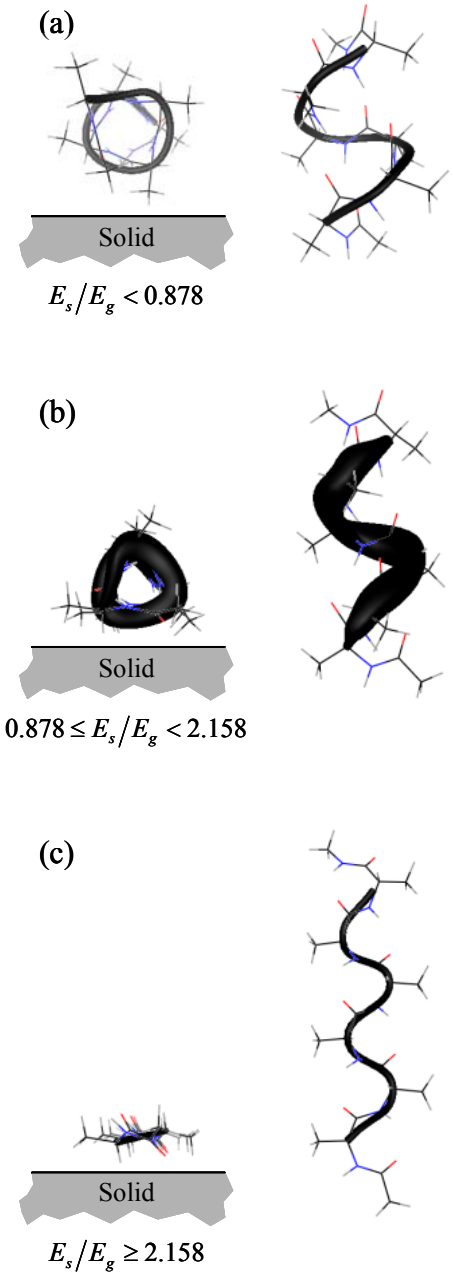


Figure 5.3 The three conformations observed for 6-alanine as surface energy is increased, from the N-term end (left) and from the top (right): (a) α-helix; (b) 3.1_{10}-helix; and (c) 2_7-helix. The corresponding surface energy ranges over which the conformations are stable are also shown.

Detailed consideration of the conformations obtained in the uppermost surface energy range like that shown in Figure 5.3(c) reveals an $i+2 \rightarrow i$ hydrogen bonding pattern with 7 atoms in the hydrogen bonded rings and with the weakest and strongest bonds characterised with energies of -1.737 and -2.268 kcal/mol, respectively. As the number of residues per turn for all the conformations in this surface energy range is identical at $S = 2.0$ despite the dihedral angles of the conformations changing slightly with surface energy, they can all be described as 2_7-helices. With the side chains being in the plane of the helix, these conformations are essentially identical to that proposed by Zahn in 1947 for α-keratin (Zahn, 1947), which Bragg and co-workers denote as 2_7b (Bragg et al., 1950) in order to differentiate it from the much less stable alternative 2_7 conformation of Huggins (Huggins, 1943). Whilst the 2_7-helical conformation obtained here has long been hypothesised for proteins in solutions or crystals (Zahn, 1947; Bragg et al., 1950), we have found only one reported experimental observation under such conditions (Pervushin and Arseniev, 1992) – the results here suggest they could possibly be more prevalent for proteins near solid surfaces.

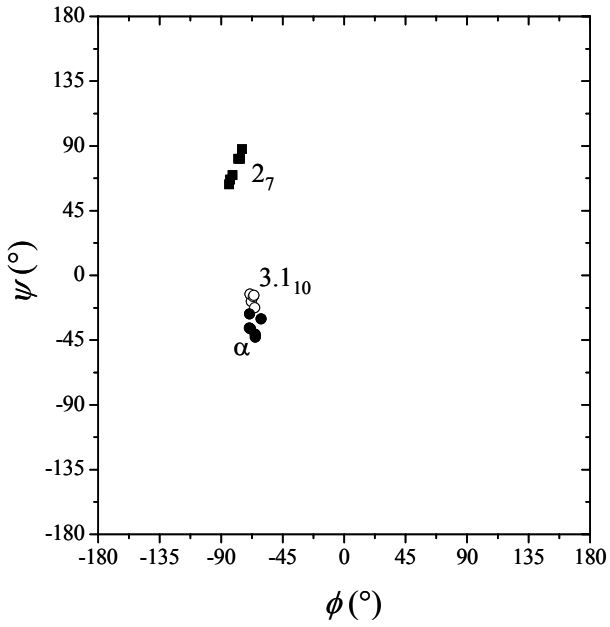


Figure 5.4 Ramachandran plot showing the backbone dihedral angles for one example of each of the conformations in Figure 5.3.

As seen in Figure 5.4 and Figure 5.5, the conformational switches are reflected in the various characteristics of the peptide/surface system. The length of the peptide undergoes a substantial change at each transition, Figure 5.5(a), with the 3.1_{10} and 2_7 conformations being ~17% and ~59% longer than the α-helical conformations, respectively. Figure 5.5(b) shows that the normal distance between the centroid of the peptide and the solid surface also undergoes a step change as the peptide switches from the α-helix ($d_c \approx 5.45\,\text{Å}$) to the 3.1_{10}-helix ($d_c \approx 4.70\,\text{Å}$) and, finally, 2_7-helix ($d_c \approx 3.88\,\text{Å}$). The variation of the normal distance between the lower surface of the peptide and the solid surface, d_s, also shown in Figure 5.5(b) indicates, however, that the peptide moves closer to the solid surface in the first conformational switch and then away again in the second switch. Figure 5.5(c) shows that whilst the angle between the peptide axis and the solid surface, θ, is always small (i.e. the molecules lay almost flat to the solid surface), it too undergoes a step change at the conformational switches.

A number of the characteristics of the α- and 3.1_{10}-helical systems appear to experience some change over their associated surface energy ranges. As the changes are much less than both those that occur at the conformational switches and the approximations inherent to the model, we focus here on only the very notable nonlinear decline in the tilt seen for the 3.1_{10}-helical conformation, Figure 5.5(c). Origins of this behavior can be explained using Figure 5.6, which shows a simplified representation of the conformations of the 3.1_{10}-helix at either end of the associated surface energy range. At the lower end of the surface energy range shown at the left of this figure, the "virtual bonds" (Quine, 1999) nearest the solid surface are inclined to the surface, with the degree of inclination decreasing from the NHMe cap to the Ac cap. An increase in the surface energy leads to a reduction in this inclination by offsetting the unfavorable change in the intramolecular energy arising from the required changes in the relevant dihedral angles. This reduction in the inclination leads to a corresponding decrease in the displacement between the NHMe and Ac caps normal to the solid surface and, hence, the angle between the peptide axis and the solid surface.

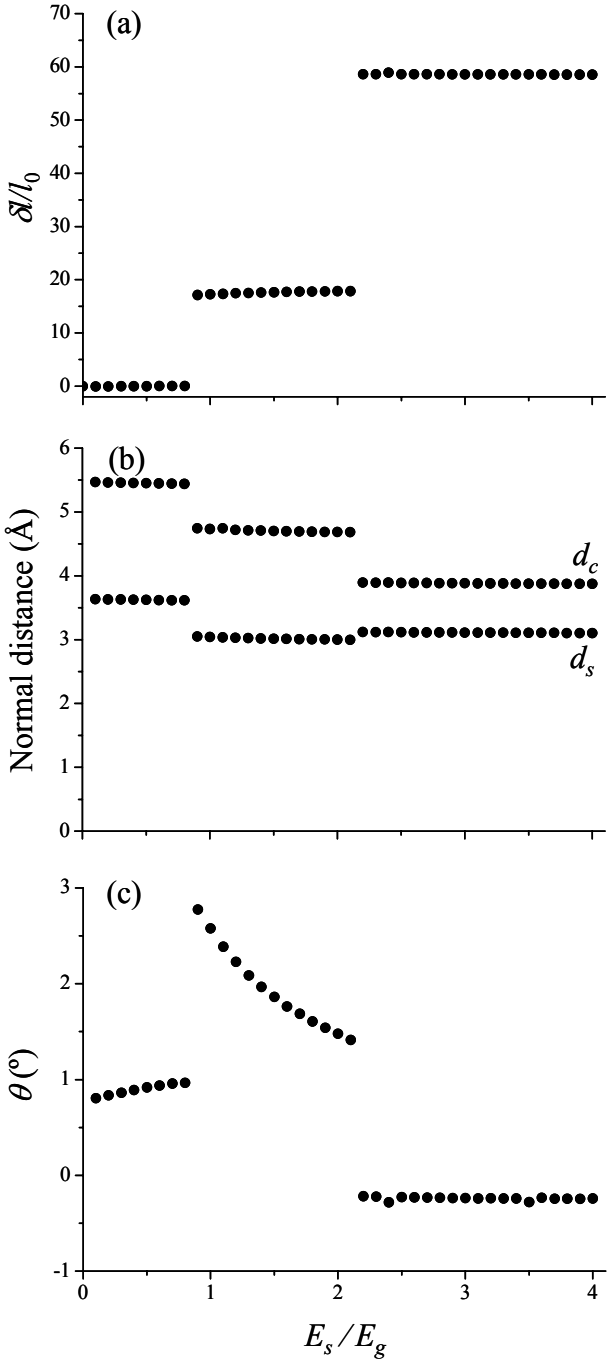


Figure 5.5 Variation of conformational measures for 6-alanine with surface energy: (a) longitudinal strain; (b) normal distance between peptide centroid and the solid surface, d_c, and peptide lower surface and the solid surface, d_s, where the latter is obtained by subtracting from the former the peptide radius of gyration component normal to the solid surface; and (c) angle between the peptide axis and solid surface (the tilt).

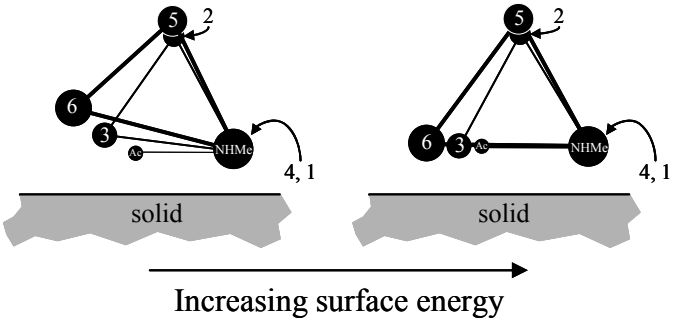


Figure 5.6 A schematic to aid explanation of the change in the tilt of the 3.1_{10}-helix as the surface energy increases, Figure 5.5(c). The residues and caps of the peptide are replaced by beads connected by what Quine (Quine, 1999) terms "virtual bonds". The beads are located at the C_α and the methyl C atoms of the caps. The size of the beads reduces from the NHMe cap in the foreground to the Ac cap in the background. Angles have been exaggerated to aid the discussion.

5.3.2. Energetics of Adsorption of 6-alanine

Figure 5.7(a) shows that whilst the conformational potential energy (PE) of the peptide changes adversely as it switches from an α-helix to a 3.1_{10}-helix and, finally, a 2_7-helix, the PE arising from the solid surface is more than sufficient to stabilise the respective conformations. The continuous decrease of the total PE of the system with rising surface energy begs the question of why the peptide conformation does not also gradually change. The reason becomes clear when the intramolecular PE is decomposed into that arising from the hydrogen bonds and that which does not, Figure 5.7(b). This figure shows that the combined effect of the torsional and non-hydrogen bond electrostatic and LJ interactions is to destabilize the 3.1_{10} and, even more so, 2_7-helical conformations relative to the α-helix. The hydrogen bonds, on the other hand, always act to stabilize the higher surface energy conformations – in short, only hydrogen bond stabilized conformations are possible. It is clear, therefore, that because continuous conformational change would lead to a breaking of the hydrogen bonds at some point, such change is not possible.

The precise switching surface energy can be identified with the point of intersection of the lines that define the variation of the total energy of the two conformers with the surface energy as illustrated for the $\alpha \rightarrow 3.1_{10}$-helix switch in the insert of Figure 5.7(a). Using this observation and assuming that between the switching points the conformational PE of the peptide is constant and the peptide-

surface PE varies in a linear manner with surface energy (i.e. the peptide conformation as a whole remains unchanged between the switching points), the switching surface energy can be estimated using

$$E_{sw} = -\frac{\Delta U}{\Delta S_{ps}} \tag{5.4}$$

where ΔU and ΔS_{ps} are the changes in the peptide conformational PE and the derivative of the protein-surface PE with respect to the surface energy, dE_{ps}/dE_s, across the switch respectively (derivation provided in the appendix). Application of this expression to estimate the switching points for a peptide in principle requires a simulation for each conformation on a single solid surface (i.e. one surface energy) only. Good estimates can be obtained by using the same solid surface but, because the peptide conformation as a whole does change with surface energy slightly (as

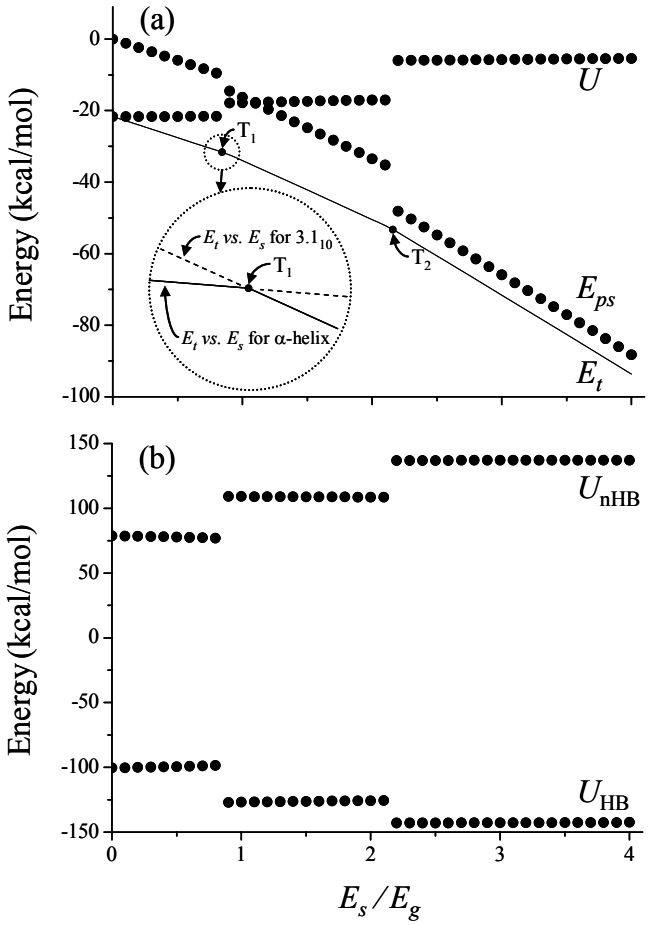


Figure 5.7 Variation of various potential energy (PE) contributions with the surface energy: (a) peptide conformational PE, U, peptide-surface PE, E_{ps}, and total PE, $E_t = U + E_{ps}$; and (b) PE associated with hydrogen bonding in the peptide, U_{HB}, and the remaining peptide conformational PE, $U_{nHB} = U - U_{HB}$.

shown in Figure 5.5), identification of the switching points to a high level of accuracy requires a small number of iterations.

Casual study of Figure 5.3 may suggest that the various conformations may possess some symmetry about the peptide axis. Detailed consideration of the potential energy surface (PES) of each conformation about this axis, shown in Figure 5.8, reveals that this is in fact not true. In the case of α-helical conformations, Figure 5.8(a) shows that the rotational path around the peptide axis, which also requires some change in the angle of tilt and (not shown) distance from the solid surface, is characterised by six non-equivalent minima with the energies given in Table 5.3. This lack of symmetry essentially arises out of the number of turns about the axis being fractional (i.e. there is an incomplete turn), which leads to one part of the peptide being more dense, and therefore more active to the solid, than the remainder. Figure 5.8(b) and Table 5.4 reveal three non-equivalent minima as the 3.1_{10}-helical conformation is rotated about its axis; this arises from the difference in the number of side chains at the three vertices of the helix (in the case of the 6-alanine peptide considered here, two have 3 groups each whilst the third has only 2) and the number of oxygen and nitrogen atoms on the three "faces" of the helix (there is 3 of each atom on the helix "face" closest to the solid and 2 of each atom on the other helix "faces" for the 6-alanine peptide here). The lack of energetic symmetry about the peptide axis for the 2_7-helical conformation, as shown in Figure 5.8(b) and Table 5.5, arises from the differing number of oxygen atoms on the two sides of the structure (for the 6-alanine peptide considered here, there are 4 on one side and 3 on the other, for example).

Inspection of Table 5.3 to Table 5.5 shows that whilst the energy differences between the global minimum and the other local minima about the peptide axis are relatively small for all three conformations, the barriers to rotation away from the global minimum are very considerable indeed (~2200 K, ~4500 K and ~6500 K for the α, 3.1_{10} and 2_7 helices respectively). This suggests that the peptides, once adsorbed, will not rotate about their axis. Consideration of Figure 5.8 shows that the barriers to the variation of the angle of tilt, θ, are even greater than those to rotation about the peptide axis, suggesting that the orientation of the peptide to the solid surface is also likely to vary little once adsorption occurs.

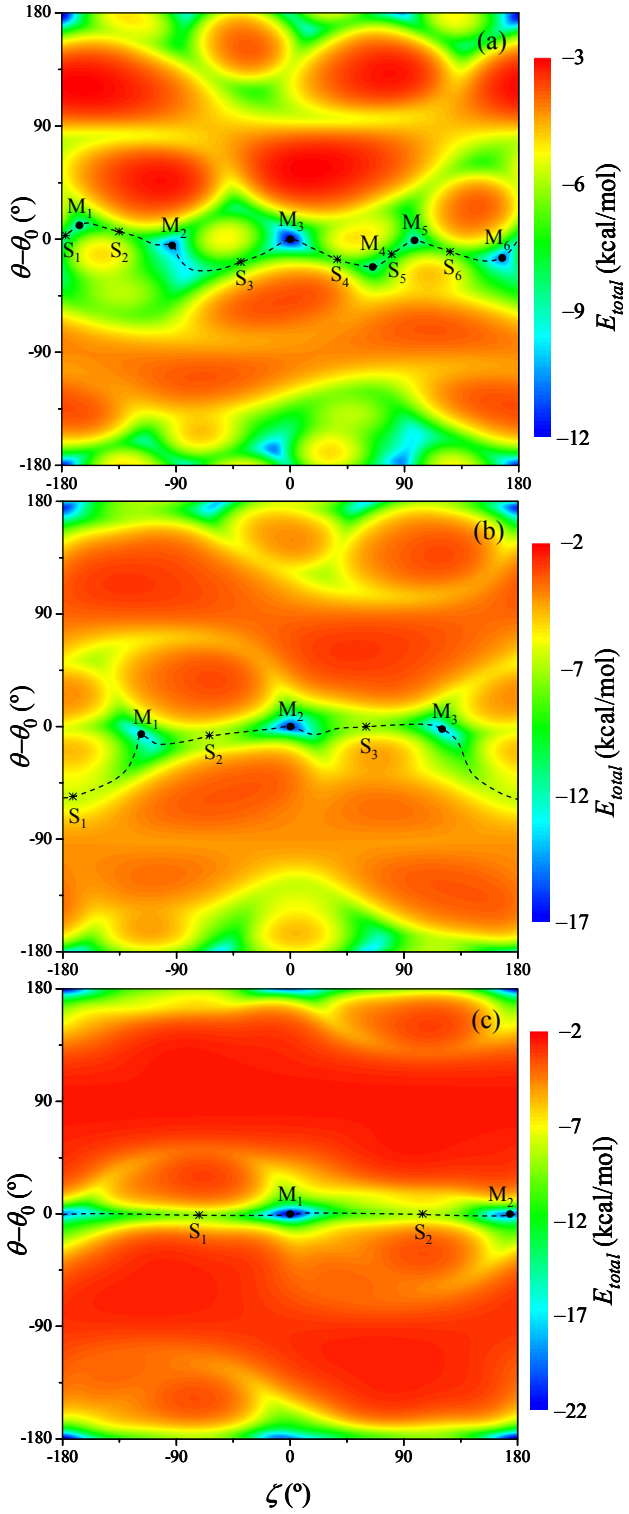


Figure 5.8 Variation of potential energy (PE) of 6-alanine with the angle of rotation about the peptide axis, ζ, and the angle of tilt, θ: (a) α-helix, (b) 3.1_{10}-helix, and (c) 2_7-helix. The angle of rotation has been arbitrarily defined relative to the global minimum, whilst the angle of tilt has been given relative to that at the global minimum, θ_0 (i.e. in each case, the global minimum is located at the centre of the PE surfaces shown here).

Table 5.3 Minima and saddle points in the potential energy surface of Figure 5.8(a) and associated energies relative to the global minimum (M_3).

Point	ζ (º)	θ (º)	E (kcal/mol)
S_1	-177	3	3.50
M_1	-166	11	2.97
S_2	-135	6	5.12
M_2	-93	-5	1.24
S_3	-39	-18	4.47
M_3	0	0	0.00
S_4	37	-16	5.06
M_4	65	-22	3.27
S_5	80	-12	3.98
M_5	98	-1	2.83
S_6	126	-10	5.04
M_6	167	-15	1.18

Table 5.4 Minima and saddle points in the potential energy surface of Figure 5.8(b) and associated energies relative to the global minimum (M_2).

Point	ζ (º)	θ (º)	E (kcal/mol)
S_1	-172	-56	9.65
M_1	-118	-6	2.90
S_2	-64	-7	9.39
M_2	0	0	0.00
S_3	60	0	9.04
M_3	120	-2	2.39

Table 5.5 Minima and saddle points in the potential energy surface of Figure 5.8(c) and associated energies relative to the global minimum (M_1).

Point	ζ (º)	θ (º)	E (kcal/mol)
S_1	-72	-1	12.99
M_1	0	0	0.00
S_2	105	0	13.70
M_2	174	0	2.52

5.3.3. Effect of Number of Alanine Residues

The results presented above for 6-alanine are qualitatively similar for the other polyalanine molecules considered in the study. The size of the peptide did, however,

quantitatively affect two key aspects of the switching phenomenon – the switching surface energy and the longitudinal strain – which are considered here in detail.

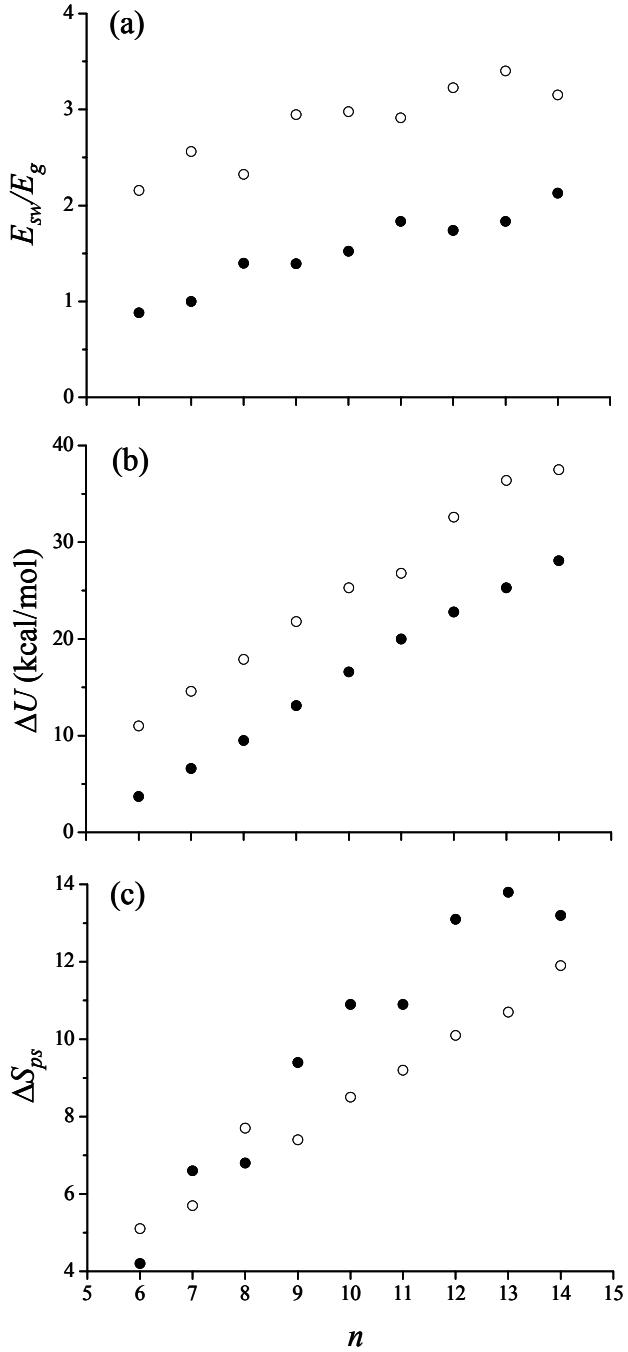


Figure 5.9 Variation of switching-related energetic characteristics with the number of alanine residues for the $\alpha \rightarrow 3.1_{10}$ (closed circle) and $3.1_{10} \rightarrow 2_7$ (open circle) switches: (a) switching surface energy, (b) change of conformational PE across switches, and (c) change of dE_{ps}/dE_s across switches.

Figure 5.9(a) shows that whilst the switching surface energies, E_{sw}, tend overall to increase with the number of residues, the increases are neither smooth nor, indeed, locally monotonic. This complex change can be understood by first considering in turn the two factors that affect the switching surface energy as indicated by equation (5.4). Figure 5.9(b) shows that the change in the conformational potential energy (PE) of the peptide across the switches, ΔU, increases relatively smoothly for both switches – this clearly is not the cause of the complex change in the switching surface energy with the number of residues. Figure 5.9(c), on the other hand, shows that the change in dE_{ps}/dE_s across the switches, ΔS_{ps}, does vary in a complex manner with the number of residues. Moreover, comparison of this variation with that in Figure 5.9(a) reveals a definite correlation – for example, the relatively small increase in the $\alpha \rightarrow 3.1_{10}$ switching surface energy as the number of residues increases from $n = 6$ to $n = 7$ corresponds to a relatively large jump in ΔS_{ps}, whilst the relatively large jump in the switching surface energy for $n = 7$ to $n = 8$ corresponds to a small change in ΔS_{ps}.

The source of the complex change in ΔS_{ps} for a switch can be understood using analysis of change in peptide-surface PE with surface energy, E_s, for the three conformations. If we assume the conformations within their range of stability remain essentially rigid – which results from §5.3.1 show to be a good approximation – then, because the variation of the peptide-surface potential energy (PE) with the surface energy must pass through the origin, the slope of this variation for a conformation on a surface of energy E_s, is given by $S_{ps} = E_{ps}/E_s$. Thus, the difference between the slopes for two conformations, A and B, that are adsorbed on the same solid surface is given by $\Delta S_{ps} = \left(E_{ps}^{\mathrm{A}} - E_{ps}^{\mathrm{B}}\right)/E_s$. As it does not matter which solid surface is involved, it is sufficient to say that the change in slope scales with the change in the peptide-surface PE of the two conformations on the same solid surface, which is denoted by $\Delta S_{ps} \sim \left(E_{ps}^{\mathrm{A}}\big|_s - E_{ps}^{\mathrm{B}}\big|_s\right) = \Delta E_{ps}\big|_s$. Thus, the irregular variation of ΔS_{ps} with the number of residues arises out of the differences in the way $E_{ps}\big|_s$ varies with the number of residues for the two conformations. The origin of these differences is: (1) the fractional periodicities of the three helices, and (2) the disparity between these

periodicities. The fractional periodicity combines with the discreteness of n to yield a complex variation of the peptide-surface PE for each helix. This complexity is then compounded when the difference between them is taken across a switch.

Figure 5.10 shows the variation of the longitudinal strain of the peptide with the number of residues for the two conformational switches. Whilst, as expected, the strain for the $\alpha \rightarrow 3.1_{10}$ switch is less than that of the $3.1_{10} \rightarrow 2_7$ switch, the change in strain with number of residues is qualitatively similar for both. Although the strain tends to increase with the number of residues, as with the switching surface energy, the change is complex, with rises in the ranges $n = 6-8$ and $n = 10-11$ being followed by shallow dips between $n = 9-10$ and $n = 12-13$ respectively. Figure 5.11 shows that the complexity comes from the fact that the length of the gas phase α-helix – which is the reference conformation when evaluating the strain – varies with the number of residues (a similar but very much weaker dependence is also observed for the 3.1_{10} and 2_7 helices on the solid surface). This is due to the rise in the strength of the collective dipole (Ripoll and Scheraga, 1988; Park and Goddard, 2000) – which acts to shorten the helix – and the degree of completeness of the helix, which is responsible for the local irregularity.

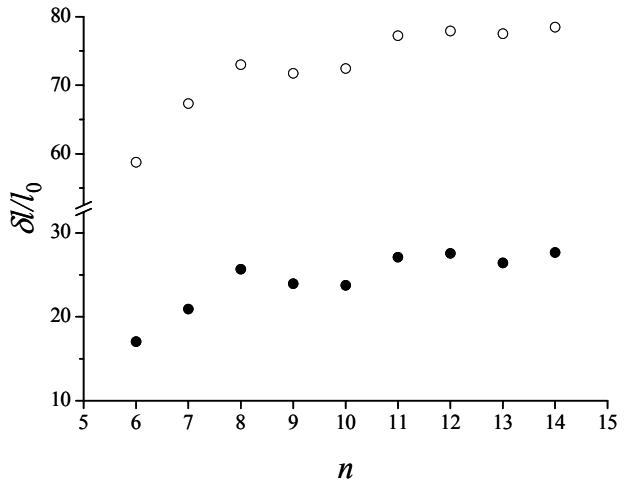


Figure 5.10 Variation of longitudinal strain with number of alanine residues for the $\alpha \rightarrow 3.1_{10}$ (closed circle) and $3.1_{10} \rightarrow 2_7$ (open circle) switches. The strain is measured relative to the gas phase α-helical conformation.

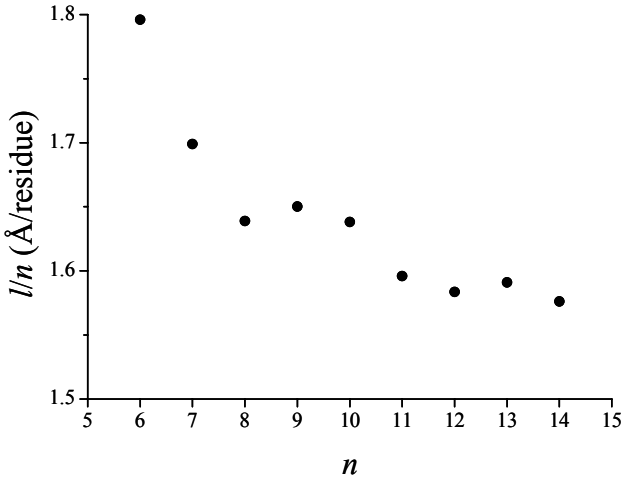


Figure 5.11 Variation of peptide length per residue with number of alanine residues for the gas-phase α-helix.

5.3.4. General Discussion

It is clear from the results above that hydrogen bonding is essential to the manifestation of the observed conformational switching. As many of the amino acids are capable of supporting hydrogen bonds, similar switching could perhaps be expected in other peptides and proteins. However, although experimental observation of switching such as that observed here is likely to be difficult to detect, lack of reports in the literature suggests that hydrogen bonding is not the only requirement. We hypothesize that a further essential requirement is a high peptide symmetry such as that obtained in a homopeptide like polyalanine, which would induce the entire peptide to switch at specific surface energies rather than in sub-elements over ranges of surface energies. Our results also suggest that switching may well be restricted to smaller peptides. These hypotheses will be tested by us in future work.

Braun and co-workers provide good arguments as to why the smooth surface used here is a satisfactory model for the study of charge-neutral molecules like polyalanine on metal surfaces (Braun et al., 2002), whilst this surface has also been widely used in the study of adsorption of uncharged molecules on the graphite basal plane (Cracknell et al., 1995; Nicholson, 1996; Bandosz et al., 2003). There are, however, some surfaces where corrugations are significant such as, for example, the armchair and other non-basal surfaces of graphite (shown, for example, in Figure 11 in (Biggs et al., 2004)). Given that the switching phenomenon observed here arises

from the symmetry of the molecule and its ability to support hydrogen bonds – the surface simply provides the energy for switching – such corrugations are unlikely to destroy either the switching or the structures observed. However, depending on the exact nature of the corrugations it may be expected that they will have some effect on the switching energies. The surface representation used here is also unlikely to be good for metal oxides and other surfaces that may act to subvert the intra-peptide hydrogen bonding that stabilize the structures observed here – future work will seek to investigate this issue.

The effect of entropy has been ignored in this study. Its inclusion is unlikely to destroy the switching phenomenon, however, as all three conformations observed here are known to exist in real proteins (Pervushin and Arseniev, 1992; Solov'yov et al., 2006), albeit at substantially different levels, whilst various theoretical studies have established the free energy minima associated with the α- and 3_{10}-helical conformations (Clark et al., 1991; Tirado-Rives et al., 1993; Huston and Marshall, 1994; Zhang and Hermans, 1994). The entropic effect is, however, likely to modify the switching surface energies. In particular, previous work on polyalanine shows that there is a greater entropic stabilization of the 3_{10}-helix relative to the α-helix, and this would be even more so for the more extended 2_7-helix. The entropic effect is, therefore, likely to depress the switching surface energies relative to those predicted here. Future work will seek to elucidate the entropic contributions further for polyalanine on the solid surface.

Effects of the entropy will be the most intensive in the vicinity of switching points, where the potential energy difference between two conformations is the lowest. This is somewhat limiting factor for the possible future applications of the conformational switching effect as any device based on the switching phenomenon would have to operate on very low temperatures in order to disable random transitions between the conformations. Moving away from the switching points, however, stabilises the adsorbed conformations as the potential energy difference between them increases, as shown in the insert of Figure 5.7(a). Thus, operating in the surface energy range away from the switching points would reduce the effects of entropy and allow application of the conformational switching phenomenon on higher temperatures.

Previous work has shown for polyalanine in the bulk phase that the relative stabilities of the α- and 3_{10}-helical conformations can be changed with the nature of the solvent, the end groups and the presence of ligands (Clark et al., 1991; Smythe et al., 1993; Zhang and Hermans, 1994). These observations could be exploited to tailor the switching surface energies for polyalanine on a solid surface, or provide a means of instigating a conformational switch without changing the nature of the solid surface. These possibilities will also be investigated in future work.

It is perhaps worthwhile mentioning some potential implications of the work reported here. Conformational switching is of relevance in molecular electronics (Rambidi, 2003) and protein-based computer memories (Birge, 1992), both of which are part of an ongoing quest to build a biomolecular computer. Molecular switches are also an important molecular mechanical element that underpins molecular rotors, brakes and motors (Feringa, 2001; Kelly, 2001). Stretches of polyalanine are very common in natural proteins and have been implicated in some diseases (Albrecht and Mundlos, 2005) – the work here raises the possibility that solid surfaces may be able to stabilize in these stretches non-native conformations such as observed here, which may have implications for their biological activity and function.

5.4. Conclusions

Using an *ab initio* structure prediction approach, we have discovered a conformational switching phenomenon for polyalanine on solid surfaces – the peptide undergoes step changes in its conformation at specific surface energies that vary in a complex manner with the peptide size. Two conformational switches were observed: (1) α-helix→3.1_{10}-helix, and (2) 3.1_{10}-helix→2_7-helix. The first always occurs at lower surface energies than the second. All three structures are characterized by hydrogen bonding – it is this hydrogen bonding and, we hypothesize, the symmetry of the homopeptide that leads to the conformational switching rather than gradual change in the structure.

Whilst all the conformational characteristics of the peptide-solid surface system undergo some step change at the switching points, the backbone dihedral angles, number of residues per turn, and strain along the peptide axis experience the most significant changes. The strain in particular sees significant changes that could well be exploited. Although the various components of the potential energy (PE) also

undergo step changes at the switches, the total energy of the peptide-solid system undergoes a continuous change, with the discontinuity being restricted to its derivative. By making some well founded assumptions, a simple expression (equation (5.4)) for the switching surface energy was obtained that can be used to estimate the switching surface energies with a small number (as little as three) simulations.

Whilst the conformational switching was observed in all the polyalanine molecules from $n = 6$ to $n = 14$ residues, the surface energy at which the switches occur and the associated longitudinal strain vary in a complex manner with the number of residues. These complex variations arise from the fractional periodicity of the helices and the disparities between these periodicities.

Although the effect of entropy has not been included here, results from gas and solution phase simulations as well as experimental evidence suggest that entropy will not destroy the switching effect. Entropy will, however, most likely lead to a reduction in the switching surface energies predicted here. Previous simulation work additionally suggests that the presence of solvents will also not destroy the switching but, rather, offer a route for tailoring the switching surface energies.

Chapter 6. Investigation of Coupling of Langevin Dipole Method with Amber PE Model

6.1. Introduction

Results of the studies presented in Chapters 4 and 5 have demonstrated the ability of an EA based approach to predict protein conformation in a gas phase and on the gas-solid interfaces. Whilst being very useful in theoretical studies and for method testing, gas phase simulations do not have many practical applications in studies of proteins. For most biomolecules, solution in water is a much more common environment. Water is believed to play an essential role in protein folding (Barron et al., 1997; Xu and Cross, 1999) as well as in the behavior of proteins in non-native environments, including solid-liquid interfaces (Mungikar and Forciniti, 2004; Carravetta and Monti, 2006), which is of particular relevance here.

As indicated in Chapter 2, the Langevin dipole (LD) model of water (Florián and Warshel, 1997) has been shown to be both fast and accurate. One of its disadvantages, however, is the need to obtain solute atomic charges from quantum mechanical (QM) methods which are computationally very expensive and, hence, inapplicable in an EA based approach. We have, therefore, investigated the possibility of replacing the expensive QM charge calculation with a set of static charges adopted from the Amber PE model, thus creating a modified model that we have termed LD-Amber. The systems used to test the LD-Amber method are defined first, along with the description of the method. The results of the LD-Amber application in prediction of solvation free energies of amino acid side chain analogues and alanine dipeptide are then presented in detail. Finally, the performance of the method is compared to traditional molecular dynamics approach.

6.2. Study Details

6.2.1. Solvation Free Energies of Amino Acid Side Chain Analogues

Amino acid side chain analogues obtained by replacing the backbone atoms of the α-amino acids with a hydrogen atom are commonly used as a basis for testing solvent models (Wolfenden et al., 1981; Edsall and McKenzie, 1983; Ben-Naim, 1990; Avbelj, 2000). The first part of this study was, therefore, focused on comparing the LD-Amber-derived free energies of solvation of the amino acid side chain analogues in Table 6.1 with published experimental and theoretical data.

Table 6.1 Amino acid side chain analogues considered in this study

Amino acid (code)	**Side chain analogue at pH 7**
Alanine (ala)	Methane
Arginine (arg)	N-propylguanidinium
Asparagine (asn)	Acetamide
Aspartate (asp)	Acetate ion
Cysteine (cys)	Methanethiol
Glutamate (glu)	Propionate ion
Glutamine (gln)	Propionamide
Histidine (his)	Methylimidazolium
Isoleucine (ile)	Butane
Leucine (leu)	Isobutane
Lysine (lys)	N-butylammonium
Methionine (met)	Methylethylsulfide
Phenylalanine (phe)	Toluene
Serine (ser)	Methanol
Threonine (thr)	Ethanol
Tryptophan (trp)	3-Methylindole
Tyrosine (tyr)	P-cresol
Valine (val)	Propane

It may be noted from Table 6.1 that the sidechain analogues for glycine and proline were not considered. In the former case, its sidechain analogue is molecular hydrogen, which is of little interest here because the hydrogen atoms carry no charge. The proline sidechain analogue was omitted because the charge distribution on what is nominally a propane structure is highly non-physical due to the cyclic sidechain of proline being connected to backbone atoms of very different electronegativity.

6.2.2. Free Energy Surface of Alanine Dipeptide in Neutral Water

In some contexts it is important to be able to correctly model the free energy surface (FES) rather than just the energy of solvation for a native conformer – the most obvious example is in the *ab initio* structure prediction context, where search methods probe the FES extensively in search for the global minimum. The second part of the study was, therefore, focused on comparing the characteristics of the LD-Amber-based FES of the alanine dipeptide in neutral water with published experimental and theoretical data.

The alanine dipeptide, which is a single alanine residue capped by acetate and amino-methyl groups on the N- and C-termini respectively (i.e. AcAlaNHMe), was selected for study here for a number of reasons. Principally, its small size – its structure can be defined in terms of just two dihedral angles – makes it possible to thoroughly probe its FES without excessive computational effort. However, as will be seen, its use in the parameterization of the Amber potential model (Cornell et al., 1995) also aids in better understanding any deficiencies revealed by our analysis. Whilst the even simpler glycine dipeptide could have been used instead for the same reason, conformational analysis would have been complicated by significant solute entropic contributions arising from its small side chain (Rappé and Casewit, 1997).

6.2.3. Electrostatic Potential Field and Water Structure Around Alanine Dipeptide

Previous theoretical work suggests hydrogen bonding networks involving the solute and solvent play a role in stabilizing solute structures (Mezei et al., 1985; Beglov and Roux, 1995). More recent theoretical work also suggests that moderation of the intrasolute electrostatic interactions by the solvent also influences solute structure (Drozdov et al., 2004). As these phenomena are dependent on the heterogeneous solvent structure within and immediately around the solute both at a local level (i.e. molecule-molecule) and over longer ranges in the form of bridges, for example, it is reasonable to suppose that accurate determination of stable conformers is dependent on correctly predicting the structure of the solvent. Whilst the LD method cannot say anything directly about the structure of the solvent because the dipoles are constrained to a regular lattice, it does predict the electric field at a local level. We have investigated this issue by comparing the electrostatic potential field

obtained from the LD-Amber approach for a particular conformation of the alanine dipeptide with that obtained from an MD simulation.

6.2.4. Computational Peformance

One of the main motivations for using the LD method is its speed. There is, however, little quantitative information available on the computational expense of the method and how it compares with competitor explicit approaches. The computational expense of the LD-Amber approach is, therefore, compared with traditional explicit approaches.

6.3. Methodology

6.3.1. LD-Amber Method

As illustrated in Figure 6.1, the volume around the solute in the LD approach (Florián and Warshel, 1997) is divided into a number of distinct regions. The solvent is absent in the first of these regions, which is located within the surface defined by the van der Waals radii of the solute atoms, σ. The solvent in the volume located between this van der Waals surface and the surface $R_O(\mathbf{x})$, Figure 6.1, is modeled by Langevin dipoles located at the nodes of cubic grids, whilst beyond this outer surface the solvent is treated using a continuum approximation. The outer surface, $R_O(\mathbf{x})$, is defined by the set of nodes where the electric field arising from the solute falls below a threshold, ξ_O, where this field at a node-*j* is given by (Florián and Warshel, 1997)

$$\boldsymbol{\xi}_j = \sum_i \frac{Q_i \mathbf{r}_{ij}}{\varepsilon_{ij} r_{ij}^3} \tag{6.1}$$

where Q_i are the atomic charges associated with the solute, $\mathbf{r}_{ij}$ and r_{ij} are the displacement between the atomic charge-*i* and node-*j* and its magnitude respectively, and ε_{ij} is the screening function given by (Florián and Warshel, 1997)

$$\varepsilon_{ij} = \frac{\sqrt{2 + r_{ij}}}{1.7} \tag{6.2}$$

where the magnitude of the displacement is in angstroms. As the electric field gradients near the solute are in general large, a fine cubic grid of spacing a_f is used between the van der Waals surface and the surface defined by the distance $\sigma + \delta$ from the solute atoms. Beyond this, a coarser grid spacing, $a_c > a_f$, is used, which

aids the computational efficiency of the method. Dipoles are discarded from nodes of this coarser grid if they fall within a distance $\left(a_c + a_f\right)/2$ of a dipole on the finer grid.

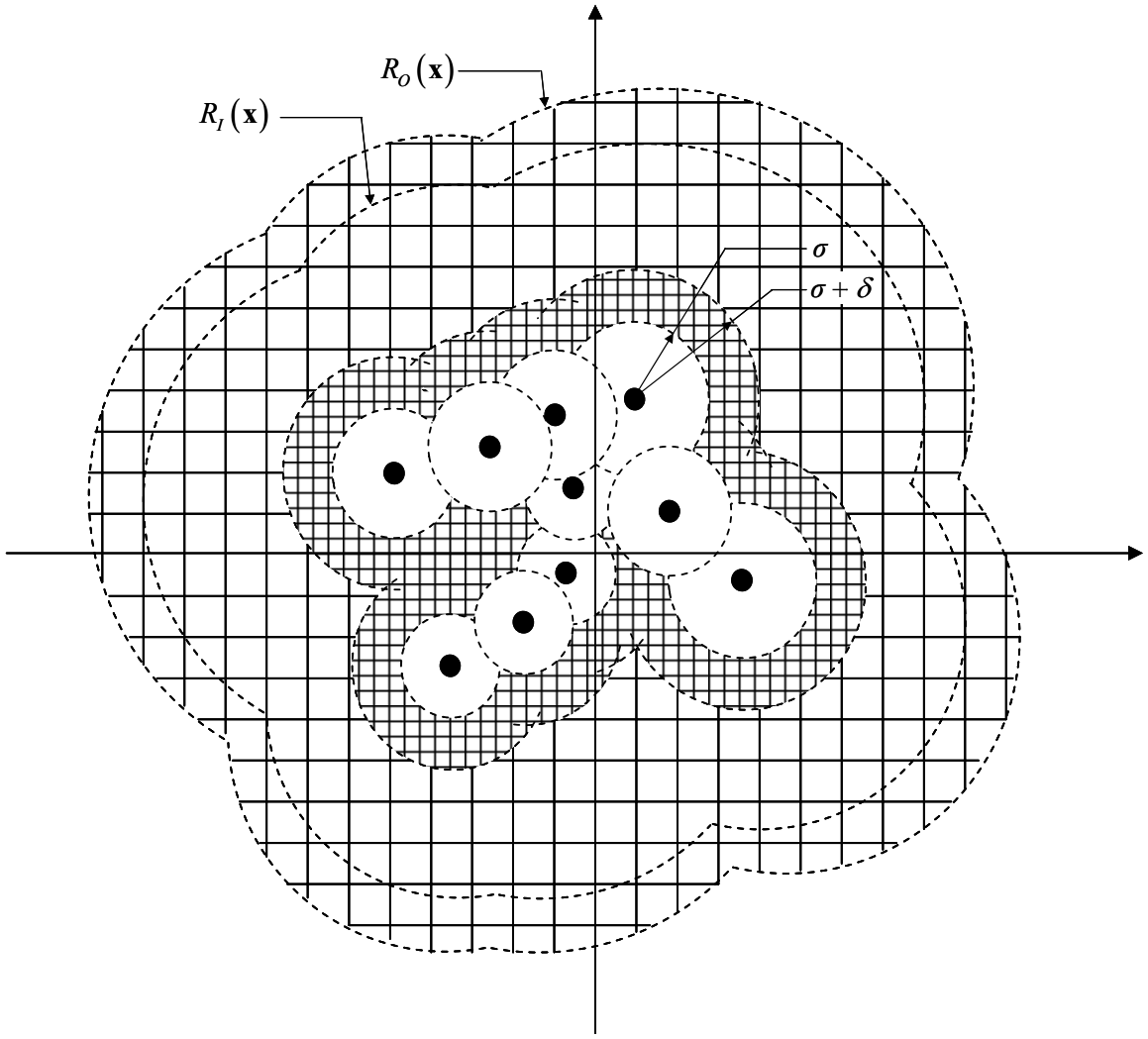


Figure 6.1 Schematic showing the five different regions around the solute in terms of solvent treatment. The solvent is excluded from the innermost region (white). The solvent in the two regions beyond this is modeled by Langevin dipoles located at the nodes of fine and coarse grids with their magnitudes and directions being determined iteratively in a self-consistent manner under the influence of the solute and the fixed dipoles located in the fourth region between the surfaces $R_I(\mathbf{x})$ and $R_O(\mathbf{x})$. The solvent in the fifth region beyond the surface $R_O(\mathbf{x})$ is modeled using a continuum approximation.

The orientation and magnitude of the dipoles in the volume $R_I(\mathbf{x}) < \mathbf{x} < R_O(\mathbf{x})$ are determined from equations (6.1) and (6.2), where the surface $R_I(\mathbf{x})$ is defined by the set of nodes where the electric field arising from the solute falls below a second

threshold, $\xi_I > \xi_O$. The orientation and magnitude of the remaining dipoles are determined iteratively in a self consistent manner under the influence of the solute and the fixed dipoles beyond $R_I(\mathbf{x})$ as described by Florián and Warshel (Florián and Warshel, 1997). The dipole j is allowed to polarise by changing its orientation and magnitude in the direction of the of the total electrostatic field, $\boldsymbol{\xi}_j$, calculated as a sum of the electrostatic field of the solute and of the neighbouring dipoles

$$\boldsymbol{\xi}_j = \boldsymbol{\xi}_j^0 + \sum_{k \neq j} \frac{3\left(\mathbf{r}_{jk} \circ \boldsymbol{\mu}_k^{(n-1)}\right)\mathbf{r}_{jk} - r_{jk}^2 \boldsymbol{\mu}_k^{(n-1)}}{r_{jk}^5} \tag{6.3}$$

where $\mathbf{r}_{jk}$ is the position vector that connects dipoles j and k and r_{jk} is its magnitude. $\boldsymbol{\mu}_k$ is the neighbouring dipole vector k, while the superscript (n-1) denotes that its value is taken from the previous iterative step. $\boldsymbol{\xi}_j^0$ is the electrostatic field of the solute at position of dipole j and its value remains constant during a single iteration procedure. Electrostatic field of the solute is calculated from the unscreened solute charges

$$\boldsymbol{\xi}_j^0 = \sum_i \frac{Q_i \mathbf{r}_{ij}}{r_{ij}^3} \tag{6.4}$$

Electrostatic field calculated in equation (6.3) is used as a basis for calculating the magnitude of the dipole j in the current, n^{th} iteration. The new magnitude of the dipole j is obtained using the Langevin function, L(x)

$$\mu_j^{(n)} = \mu_0 L\left(x\right) \tag{6.5}$$

where μ_0 is the magnitude of dipole j at saturation (i.e. its maximal magnitude – 0.05 and 0.26 $\mathrm{e}\cdot\text{Å}$ for finer and coarser grids, respectively), while the Langevin function is defined as

$$L\left(x\right) = \coth\left(x\right) - \frac{1}{x} \tag{6.6}$$

x is a compound term defined as

$$x = \frac{\mu_0 \xi_j}{k_B T} \tag{6.7}$$

where ξ_j is the magnitude of the electrostatic field at point dipole j. k_B in the denominator of the equation is the Boltzmann constant, while T is the absolute

temperature of the system. The orientation of the dipole j is assumed to be the same as the orientation of the electrostatic field ξ_j at the n^{th} iteration.

Equation (6.7) expresses the effect of the solvent temperature on the energetics of solvation process. The Langevin function shown in equation (6.6) asymptotically tends to 1 as x tends to infinity (i.e. for very low temperatures). As the temperature is reduced, the dipole fluctuation is decreased, thus increasing their effective magnitude in the direction of the electrostatic field, as expressed by equation (6.5). Increase in the temperature, on the other hand, increases the dipole fluctuations, thus reducing their effective magnitude. The effect of the temperature on the magnitude of the Langevin dipole is shown in Figure 6.2.

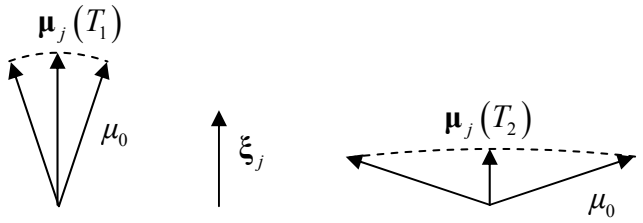


Figure 6.2 Effective magnitude of the Langevin dipole changes with the temperature. Lower temperatures (T_1) reduce fluctuations and increase the effective magnitude, while higher temperatures (T_2) promote dipole fluctuations, thus reducing its effective magnitude. In both cases, the effective orientation of the dipole j is in the direction of the electrostatic field ξ_j.

Knowing the dipole vectors, the free energy that arises from the interaction between the solute and that part of the solvent modeled by the dipoles is determined straightforwardly by assuming the solvent responds in a linear manner to the solute (i.e. linear response theory) (Florián and Warshel, 1997)

$$\Delta G_{\mathrm{es(LD)}} = 722.5 \sum_{j=1}^{N} \boldsymbol{\mu}_j \circ \boldsymbol{\xi}_j^0 \tag{6.8}$$

where N is the total number of the dipoles and coefficient 722.5 is applied to obtain the value in kJ/mol. The value of the coefficient implicitly includes energy needed to polarise the solvent molecules, as determined by Florián and Warshel (1997).

Fixing the dipoles to a lattice means the free energy change due to the dipoles varies with the position of the solute relative to the grid. It is, therefore, necessary to average the free energy change obtained from a small number of simulations, S, in which the position of the solute molecule is randomly generated within the central

cell of the fine grid. It should be noted that changes from the gas-phase solute structure due to solvation are sometimes included by coupling the internal degrees of freedom of the solute and the solvent in a self-consistent manner. However, inline with common practice, this is not done here as it involves an iterative process.

The total free energy of solvation of the solute, ΔG_s, is obtained by adding to the average free energy change arising from the volume modeled by Langevin dipoles, $\Delta G_{\mathrm{es(LD)}}$, the free energy change due to the electrostatic interaction between the solute and solvent volume beyond the surface $R_O(\mathbf{x})$, $\Delta G_{\mathrm{es(c)}}$, the van der Waals interactions between the solute and dipoles, ΔG_{vdW}, the interaction between the solvent and non-polar part of the solute surface (i.e. hydrophobic contributions), ΔG_{phob}, and solute polarization due to the solvent, ΔG_{pol} (Florián and Warshel, 1997)

$$\Delta G_{\mathrm{s}} = \Delta G_{\mathrm{es(LD)}} + \Delta G_{\mathrm{es(c)}} + \Delta G_{\mathrm{vdW}} + \Delta G_{\mathrm{phob}} + \Delta G_{\mathrm{pol}} \tag{6.9}$$

The models proposed by Florián and Warshel for all but the solute polarization term were used here unchanged.

The contribution of implicitly represented solvent, $\Delta G_{\mathrm{es(c)}}$, is calculated using Born's formula if the solute is charged (Florián and Warshel, 1997)

$$\Delta G_{\mathrm{es(c)}} = -695\left(1-\frac{1}{\varepsilon_r}\right)\frac{Q^2}{R} \tag{6.10}$$

where Q is the net charge of the solute (in e), while R is the average radius of the domain represented with explicit dipoles (in Å). ε_r is the relative dielectric constant and the coefficient 695 is applied to express the energy in kJ/mol. If the solute carries no net charge, its total dipole moment μ is used to calculate the contribution of continuous solvent (Florián and Warshel, 1997)

$$\Delta G_{\mathrm{es(c)}} = -695\frac{2\varepsilon_r - 2}{2\varepsilon_r + 1}\frac{\mu^2}{R^3} \tag{6.11}$$

The van der Waals energy of solvation is expressed through a 9-6 interaction term (Florián and Warshel, 1997)

$$\Delta G_{\mathrm{vdW}} = k_{\mathrm{vdW}}\sum_{i,j} C_i N_j \left[2\left(\frac{r_i^*}{r_{ij}}\right)^9 - 3\left(\frac{r_i^*}{r_{ij}}\right)^6\right] \tag{6.12}$$

where k_{vdW} is an empirical parameter equal to 3.5 kJ/mol, and r_i* and C_i are the radius and the London coefficient of the solute atom i, while N_j is the normalisation

factor whose purpose is to scale down the strength of the interactions with the dipoles of the denser fine grained grid

$$N_j = \left(\frac{a_j}{a_c} \right)^3 \tag{6.13}$$

where, as above, a_c is the node distance in the coarse grid while a_j is the node distance of the grid to which the dipole j belongs.

The hydrophobic term, ΔG_{phob}, represents the energy invested in the formation of the solvent cavity. It is proportional to the number of the Langevin dipoles on a distance less than 1.5 Å from the van der Waals surface of the solute atoms

$$\Delta G_{\text{phob}} = k_{\text{phob}} \sum_j f\left(\Phi_j\right) \tag{6.14}$$

where k_{phob} is an empirical parameter equal to 0.050 kJ/mol, while Φ_j is the electrostatic potential calculated at the position of the dipole j. f is a complex function of the electrostatic potential

$$f\left(\Phi_j\right) = \begin{cases} 1, & \left|\Phi_j\right| \le \Phi_{\min} \\ 1 - \dfrac{\left|\Phi_j\right| - \Phi_{\min}}{\Phi_{\max} - \Phi_{\min}}\left(1-\chi\right), & \Phi_{\min} < \left|\Phi_j\right| < \Phi_{\max} \\ \chi, & \Phi_{\max} \le \left|\Phi_j\right| \end{cases} \tag{6.15}$$

where $\Phi_{\min}$ = 0.002 e/Å, $\Phi_{\max}$ = 0.015 e/Å and χ = 0.08 are all empirical parameters (Florián and Warshel, 1997).

Solute polarization was ignored as the model of Florián and Warshel for ΔG_{pol} requires either access to high level QM results (indeed, even higher than those used to determine the atomic charges) or empirical data for the solute, neither of which are desirable in the contexts of interest here. This neglect of solute polarization is in part justified by the fact that the QM method used to determine the Amber atomic charges over-predicts the gas-phase molecular dipole moment by 10-20%, effectively mimicking the polarization effect in some approximate meanfield way (Cornell et al., 1995; Florián and Warshel, 1997).

A variety of parameters are required to be specified before simulations can be done, including the magnitude of the dipoles on the fine and coarse grids at saturation, $\mu_{0,f}$ and $\mu_{0,c}$ respectively. As already indicated, the atomic charges, Q_i, were taken from the *ff*94 Amber parameter set (Cornell et al., 1995). This use of the

Amber charges in principle requires us to determine the remaining parameters afresh rather than use those of Florián and Warshel (Florián and Warshel, 1997). Such a re-parameterization is, however, contrary to the spirit of the approach being investigated here and we, therefore, used the parameters of Florián and Warshel as summarized in Table 6.2 – part of the motivation for this study is, of course, to determine how good this approximation is.

Table 6.2 LD parameter values used in the work reported here beyond the atomic charges, which were taken from the *ff*94 Amber parameter set (Cornell et al., 1995)

Parameter	**Value**[a]
a_f	1 Å
a_c,	3.1043 Å
δ	2 Å
ξ_I	0.0021 e/Å^2
ξ_O	0.0015 e/Å^2
$\mu_{0,f}$	0.05 eÅ
$\mu_{0,c}$	0.26 eÅ
S	10[b]
$\sigma_{C(sp^3)}$	2.65 Å
$\sigma_{C(sp^2)}$	3.00 Å
$\sigma_{O(sp^3)}$	2.20 Å
$\sigma_{O(sp^2)}$	2.65 Å
σ_N	2.65 Å
σ_S	3.20 Å
σ_H	c

a. Unless indicated otherwise, all values are taken from Florián and Warshel (Florián and Warshel, 1997).

b. Although this value is smaller than that used by Florián and Warshel, experimentation showed this to be sufficient for accurate results.

c. The van der Waals radius of hydrogen is determined using $\sigma_H = k\sigma_h$, σ_h is the radius of the nearest heavy atom, and k is a constant that takes a value of 0.88 or 0.78 when the heavy atom is in the first or second row of the periodic table respectively.

The van der Waals parameter for the oxygen atoms of the carboxyl group of the aspartate and glutamate side chains presented a problem. In the neutral form of these side chains, the oxygen atom in the -OH group is sp^3 hybridized whilst that in –C=O is sp^2 hybridized. In neutral water, on the other hand, deprotonation occurs (Rappé and Casewit, 1997) to leave behind a spare electron that is delocalized over

the two oxygen atoms; in this case the oxygen atoms are neither sp^2 nor sp^3 hybridized (Pauling, 1940). Whilst Florián and Warshel provide van der Waals radii for interactions between the dipoles and both sp^2 and sp^3 hybridized forms of the oxygen atom (Florián and Warshel, 1997), they do not provide parameters for the resonant case. Tests were, therefore, undertaken as part of this work to determine the most appropriate parameters. These tests revealed that the sp^3 oxygen parameter of Florián and Warshel lead to acceptable results – they were, therefore, used for the work reported here.

6.3.2. Generation of Solvation Free Energies of Amino Acid Side Chain Analogues

The solvation energy for each of the side chain analogues in Table 6.1 was generated using the LD-Amber approach as described in §6.3.1. The analogue structures were derived from the acetyl and amino-methyl capped dipeptide of the associated amino acids using a two-stage process. In the first stage, 1296 structures obtained by systematically varying the backbone dihedral angles of the dipeptide in 10° increments over the range of [–180°, 180°] were locally relaxed using the algorithm of Davidon (Davidon, 1975; Ponder, 2004) with an RMS gradient cutoff criterion of 0.01 Å; all other initial angles and bond lengths were defined by the Amber *ff*94 parameter set (Cornell et al., 1995). The analogue structure used in the LD-Amber simulations was then obtained by replacing the backbone atoms (including the caps) of the lowest energy member of the set of 1296 locally relaxed structures with an H atom carrying a charge equal to that of the other H_β atoms, and then subtracting the excess charge from the C_β atom to ensure charge neutrality (Shirts et al., 2003).

6.3.3. Generation of Free Energy Surface of Alanine Dipeptide in Neutral Water

As direct determination of the free energy difference between conformer C_1 and conformer C_2 in solution, ΔG^s_{12}, is computationally very demanding, it is sometimes determined via the thermodynamic cycle illustrated in Figure 6.3 (Ben-Naim, 1990), which gives

$$\Delta G^s_{12} = \Delta G^g_{12} + \Delta G_{s(2)} - \Delta G_{s(1)} \tag{6.16}$$

where $\Delta G_{s(i)}$ is the free energy of solvation of the conformer C_i, and ΔG_{12}^{g} is the free energy difference between the conformers in the gas phase. Whilst determination of ΔG_{12}^{g} is computationally less expensive than its solution-phase counterpart, it is still a non-trivial exercise and approximations are, therefore, often made. These approximations may be understood by considering the two aspects of the free energy difference between the two conformers in the gas phase at a temperature T

$$\Delta G_{12}^{g} = \Delta U_{12}^{g} - T\Delta S_{12}^{g} \tag{6.17}$$

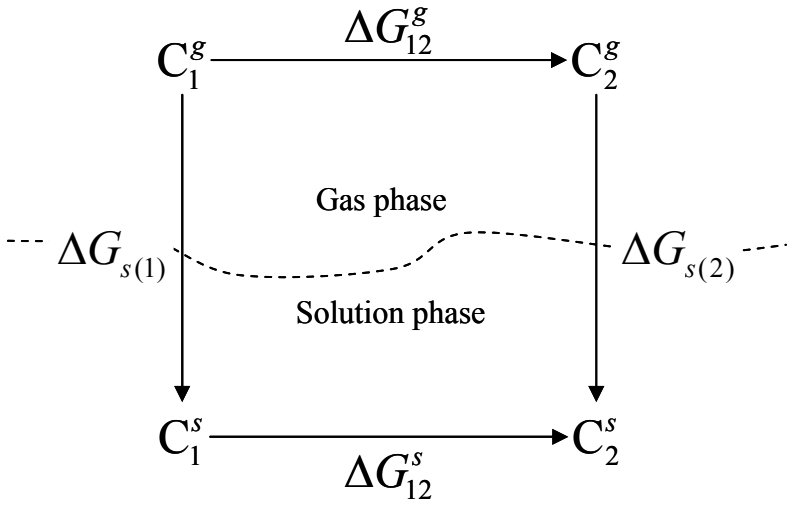


Figure 6.3 The thermodynamic cycle commonly used to determine the difference in free energies of two conformers C_1 and C_2 in solution, ΔG_{12}^{s}, given knowledge of the free energy of solvation of the two conformations, $\Delta G_{s(i)}$, and the free energy difference between the two conformers in the gas phase, ΔG_{12}^{g}.

The potential energy component, ΔU_{12}^{g}, can be evaluated very easily from knowledge of the three-dimensional structure of the two conformers and a potential energy model. The solute entropic component, $T\Delta S_{12}^{g}$, on the other hand is more difficult to determine and is, hence, often neglected to give

$$\Delta G_{12}^{s} \approx \Delta U_{12}^{g} + \Delta G_{s(2)} - \Delta G_{s(1)} \tag{6.18}$$

This approximation is made here.

Inline with common practice, the FES of the alanine dipeptide was determined as a function of the two backbone dihedral angles, ϕ and ψ, shown in Figure 6.4. The gas-phase PES over these angles was obtained from the Amber model using the associated *ff*94 parameter set (Cornell et al., 1995) by evaluating the potential energy

of the 32400 structures obtained by varying these two angles in 2° increments over the range [−180°, 180°]. All bond lengths and angles, including the side chain dihedral angle, were kept fixed at their corresponding values obtained by locally relaxing the alanine amino acid structure from the equilibrium values specified by the Amber *ff*94 parameter set (Cornell et al., 1995).

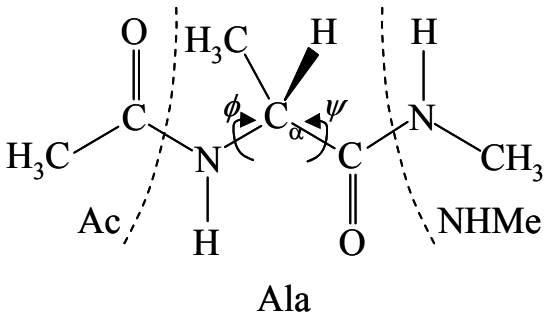


Figure 6.4 Alanine dipeptide structure showing the two dihedral angles that define its backbone conformation.

As indicated by equation (6.18), the FES of the alanine dipeptide in neutral water was determined by adding to the gas phase PES the free energies of solvation. These were determined for the 32400 conformers of the dipeptide using the LD-Amber approach as described in §6.3.1.

6.3.4. Generation of Electrostatic Potential Field from the LD-Amber Approach and MD

Given a set of n charges, Q_i, and m dipoles, $\boldsymbol{\mu}_i$, at positions $\mathbf{r}_i$ relative to $\mathbf{r}$, the electrostatic potential at that position, $\Phi(\mathbf{r})$, can be determined by

$$\Phi(\mathbf{r}) = \frac{1}{4\pi\varepsilon_0}\left[\sum_{i=1}^{n}\frac{Q_i}{r_i} + \sum_{i=1}^{m}\frac{\boldsymbol{\mu}_i \circ \mathbf{r}_i}{r_i^3}\right] \tag{6.19}$$

where r_i is the magnitude of the position vector, and ε_0 is the permittivity of free space. This expression was used to determine the electrostatic potential field (EPF) around the α_L conformer of the alanine dipeptide as identified from the second part of this study using solvent configurations obtained from LD-Amber and MD simulations.

The EPF for the LD-Amber approach was determined by averaging over the fields obtained from the final configurations of $S = 10$ LD-Amber simulations undertaken as described in §6.3.1.

The EPF field for the MD approach was determined by averaging over 1000 snapshots taken at evenly spaced intervals from an MD simulation of 50,000 timesteps of size 2 fs. The MD simulation was done in the canonical ensemble using the algorithm of Berendsen and co-workers (Berendsen et al., 1984) as implemented in the Tinker code (Ponder, 2004). The solute and water molecules, which were modeled using the TIP3P molecule (Jorgensen, 1981), were treated as rigid bodies. The solute-water interactions were modeled with the Amber potential with the associated *ff*94 parameter set (Cornell et al., 1995). A cutoff radius of 9 Å was used for all van der Waals interactions, while electrostatic interactions were evaluated using particle mesh Ewald method (Darden et al., 1993) with the same cutoff. The initial state of the MD simulation was generated by placing the solute molecule centrally into a cubic volume and then inserting water molecules into volume using grand canonical MC simulation with a chemical potential corresponding to a bulk water density of 1000 kg/m^3 and a temperature of 298 K. The linear dimension of the volume was set to 30 Å more than the largest dimension of the solute. A total of 1682 water molecules were simulated, which represented more than four complete hydration layers around the solute. The system obtained from the MC simulation was relaxed for 2500 MD timesteps before being used for production purposes.

6.3.5. Comparison of Computational Performance

The computational cost of the LD-Amber approach for determining the free energy difference between two conformations in solution is compared with that of traditional explicit approaches based on use of thermodynamic integration or similar strategies with a series of molecular dynamic (MD) simulations along a reaction coordinate between the two conformations (Anderson and Hermans, 1988; Tobias and Brooks, 1992; Chipot and Pohorille, 1998; Smith, 1999b). The computational expense of this approach depends on a number of issues – to ease comparison, the timing for a single MD simulation was scaled using conservative assumptions about these issues so as to give a reasonable best case estimate for the computational expense of the traditional explicit approach. Because timings are very dependent on the machine used amongst other things, all MD and LD-Amber simulations were performed using the same machine based on an AMD Athlon MP 1.4 GHz CPU with 2 GBytes of RAM and running under Linux.

6.4. Results and Discussion

6.4.1. Solvation of Amino Acid Side Chain Analogues

The free energy changes for transfer of the amino acid side chain analogues from the gas phase to water at pH 7 estimated by the LD-Amber approach are given in Table 6.3 along with the constituent parts. These predicted solvation energies are generally inline with the hydrophobicity of the associated amino acids, bearing in mind that there are many ways in which this can be defined (Cornette et al., 1987). Analysis of the contributions to the solvation energy indicates that the change in free energy arising from the electrostatic and van der Waals interactions is in all cases offset by a decrease in the free energy of the water as reflected in the consistently positive hydrophobic contribution. Inline with their largely non-polar character, the analogues of Ala, Ile, Leu and Val are characterized by some of the largest hydrophobic contributions and smallest electrostatic contributions. Also as expected, the electrostatic contributions from the five charged analogues are an order of magnitude greater than the closest neutral analogues, and are the only analogues to see a significant contribution from the volume beyond that modeled by the Langevin dipoles. Amongst the neutral analogues, those of Asn, Gln, Ser and Thr are characterized by some of the largest electrostatic contributions and smallest hydrophobic contributions, inline with their known polar character. A similar balance between the contributions could also perhaps be expected for the analogue of Tyr given its known polar character – this somewhat anomalous result will be considered further below.

Comparison of the results obtained here against experiment, Table 6.4 (Wolfenden et al., 1981; Florián and Warshel, 1997; Smith, 1999a; Shirts et al., 2003), shows that the LD-Amber approach leads to very good results for most of the analogues and acceptable results for the remainder. Although the largest relative deviation from experiment is 65% for the Phe analogue, the corresponding deviation of 2.07 kJ/mol is small compared to experimental uncertainty. The 15% and 26% relative deviations for the Ser and Met analogues, respectively, are similarly associated with small deviations and are, therefore, acceptable. The most disappointing results are those of the Asn, Trp, Gln and Tyr analogues, where the

Table 6.3 Free energy change for transfer of the amino acid sidechain analogues from the gas phase to water at pH 7 estimated by the LD-Amber approach, $\Delta G_s^{\text{LD-Amber}}$, and its constituent parts – electrostatic contribution due to solvent modeled by dipoles, $\Delta G_{es(LD)}$, electrostatic contribution due to solvent volume beyond that modeled by dipoles, $\Delta G_{es(c)}$, contribution due to van der Waals interactions between solute and dipoles, ΔG_{vdW}, and hydrophobic contribution, ΔG_{phob}.

Side chain analogue of	$\Delta G_{es(LD)}$ (kJ/mol)	$\Delta G_{es(c)}$ (kJ/mol)	ΔG_{vdW} (kJ/mol)	ΔG_{phob} (kJ/mol)	$\Delta G_s^{\text{LD-Amber}}$ (kJ/mol)
Ala	-0.05	0.00	-5.90	14.51	8.56
Arg^+	-184.91	-41.49	-21.23	2.22	-245.41
Asn	-23.93	-0.94	-11.85	5.67	-31.06
Asp^-	-282.75	-42.19	-11.22	1.38	-334.77
Cys	-5.01	-0.10	-8.55	8.13	-5.53
Gln	-24.26	-0.37	-13.69	7.52	-30.80
Glu^-	-276.46	-42.42	-13.75	1.62	-331.02
His^+	-183.87	-42.03	-18.58	2.00	-242.49
Ile	-0.08	0.00	-12.83	21.99	9.07
Leu	-0.24	0.00	-12.95	21.74	8.55
Lys^+	-212.98	-42.20	-16.88	1.98	-270.07
Met	-4.44	-0.05	-13.05	12.97	-4.57
Phe	-4.12	0.00	-18.68	21.69	-1.11
Ser	-21.74	-0.85	-7.38	5.61	-24.36
Thr	-19.99	-0.30	-9.80	7.96	-22.13
Trp	-12.91	-0.10	-23.66	17.50	-19.17
Tyr	-15.12	-0.03	-19.28	15.67	-18.76
Val	-0.12	0.00	-10.92	20.02	8.99

deviations from experiment are greater than experimental uncertainty and not an insignificant fraction – 22% to 27% – of the total. However, as will be seen below, the deviations are certainly no worse than those associated with other methods. It is interesting to note that three of the six analogues whose relative deviations from experiment exceed 20% are the only molecules that contain a benzene ring. This could suggest a particular incompatibility between the Amber charge distribution for the aromatic carbon atoms and the associated LD van der Waals radius or, alternatively, the difficulties faced in identifying unambiguously this radius for atoms

Table 6.4 Comparison of solvation free energies from LD-Amber, $\Delta G_s^{\text{LD-Amber}}$, the LD method (Florián and Warshel, 1997), ΔG_s^{LD}, a continuum method (Smith, 1999a), ΔG_s^{C}, a traditional explicit method (Shirts et al., 2003), ΔG_s^{ES}, and experiment, ΔG_s^{e}.

Sidechain analogue of	$\Delta G_s^{\text{LD-Amber}}$ (kJ/mol)	ΔG_s^{LD} (kJ/mol)	ΔG_s^{C} (kJ/mol)	ΔG_s^{ES} (kJ/mol)	ΔG_s^{e} (kJ/mol)
Ala	8.56 $^{0.44}_{5\%}$	7.53 $^{-0.59}_{7\%}$	6.4 $^{-1.72}_{21\%}$	9.37 $^{1.25}_{15\%}$	8.12†
Arg^+	-245.41 $^{-}_{-}$	-	-246.3 $^{-}_{-}$	-	None available
Asn	-31.06 $^{9.44}_{23\%}$	-37.24 $^{3.26}_{8\%}$	-47.4 $^{-6.90}_{17\%}$	-35.61 $^{4.89}_{12\%}$	-40.50†
Asp^-	-334.77 $^{10.33}_{3\%}$	-326.35 $^{18.75}_{5\%}$	-343.7 $^{1.40}_{<1\%}$	-	-345.10‡
Cys	-5.53 $^{-0.34}_{7\%}$	-6.28 $^{-1.09}_{21\%}$	-0.9 $^{4.29}_{83\%}$	-2.30 $^{2.89}_{56\%}$	-5.19†
Gln	-30.80 $^{8.45}_{22\%}$	-	-42.8 $^{-3.55}_{9\%}$	-36.11 $^{3.14}_{8\%}$	-39.25†
Glu^-	-331.02 $^{6.58}_{2\%}$	-	-336.1 $^{1.50}_{<1\%}$	-	-337.60‡
His^+	-242.49 $^{5.11}_{2\%}$	-	-256.3 $^{-8.70}_{4\%}$	-	-247.60‡
Ile	9.07 $^{0.07}_{1\%}$	8.37 $^{-0.63}_{7\%}$	8.3 $^{-0.70}_{8\%}$	10.17 $^{1.17}_{13\%}$	9.00†
Leu	8.55 $^{-0.99}_{10\%}$	-	7.3 $^{-2.24}_{23\%}$	9.50 $^{-0.04}_{<1\%}$	9.54†
Lys^+	-270.07 $^{7.73}_{3\%}$	-	-274.9 $^{2.90}_{1\%}$	-	-277.80‡
Met	-4.57 $^{1.62}_{26\%}$	-	-0.4 $^{5.79}_{94\%}$	-1.46 $^{4.73}_{76\%}$	-6.19†
Phe	-1.11 $^{2.07}_{65\%}$	-	-2.1 $^{1.08}_{34\%}$	-3.60 $^{-0.42}_{13\%}$	-3.18†
Ser	-24.36 $^{-3.19}_{15\%}$	-25.10 $^{-3.93}_{19\%}$	-25.7 $^{-4.53}_{21\%}$	-18.87 $^{2.30}_{11\%}$	-21.17†
Thr	-22.13 $^{-1.71}_{8\%}$	-22.18 $^{-1.76}_{9\%}$	-26.4 $^{-5.98}_{29\%}$	-17.66 $^{2.76}_{14\%}$	-20.42†
Trp	-19.17 $^{5.43}_{22\%}$	-	-22.4 $^{2.20}_{9\%}$	-20.42 $^{4.18}_{17\%}$	-24.60†
Tyr	-18.76 $^{6.80}_{27\%}$	-	-33.4 $^{-7.84}_{31\%}$	-20.84 $^{2.72}_{11\%}$	-25.56†
Val	8.99 $^{0.66}_{8\%}$	8.37 $^{0.04}_{<1\%}$	7.9 $^{-0.43}_{5\%}$	9.79 $^{1.46}_{18\%}$	8.33†

Deviations and relative deviations (%) from experiment are given as superscripts and subscripts to predictions respectively.

† (Wolfenden et al., 1981)

‡ (Smith, 1999a)

that are part of a resonant structure (Florián and Warshel, 1997). Two of the remaining analogues whose relative deviations from experiment exceed 20% – those of Asn and Gln – are the only molecules that contain a carboxyamide group. This and the fact that the associated deviations from experiment are similar to those for the (charged) analogues that contain constituent parts of the carboxyamide group (i.e. NH_2 or sp^2 hybridized oxygen) suggest that one or more of the constituent parts of this group may be the source of the discrepancy.

Table 6.4 includes results obtained by Florián and Warshel (Florián and Warshel, 1997) for the eight analogues considered by them. Comparison of these results with those generated here reveals the latter to be better in all but two cases. The average deviation from experiment for the eight analogues is 3.27 kJ/mol compared to 3.76 kJ/mol for those of Florián and Warshel, whilst the corresponding average relative deviations are 9% and 10% respectively. Comparison of our results with those obtained by Smith (Smith, 1999a) using a continuum solvent based model (column 4 of Table 6.4) provides, on first glance, a more mixed picture. The average deviation of our results from experiment is 4.17 kJ/mol compared to 3.63 kJ/mol for those of Smith. This larger average deviation should, however, be contrasted with the average relative deviation of our results which, at 15%, is some 8% less than that associated with the results of Smith. This reflects the fact that, whilst the continuum solvent based method used by Smith works particularly well for the charged analogues (average deviation and relative deviation from experiment are 3.63 kJ/mol and 1% compared to 7.44 kJ/mol and 2% for the results obtained here), it is particularly poor for the uncharged analogues where the average deviation and relative deviation are 3.63 kJ/mol and 30% compared to 3.17 kJ/mol and 18% for our results. Comparison of the results generated here with those obtained by Shirts and co-workers (Shirts et al., 2003) for the neutral analogues using a traditional explicit solvent based approach, column 5 of Table 6.4, reveals our results to be better in 7 out of the 13 cases. This is reflected in a slightly better average relative deviation here of 18% against 20% for Shirts *et al*., although the average deviation from experiment for our results is 0.71 kJ/mol worse at 3.17 kJ/mol.

It can be concluded from the above analysis that use of the Amber potential model within the LD framework of Florián and Warshel (Florián and Warshel, 1997)

produces results that are consistent and of an accuracy similar to those produced by other methods. Perhaps the only cause for concern in using Amber within the LD framework are the less than accurate results for the three analogues that contain a benzene ring and the two that contain the carboxyamide group. Although the accuracy of the results obtained here for these analogues are not substantially worse than those of other methods or excessive compared to experimental uncertainty, improved results may follow re-parameterization of the van der Waals radius associated with the interaction between the dipoles and the aromatic carbon atom and constituent atoms of the carboxyamide group.

6.4.2. Free Energy Surface of Alanine Dipeptide in Neutral Water

Experimental and theoretical studies all suggest the FES of alanine dipeptide is characterized by a number of local minima depending on the environment. The first group of commonly cited minima is associated with seven and five membered ring structures – denoted by C_{7eq}, C_{7ax} and C_5 – formed by an intramolecular hydrogen bond between the CO and NH groups at the ends of the peptide. The remainder of the commonly cited minima are associated with more extended structures, with their dihedral angles being similar to those of the left-handed polyproline II helix, denoted by P_{II}, and right- and left-handed alpha helices, which are denoted by α_R and $\alpha_{L,}$ respectively (it should be noted that whilst the angles are similar to those found in these helical structures, they are not P_{II}- or α-helices per se as the intramolecular hydrogen bond pattern that defines them is absent due to, clearly, an insufficiency of residues). As the dihedral angles of the C_{7eq}, C_5 and P_{II} conformers are all located within the β region of the Ramachandran plot, one or more of them are sometimes collectively denoted as β conformers (often without distinction).

The PES of the alanine dipeptide in the gas phase shown in Figure 6.5 reveals five separate minima. As with most previous theoretical studies (Pettitt and Karplus, 1988; Tobias and Brooks, 1992; Gould et al., 1994; Schmidt and Fine, 1994; Buesnel et al., 1997; Chipot and Pohorille, 1998; Apostolakis et al., 1999; Rosso et al., 2005), the global potential energy minimum is associated with the C_{7eq} conformer, located here at $\phi = -80^{o}$ and $\psi = 76^{o}$ where $U = -108.45\,\text{kJ/mol}$. The energy of the C_5 conformer, located here at $\phi = -152^{o}$ and $\psi = 166^{o}$, is only slightly higher at

$U = -107.68\,\text{kJ/mol}$ – this result is also very much inline with most of the previous studies already cited which consider more than the two C_7 conformers (Tobias and Brooks, 1992; Gould et al., 1994; Buesnel et al., 1997; Chipot and Pohorille, 1998).

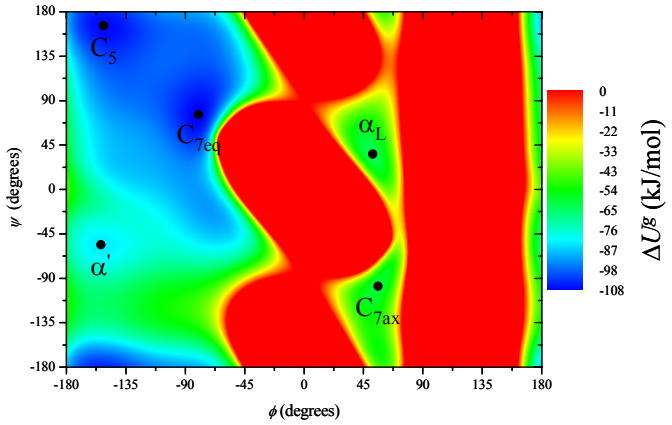


Figure 6.5 Potential energy surface (PES) of alanine dipeptide determined using the Amber potential model with the associated *ff*94 parameter set (Cornell et al., 1995).

The potential energies of the other two commonly known conformers, α_L and C_{7ax}, are substantially higher at $U = -63.00\,\text{kJ/mol}$ ($\phi = 52^\circ$, $\psi = 36^\circ$) and $U = -60.41\,\text{kJ/mol}$ ($\phi = 56^\circ$, $\psi = -98^\circ$) respectively. This order is the reverse of those studies already cited which consider these conformers (Gould et al., 1994; Chipot and Pohorille, 1998). The minimum with $U = -83.04\,\text{kJ/mol}$ at $\phi = -154^\circ$ and $\psi = -56^\circ$ is located in the α-helix region of the Ramachandran plot, but does not correspond to the traditional α_R structure, which is typically located at much lower values of ϕ. It does, however, correspond to the α' conformer identified by Head-Gordon and co-workers (Head-Gordon et al., 1991) for an analogue of the alanine dipeptide obtained by replacing the terminal methyl groups with hydrogen atoms; this conformer is, therefore, denoted here accordingly.

It is interesting to compare in greater detail the PES obtained here with that of Gould and co-workers (Gould et al., 1994), as their results form the basis for the Amber *ff*94 parameter set. The potential energy difference obtained here between the two most stable conformers and their associated angles are very much inline with those of these workers, which is particularly encouraging given our interests lie primarily in this part of the Ramachandran plot. As already indicated, the order of

stability of the α_L and C_{7ax} conformers obtained here is the reverse of that obtained by Gould and co-workers (Gould et al., 1994) – this difference most likely arises

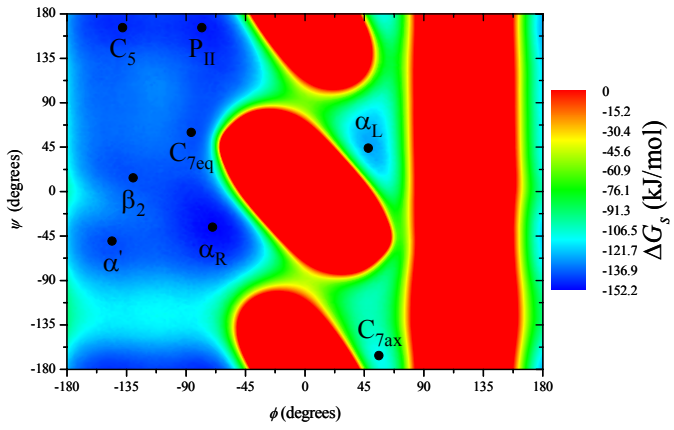


Figure 6.6 Free energy surface (FES) of alanine dipeptide in neutral water determined using the PES in Figure 6.5 and solvation energies obtained from the LD-Amber approach.

from our decision to not locally relax the sidechain structure when generating the PES. Gould *et al.* identify three conformers $-\beta_2$ ($\phi = -131^o$, $\psi = 22^o$), α_R ($\phi = -61^o$, $\psi = -41^o$) and β ($\phi = -58^o$, $\psi = -134^o$) in order of decreasing stability – which do not appear to exist in the gas-phase PES obtained here. However, as all the angles associated with these minima lie within a rather flat bottomed low-energy region as shown in Figure 6.5, these minima may well appear if the side chain were to be locally relaxed. Indeed, as will be seen below, two of these minima subsequently appear on solvation of the gas-phase structure, suggesting that nascent minima may in fact exist in the PES.

Figure 6.6 shows the FES of the alanine dipeptide in neutral water evaluated using the PES of Figure 6.5 and the solvation energies obtained from the LD-Amber approach. This figure reveals all the expected minima as well as two additional minima. The first of these additional minima is located in the lower left hand corner of the β sheet region and is connected to both the C_5 and α_R regions – this minimum has been denoted here by β_2 following Gould *et al.* who predicted a very similar structure (Gould et al., 1994). The second additional local minimum is to the left of, and directly connected to, the traditional α_R minimum. As already indicated in the discussion above, this minimum appears to arise from the α' minimum in the gas-phase PES and, as such, has been denoted here accordingly.

Table 6.5 Minima identified here for the alanine dipeptide in neutral water and saddle points between a selection of these minima with associated dihedral angles and free energy values.

Conformer	ϕ	ψ	ΔG_s kJ/mol
α_R	-70°	-36°	-152.157
P_{II}	-78°	166°	-149.438
C_5	-138°	166°	-147.076
C_{7eq}	-86°	60°	-141.318
α'	-146°	-50°	-141.197
β_2	-130°	14°	-140.877
α_L	48°	44°	-127.672
C_{7ax}	56°	-166°	-113.663
$P_{II}\leftrightarrow C_5$	-112°	168°	-143.718
$P_{II}\leftrightarrow C_{7eq}$	-80°	76°	-140.056
$C_{7eq}\leftrightarrow\alpha_R$	-88°	34°	-138.349
$C_5\leftrightarrow\beta_2$	-152°	74°	-135.855
$\alpha_R\leftrightarrow\beta_2$	-118°	10°	-139.173
$\alpha_R\leftrightarrow\alpha'$	-124°	-50°	-138.227

Table 6.5, which gives details of the free energy minima of Figure 6.6 and the saddle points between them, shows the α_R conformer to be the most stable. The free energy difference between this and the P_{II} conformer and the low barriers along the pathways between them (shown in Figure 6.7) suggests, however, that the latter may also be well populated at equilibrium. The small free energy difference between the C_5 and P_{II} conformers and the low barrier between them (Figure 6.7) suggests that the former may also be partially populated, although clearly less so than the α_R and P_{II} states. The remaining conformers are unlikely to be significantly populated at equilibrium because, as illustrated in Figure 6.7, the energy levels are substantially higher than those of the three lowest states and the barriers for movement into them from these states are considerable.

Whilst there is still some debate, the vast majority of the experimental work to date suggests the alanine dipeptide in water takes on either the P_{II} conformation (Poon et al., 2000; Kim et al., 2005), or rapidly switches back and forth between this and the α_R conformation (Madison and Kopple, 1980; Han et al., 1998; Poon et al., 2000; Gnanakaran and Hochstrasser, 2001; Mehta et al., 2004). One of the

experimental studies (Takekiyo et al., 2004) also suggests the C_5 conformer may exist at equilibrium. There is very little recent experimental evidence for the existence of any other of the conformers at equilibrium. The free energy surface obtained here appears, therefore, to be inline with the experimental evidence excepting that which indicates the P_{II} conformation dominates (Poon et al., 2000; Kim et al., 2005).

There is an abundance of theoretical studies concerned with the alanine dipeptide in water. Unfortunately, the predicted energy levels, and even the differences between them, vary from study to study. These disagreements arise for a variety of reasons including differences in the solvent models (Smith, 1999b; Freedman and Truong, 2004), potential energy models (Resat et al., 1997; Hu et al., 2003), free energy components included (e.g. some include the solute entropic contribution whilst many do not), levels of accuracy (e.g. quantum model level; number of MC steps), and structures used (e.g. gas phase structures). All these make quantitative comparison between model predictions very difficult. There is scope, however, for a qualitative comparison at least in terms of order of conformer stability and the range in which the angles fall.

Analysis of those studies that give the relative free energies of the dipeptide in water, which are summarized in Table 6.6, (Stillinger and Rahman, 1974; Rossky et al., 1979; Jorgensen, 1981; Pettitt and Rossky, 1982; Jorgensen et al., 1983; Hermans et al., 1984; Anderson and Hermans, 1988; Pettitt and Karplus, 1988; Tobias and Brooks, 1992; Gould et al., 1994; Schmidt and Fine, 1994; Cornell et al., 1995; Buesnel et al., 1997; Florián and Warshel, 1997; Smart et al., 1997; Chipot and Pohorille, 1998; Apostolakis et al., 1999; Smith, 1999b; Hu et al., 2003; Rosso et al., 2005) reveals that 11 predict one of the β-sheet conformers ($5 \times P_{II}$, $5 \times \beta$ and $1 \times \beta_2$) to be the most stable against six for the α_R conformer. Further analysis also shows that the β-sheet and α_R conformers are predicted to be the second most stable in nine ($2 \times P_{II}$, $2 \times \beta$, $4 \times C_5$ and $1 \times C_{7eq}$) and six cases, respectively. Bearing in mind the many differences between the models of Table 6.6, the results obtained here are in line with these previous studies.

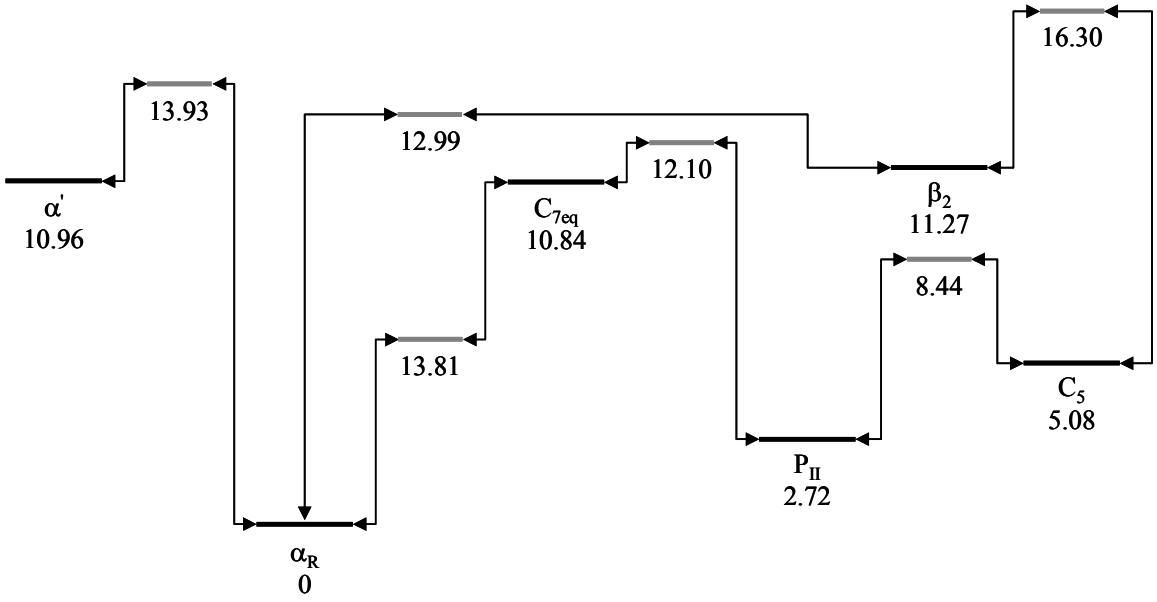


Figure 6.7 Schematic showing to scale the free energies (kJ/mol) associated with the minima on the left hand side of the FES in Figure 6.6 and the transition states between them. All energies are given relative to the global energy minimum associated with α_R conformation.

The dihedral angles of the α-helical and β-sheet conformers obtained here all fall within their respective regions as normally defined. Analysis of Table 6.6 also indicates that the vast majority of the angles predicted here fall within the range of values obtained in the previous studies. The only two exceptions are the ψ angle of the α_R conformer, which falls some 5° above the highest previously predicted value of –41° (Buesnel et al., 1997), and the same angle for the C_{7ax} conformer, which is 26° below the lowest previously predicted value of –140° (Smart et al., 1997) – as discussion above suggests, these differences most likely arise from our not relaxing the sidechain when generating the gas-phase PES. The dihedral angles of the C_{7eq} conformer are also both slightly below the range of angles predicted by others (Pettitt and Karplus, 1988; Gould et al., 1994; Buesnel et al., 1997). Whilst the approximate

Table 6.6 Comparison of order of stability of alanine dipeptide conformers as indicated by various predictions of relative free energies in water

Reference	Method[†]	SM[‡]	PE	Conformations in order of most stable (left) to least stable (right) in water[*]						
(Pettitt and Karplus, 1988)	XRISM	PR-TIP	(Rossky et al., 1979)	P_{II} -56°, 171°	$\sim C_{7ax}$ 63°, -69°	C_5 -177°, 180°	C_{7eq} -66°, 70°	α_L 63°, 49°	α_R -68°, -56°	
(Anderson and Hermans, 1988)	MD-TI	SPC	CEDAR (Hermans et al., 1984)	β -110°, 120°	α_R -120°, -40°	α_L 60°, 100°	C_{7ax} 70°, -60°			
(Tobias and Brooks, 1992)	MD-FEP	TIP3P	CHARMM19	β (≅P_{II}) -80°, 120°	α_R -80°, -60°	C_{7ax} 60°, -80°	α_L 60°, 60°			
(Gould et al., 1994)	QM	SCRF	HF/6-31G**	β_2 -112°, 23°	β -118°, 133°	C_{7eq} -73°, 75°	C_{7ax} 75°, -73°	α_L 68°, 39°		
(Schmidt and Fine, 1994)	-	PB/ NP-SAS	CFF91	P_{II} -70°, 150°	C_5 -150°, 150°	α_R -90°, -60°	α_L 70°, 70°	C_{7ax} 80°, -70°		
(Buesnel et al., 1997)	MC-FEP QM/MD	TIP3P	CHARMM AM1 MP2/6-31G**	α_R -61°, -41°	C_{7eq} -73°, 75°	β_2 -112°, 23°	C_5 -179°, 180°	β -118°, 133°	α_L 69°, 39°	C_{7ax} 75°, -73°
(Smart et al., 1997)	SD/MC	PB/ NP-SAS	CHARMM	β (≅P_{II}) -70°, 120°	α_R -70°, -50°	C_5 -160°, 150°	α_L 50°, 50°	C_{7ax} 50°, -140°		
(Chipot and Pohorille, 1998)	MD-US	TIP4P	AMBER	β	$\sim\alpha_R$	C_5	C_{7eq}			
(Smith, 1999b)	MD-US	PB	CHARMM22	β (≅P_{II}) -73°, 132°	α_R -69°, -61°	C_{7ax} 56°, -88°				
		TIP3P		α_R -72°, -56°	β (≅P_{II}) -80°, 162°	C_{7ax} 61°, -133°	α_L 59°, 57°			

Table 6.6 continued

Reference	Method†	SM‡	PE	Conformations in order of most stable (left) to least stable (right) in water*						
(Apostolakis et al., 1999)	MD-US	CHARMM19/23		P_{II}# -75°, 136°	C_5 -147°, 152°	α_R -76°, -50°	C_{7ax} 57°, -84°	α_L 51°, 81°	$\alpha_{L'}$ 65°, 143°	
(Hu et al., 2003)	MD	TIP3P	CHARMM22	α_R	~β					
			AMBER98	α_R	β					
			OPLS	β	α_R	α_L	C_{7ax}			
			SCCDFTB AMBER98	β	α_R	α_L	C_{7ax}			
(Rosso et al., 2005)	MD-AFED	?	CHARMM22	α_R -81°, -63°	β (≅P_{II}) -81°, 153°	C_{7ax} 63°, -117°				
This study	-	LD-Amber	AMBER94	α_R -70°, -36°	P_{II} -78°, 166°	C_5 -138°, 166°	C_{7eq} -86°, 60°	α_{R2} -146°, -50°	β_2 -130°, 14°	α_L 48°, 44°

† Method: Extended reduced interaction-site method (XRISM) (Pettitt and Rossky, 1982); Monte Carlo (MC); molecular dynamics (MD); quantum mechanics (QM); free energy perturbation (FEP); thermodynamic integration (TI); umbrella sampling (US).

‡ Solvent model: self consistent reaction field (SCRF); Poisson-Boltzmann only (PB); Poisson-Boltzmann with non-polar contributions included via solvent-accessible surface area term (PB/NP-SAS); Langevin dipole (Florián and Warshel, 1997) with Amber and associated *ff*94 parameter set (Cornell et al., 1995) (LD-Amber); explicit model using the following water molecule models: TIPS model (Jorgensen, 1981) as modified by Pettitt and Rossky (PR-TIP) (Pettitt and Rossky, 1982); TIP3P model (TIP3P) (Jorgensen et al., 1983); TIP4P model (TIP4P) (Jorgensen et al., 1983); SPC model (SPC) (Stillinger and Rahman, 1974).

* Only conformers considered in the studies are shown. Associated dihedral angles ϕ and ψ are shown, respectively, when given by authors. Where authors indicate a β-sheet structure only, the conformer with angles nearest those given by the authors is indicated in parentheses.

Apostolakis and co-workers (Apostolakis et al., 1999) assigned the dihedral angles of $\phi = -75^{o}$ and $\psi = 136^{o}$ to C_{7eq}. These angles are, however, more properly associated with the P_{II} conformer.

nature of the gas-phase PES undoubtedly has a part to play in this, the small differences can also be blamed on the difficulties faced in identifying precisely the location of this weak and broad minimum.

6.4.3. Electrostatic Potential Field and Water Structure around Alanine Dipeptide

The electrostatic potential fields (EPF), $\Phi(\mathbf{r})$, generated from MD and LD-Amber on a plane through the α_L conformation of the alanine dipeptide in neutral water are compared in Figure 6.8. In order to better enable comparison, the solute contribution to the field, which is the same in both cases, has been removed and the fields normalized (this was done because the numerical values were somewhat different as expected from two different solvent models). The focus on the α_L conformation and the plane shown in Figure 6.8 is motivated by the work of Beglov and Roux (Beglov and Roux, 1995), which indicates hydrogen bonded water bridges exist between the hydrogen atoms in this plane and the oxygen atoms just above and below the plane as indicated at the bottom of Figure 6.8.

Although the EPF obtained from the LD-Amber approach is smoother and somewhat more diffuse than that obtained from MD, there are many similarities between them. Both are clearly characterized by regions of significant negative and positive potential above and below $y = 0$, respectively. Whilst these regions in the MD field consist of two separate extrema connected by a saddle point, it is clear that spatial smoothing of this field would lead to a single extremum in each region similar to that seen in the LD-Amber field. There are also striking similarities in the symmetry of the two fields about $x = 0$ – in both cases the global minimum and maximum are located above $x = 0$, the positive regions are cusped upwards at the left-hand end and rounded at the other, and all the regions are tilted slightly upwards from left to right. The LD-Amber-related 3D contour plot shown in Figure 6.8 indicates the existence of a number of smaller extrema distributed around the two global extrema. Although noise due to poor sampling makes it difficult to discern similar extrema in the MD field with any precision, there are hints of such extrema as indicated by the arrows in the MD-related 3D contour plot in Figure 6.8.

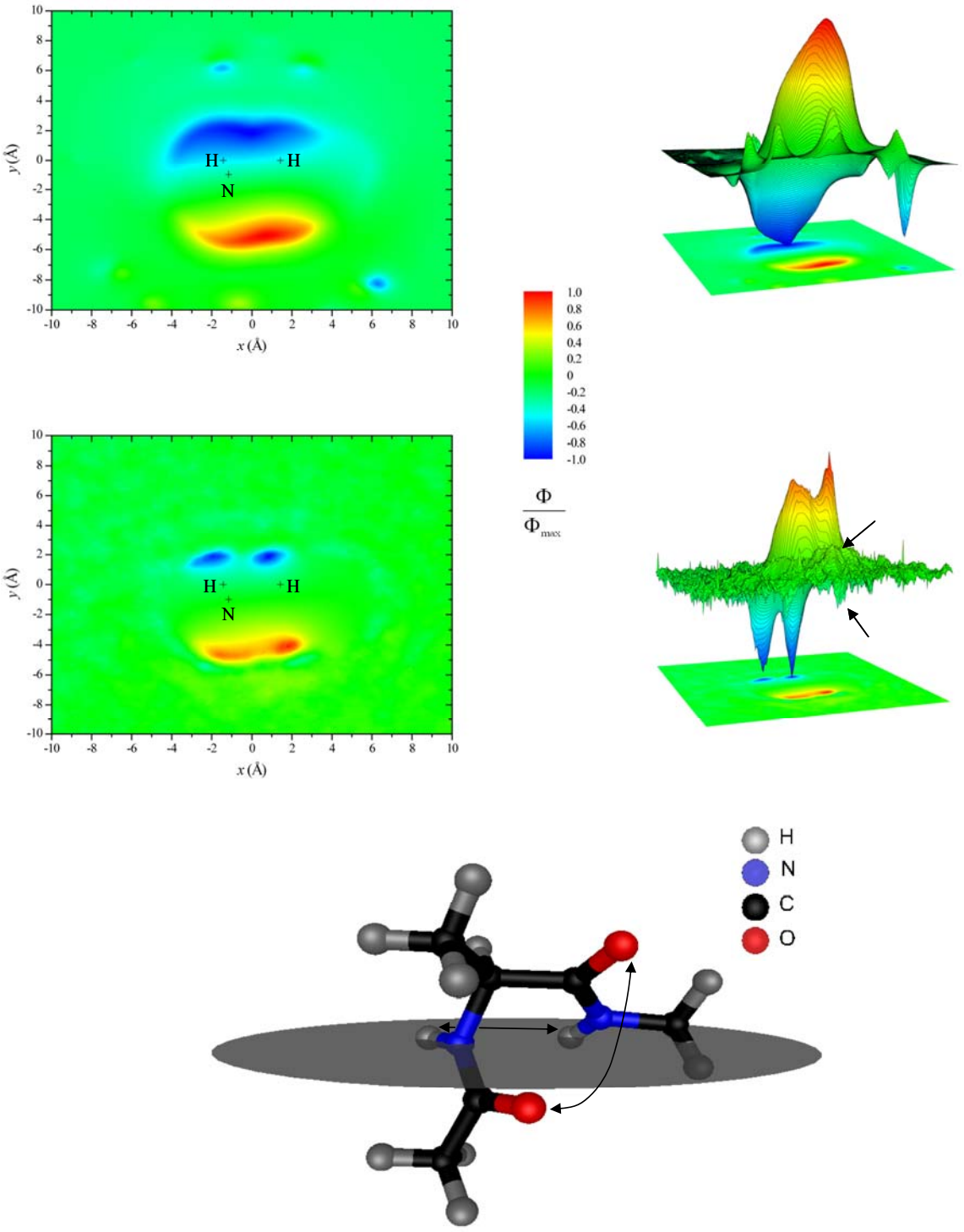


Figure 6.8 Two-dimensional (2D) and three-dimensional (3D) contour plots of the electrostatic potential field (EPF) on the plane through the α_L conformation of the alanine dipeptide shown at the bottom of the figure when in neutral water as predicted by an explicit approach based on the MD (top) and the LD-Amber approach (middle); the position of the hydrogen and nitrogen atoms in the plane are shown on the 2D contour plots. The termini of the two hydrogen bridges that are responsible for the EPF extrema are shown by double-ended arrows in the figure at the bottom.

Geometric analysis of the MD-based field suggests the negative region above $y = 0$ can be attributed to a water bridge between the two solute hydrogen atoms in the plane. In particular, the distances between the left-hand and right-hand solute hydrogen atoms and the corresponding peaks are $\sim 1.96\,\text{Å}$ and $\sim 1.91\,\text{Å}$, respectively – almost exactly inline with the theoretical values – whilst the distance between the two negative extrema is $\sim 2.7\,\text{Å}$, very much inline with the structure suggested by Beglov and Roux (Beglov and Roux, 1995) (Figure 4 in their paper suggests the distance is less than (because the O-H$\cdots$O distance is not co-linear) $1.92 + 0.95 = 2.87\,\text{Å}$, where the second value in this sum is the canonical O-H bond length).

Geometric analysis of the positive region below $y = 0$ is more difficult. It is, however, not unreasonable to attribute this region to the hydrogen atoms of one of the two water molecules that constitute a bridge between the solute oxygen atoms as indicated at the bottom of Figure 6.8. The existence of the two maxima indicates the oxygen atom of the water molecule "bonded" to the solute oxygen above the plane is in a downwards position, whilst the differences in the heights of the maxima suggests the distance between the hydrogen atoms of the water molecule are different.

It is clear that comparison of the electrostatic potential field derived from MD with that obtained from the LD-Amber approach can aid in the interpretation of the latter in terms of the solvent structure despite its smoothed character. Of course, it would make no sense to undertake such a comparison in general! However, as inversion of smoothed data to obtain atomic-level (albeit non-unique) detail has long been practiced in a variety of other fields, it is reasonable to suppose that such inversion processes may well work here.

6.4.4. Computational Performance

The computational expense of traditional explicit approaches based on use of thermodynamic integration or similar strategies with a series of Monte Carlo (MC) or molecular dynamic (MD) simulations along a "reaction coordinate" between the two conformations (Anderson and Hermans, 1988; Tobias and Brooks, 1992; Chipot and Pohorille, 1998; Smith, 1999b) depends on a number of issues including the number of steps (i.e. simulations) along the reaction coordinate, the length of the simulations,

and the number of water molecules involved. So as to get a reasonable best case estimate for the expense of the traditional explicit approach, it is assumed that just three simulations of 120k timesteps each are required along the reaction coordinate, and that only the first solvation layer is explicitly included, as proposed by Beglov and Roux (Beglov and Roux, 1995).

An 120k timestep MD simulation involving 165 rigid TIP3P (Jorgensen et al., 1983) water molecules and a single rigid alanine dipeptide molecule in the α_L conformation undertaken using Tinker (Ponder, 2004) took 464 minutes on the machine described in §6.3.5. Assuming the execution time of an MD simulation is a quadratic function of the number of atoms, this time would be reduced to approximately 40 minutes using the approach of Beglov and Roux (Beglov and Roux, 1995) in which only 43 water molecules are required. A good best case estimate of the total time required is, therefore, around 120 minutes. Whilst this estimate will be used here for comparison, it is recognized that it is optimistic, as more than three simulations would typically be required for results of better quality. Furthermore, there is evidence that explicit modeling of only the first hydration layer may be inadequate (Pal et al., 2002; Lee and Olson, 2005) and could lead to poor results (Frimand et al., 2000).

As indicated by equation (6.18), determination of the free energy difference between two conformers in the solution phase using the LD-Amber approach involves first evaluating the free energies of solvation of the two conformers using the method and then adding in the free energy difference between the conformers in the vapor phase. With each LD-Amber simulation taking approximately 0.6 seconds on the machine used here (described in §6.3.5) the total time required to evaluate the solvation free energies of the two conformers, assuming 10 simulations per conformer is sufficient, is 12 seconds. Assuming the free energy difference between the two conformers in the gas phase is evaluated following the protocol outlined above for the solution phase calculations leads to a total time of around 165 seconds for the LD-Amber-based approach, which is approximately 3% of the time estimated for the traditional explicit approach.

6.5. Conclusions

A thorough assessment of the use of the Amber potential model within the Langevin dipole (LD) framework of Warshel and co-workers – which we have termed LD-Amber to differentiate it from the various LD incarnations of these workers – was undertaken to assess the accuracy of this approach and its speed. The first part of the assessment involved comparison of the LD-Amber predictions for 18 amino acid side chain analogues with experimental and other theoretical results. This comparison showed the approach is able to produce results consistent with the experimental data and of similar accuracy to those produced by the best implicit and explicit methods. The second part of the assessment involved comparison of the LD-Amber-based free energy surface (FES) of the alanine dipeptide in neutral water with the published experimental and theoretical data. This comparison showed that the LD-Amber approach is able to produce a FES consistent with the vast majority of the experimental and theoretical results available in the literature. An approximate analysis undertaken here showed that this could be done with just 3% of the computational effort required if traditional explicit approaches are used. Finally, by comparing the electrostatic field for an alanine dipeptide conformer in neutral water obtained from the LD-Amber approach with that obtained from a molecular dynamics simulation, it was shown that the LD-Amber approach (and, therefore, LD method in general) is able to recover the correct field at a local level – this may offer the opportunity to establish the solvent structure from LD results using an inverse process. This ability to capture the solvent restructuring phenomenon gives the LD-Amber and other LD-based methods an advantage over somewhat computationally cheaper implicit solvent techniques.

Chapter 7. EA Based Study of Met-enkephalin in Water and at a Graphite-Water Interface

7.1. Introduction

It was shown in Chapters 4 and 5 that protein 3D structure in a gas phase and at the gas-solid interface could be predicted successfully using an EA based approach. Chapter 6, on the other hand, demonstrates that effects of protein solvation can be accurately incorporated into the free energy of the system with a very low computational cost using our LD-Amber model. The LD-Amber model is, however, used only to calculate the solvation free energy associated with a single 3D structure of a protein. In this chapter, we develop a method in which previous evolutionary algorithms are enhanced by embedding the contribution of the free energy of solvation into the EA fitness function. The LD-Amber facilitated EA method has been termed as LD-EA.

The LD-EA method has been applied to determine the 3D structure of met-enkephalin molecule (used in gas phase studies in Chapter 4) in water solution and at the interface between graphite and water. Met-enkephalin has previously been studied in a capped form. In water solutions, however, effect of explicit charges may be very important and the molecule is studied in both capped and zwitterionic forms here. The differences between the two forms are first briefly introduced, along with the description of the other elements of the system and a detailed explanation of the LD-EA method. The LD-EA method is then utilised to predict the structures of met-enkephalin in water solution and at the graphite-water interface. The structures are analysed and compared to the corresponding results obtained in the gas phase and at graphite-vacuum interface.

7.2. Model Details

7.2.1. Peptide

A large part of our work has been based on application of EA in prediction of 3D structure of the small polypeptide, met-enkephalin (Hughes et al., 1975). Since we have accumulated significant amount of experience on this system, we have decided to extend its study within a modified environment. Although it is beyond the scope of this specific study, further investigation of met-enkephalin may help us in analysing the differences between EA performance in predicting the vacuum and solvated structures of the same molecule. It should be noted that whilst our previous work was focused on investigation of EA ability to locate the global minimum of a function, this study is also oriented towards application of EA in predicting experimentally observed 3D structure of met-enkephalin. This has inevitably led to some changes in modelling of the molecule. Whilst the earlier study was conducted on polypeptide molecule with termini capped by electroneutral groups, most of the experimental studies that involved met-enkephalin have been performed using regular $–NH_2$-group on N-term and –COOH on C-term. However, Amber force field that we used in the earlier study of the LD method does not contain parameters for such terminated proteins. Rather than that, N- and C-termini are protonated and deprotonated, respectively, to form a zwitterionic form characterised by the presence of $–NH_3^+$ and $–COO^-$ groups on opposite ends of the molecule. Justification for investigation of zwitterionic molecule comes from some experimental studies (Roques et al., 1976; Jones et al., 1977), which have been performed in pH conditions that favour ionisation of end groups and formation of zwitterionic met-enkephalin molecule. Since the Amber force field that we have used in our previous LD study operates only with polarised termini and experimental results for this form of the molecule have already been collected, we have decided to use the zwitterionic form for testing the LD-EA method and its applicability in prediction of protein 3D structure. However, since our secondary goal was to compare the optimal structures obtained in vacuum and in water solution, we have also conducted one group of solvent-based simulations using met-enkephalin molecule capped with acetyl- and amino-methyl-groups – the same form as the one used in our previous, vacuum-

based investigation. Chemical structures of the two met-enkephalin forms are shown in Figure 7.1.

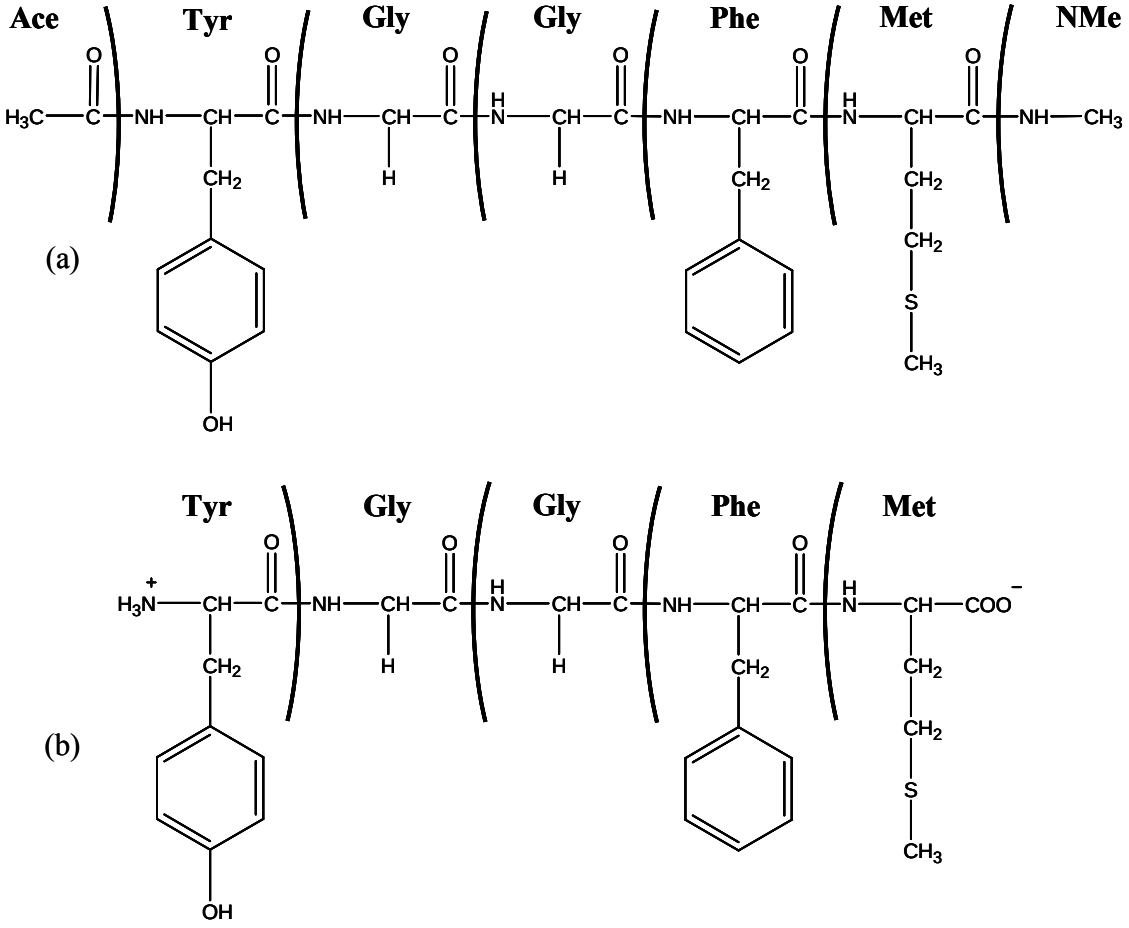


Figure 7.1 Molecular structure of two forms of met-enkephalin considered in this study: (a) capped and neutral; (b) zwitterionic.

Most of the atom parameters were taken from the Amber PE model (Cornell et al., 1995). As in our earlier LD-Amber study, atomic van der Waals radii for exclusion of water dipoles in the inner LD grid were taken from the original description of the method by Florián and Warshel (Florián and Warshel, 1997). Initial coordinates of atoms were calculated using the procedure already described in EA performance study. In short, each residue was capped with acetyl and amino-methyl groups on N- and C-termini, respectively, and allowed to relax using BFGS local minimisation algorithm. Bond lengths, angles and dihedrals obtained in the procedure were then translated into atomic coordinates and incorporated in met-enkephalin molecule. The ending residues, tyrosine and methionine, were subjected to somewhat modified procedure for calculation of their atomic coordinates in

zwitterionic form of met-enkephalin. In order to evaluate its bond characteristics in a more precise way, tyrosine was modified only by adding amino-methyl group to its C-terminus, whilst the N-terminus was modelled as ammonium group, $–NH_3^+$. Analogously, initial atomic coordinates for methionine in zwitterion were obtained from a residue capped with acetyl group on N-terminus and representing C-terminus as $–COO^-$.

7.2.2. Solvent

Description of the solvent is the same as that used in the development of the LD-Amber method (Mijajlovic and Biggs, 2007b). The only substantial difference in modelling is in the size of the system. Whilst the previous study was conducted on smaller molecules with up to 22 atoms for alanine dipeptide, zwitterionic met-enkephalin has 75, whilst the capped form includes 84 atoms. This has serious implications on computational cost as it extends duration of a single solvation energy calculation by 5-10 times. This problem has been approached by implementation of code execution in parallel environment.

The initial step of the procedure is equivalent to a regular, sequential execution of LD-Amber method. As a reminder, the calculation starts by taking a set of atomic coordinates as input and displacing the molecule in 10 random positions around the origin of the coordinate system. In the parallel version, each of these 10 random positions is sent to a different processor which calculates Langevin, hydrophobic, Lennard-Jones and bulk solvent energy for a single position. This enables synchronous calculation of solvation energies for all these positions. Finally, when the last of the parallel processes finishes, the results are returned to the master CPU, which averages them up. Although the speed of the whole procedure depends on the slowest of the solvation energy calculations, the whole procedure is considerably faster than serial evaluation of solvation energies for different positions in a loop.

Free energy surface (FES) of met-enkephalin in water is constructed analogously to that of alanine dipeptide in the LD-Amber study (Mijajlovic and Biggs, 2007b). The difference in the free energies of two solvated conformations can be approximated as the sum of the difference in their potential energies and the difference of their free energies of solvation

$$\Delta G_{12}^{s} \approx \Delta U_{12}^{g} + \Delta G_{s(2)} - \Delta G_{s(1)} \tag{7.1}$$

where ΔU_{12}^{g} is the potential energy difference between conformations 1 and 2, calculated using Amber PE model (Cornell et al., 1995), and $\Delta G_{s(i)}$ is free energy of solvation for structure *i*.

In order to utilise equation (7.1) in evaluation of fitness function for evolutionary algorithm, the free energy of a single conformation in water is calculated from an analogous expression

$$G_i^s = U_i^g + \Delta G_{s(i)} \tag{7.2}$$

where U_i^g is now the potential energy of a molecule in conformation *i*, while $\Delta G_{s(i)}$ is its free energy of solvation, as above.

Potential energy of the molecule is, like in the vacuum studies, expressed as a sum of electrostatic, U_{es}, van der Waals, U_{vdW}, and torsional term, U_{tor}, while bond lengths and angles between neighbouring chemical bonds are kept rigid, for which reason their contribution remains constant throughout the simulation and is not calculated. Free energy of solvation is decomposed as in our LD-Amber study and consists of proper Langevin dipole term, $\Delta G_{es(LD)}$, hydrophobic, ΔG_{phob}, and van der Waals, ΔG_{vdW}, terms and contribution of implicitly represented bulk solvent, $\Delta G_{es(c)}$. As before, the solute polarisation term in the original work of Florián and Warshel has been ignored, as its neglect has provided satisfactory results in the study of solvation of smaller molecules.

7.2.3. Solid Surface

Our earlier study of polyalanine adsorption (Mijajlovic and Biggs, 2007c) was based on smooth representation of the solid surface and application of Steele potential for calculation of protein-surface interactions (Bojan and Steele, 1987; Steele, 1993). Polyalanines are simple molecules and their repetitive structure allows utilisation of simple surface models. Met-enkephalin, on the other hand, has much more diversity and complexity in its side chain groups and is expected to show different adsorption behaviour on surfaces with different atomic structures. It is, therefore, necessary to model the surface in atomistic detail in order to capture all the characteristics of this interaction.

Met-enkephalin molecule features two aromatic rings. Experimental studies of other molecules that exhibit similar structural units have shown that aromatic rings can be involved in π-stacking mediated adsorption on graphitic layers (Zheng et al., 2003). In order to investigate the influence of this effect on met-enkephalin adsorption, we have decided to simulate its interaction with graphite substrate. Since π-stacking is established by interactions between delocalised electrons of aromatic rings, the necessity to capture this structural detail in graphite is obvious and it is clear that in this and similar systems smooth surface representation would be inferior to a model with full atomistic details. Distance between carbon atoms in graphite hexagonal rings is taken as 1.42 Å, while the distance between graphene layers is 3.35 Å (Trucano and Chen, 1975). Lennard-Jones parameters of carbon atoms for van der Waals interactions between graphite and protein are assumed to be the same as parameters of aromatic carbon in Amber force field (Cornell et al., 1995) – $\sigma = 3.3997\,\text{Å}$, $\varepsilon = 0.0860\,\text{kcal/mol}$. Parameters for van der Waals interactions between graphite and water dipoles are taken from the work of Florián and Warshel (Florián and Warshel, 1997) – $r^* = 3.0\text{Å}$ (sp^2 hybridised carbon), $C = 1.5$. Carbon atoms in graphite have been treated as uncharged and unpolarisable.

The main disadvantage of all-atom models is their high computational cost. Whilst interactions between protein and a smooth surface vary only as a function of height of the protein, the energy of adsorption on a structured surface is obtained by summing up interactions of individual protein atoms with each surface atom in turn. One of the ways to alleviate this obstacle, whilst still keeping a high level of structural detail, is to use a hybrid approach that represents a compromise between structured and smooth surface representations. Since the strongest protein-surface atomic interactions occur with the topmost layer of surface atoms, it is of crucial importance to represent this layer in atomistic detail, while any lower layers could be replaced with smooth planes of appropriate properties. As in our polyalanine adsorption study (Mijajlovic and Biggs, 2007c), surface energy has been calculated only from interactions with the two uppermost surface layers, which is justified by very small contributions of lower surface planes (Braun et al., 2002). Figure 7.2 illustrates the hybrid model of graphite surface used in this study. It should be noted that structural details of the surface are considered important only in its interaction

with the protein. Water model in the LD-EA approach is already simplified and using highly accurate surface model with it would not lead to significant improvements in accuracy. We have, therefore, decided to utilise a smooth representation in calculation of surface-solvent interactions.

Presence of solid surface dictates implementation of several phenomenological changes in procedure for calculation of free energy of the system. The most obvious new term that has to be added to overall free energy is the contribution of interaction between protein and solid surface, E_{surf}. Due to utilization of hybrid surface model, the surface interaction energy is calculated as a sum of explicit and implicit terms. Explicit term represents a sum of interactions between all protein atoms and carbon atoms in the first graphite layer. Interaction between a protein atom i and a surface atom j is discarded if the distance between the two is greater than a prespecified cutoff radius, r_{cut}. The cutoff radius is calculated as $2.5\sigma_{Cs^+}$, where cesium ion is the

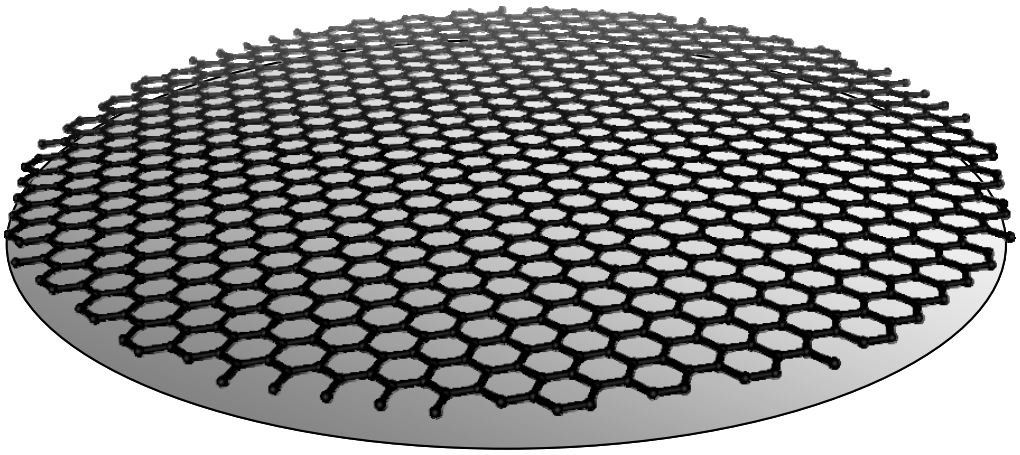


Figure 7.2 Hybrid model of graphite surface used in this study. The uppermost layer is modelled in full atomistic detail, while the lower is represented as smooth.

species with highest σ in Amber force field. In mathematical notation, the explicit surface energy contribution can be expressed as

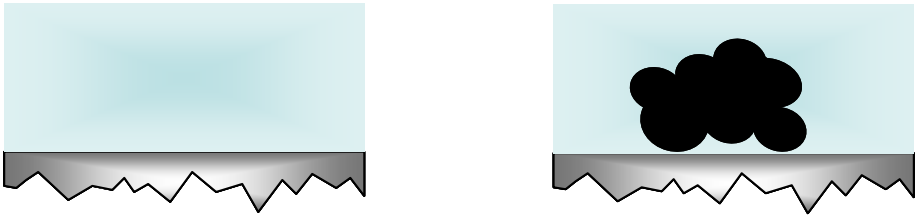
$$E_{surf}^{expl} = \sum_{i=1}^{N_a} \sum_{j\left(r_{ij} \le r_{cut}\right)} 4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{7.3}$$

where N_a is the number of atoms in adsorbed protein and σ_{ij} and ε_{ij} are obtained using arithmetic and geometric mixing rules, respectively, for protein, i, and graphite atoms, j. The implicit term in protein surface interaction describes interaction of the protein with lower, smoothly represented surface layers. Since our work has been

based on using only two layers of solid surface, one of which is represented explicitly, the implicit contribution to protein-surface interaction comes from a single layer of surface atoms and can be calculated using the following equation

$$E_{surf}^{impl} = 4\pi\rho\sum_{i=1}^{N_a}\varepsilon_{iC}\sigma_{iC}^2\left[\frac{1}{5}\left(\frac{\sigma_{iC}}{z_i+\Delta}\right)^{10} - \frac{1}{2}\left(\frac{\sigma_{iC}}{z_i+\Delta}\right)^{4}\right] \qquad (7.4)$$

where σ_{iC} and ε_{iC} are equal to corresponding coefficients σ_{ij} and ε_{ij} in equation for the explicit protein-surface interaction term, ρ is the surface density of atoms in a layer, whose value for graphite has been calculated as 0.3818 atoms/ Å^2, Δ is the distance between the graphite layers (Trucano and Chen, 1975), and z_i is the distance of atom i from the solid surface.



$\Delta G_{s(1)}^{surf}$ $\Delta G_{s(2)}^{surf}$

Figure 7.3 Solvation of smooth solid surface by Langevin dipoles in the absence (left) and presence (right) of an additional solute molecule.

Apart from its effect on protein, introduction of solid surface also affects structural features of solvent. Contribution of this effect to overall free energy of the system is, however, more difficult to evaluate. A rational approach in estimation of contribution of solid surface solvation is to calculate the solvation free energy of an area of solid surface without the presence of protein and then, the solvation free energy of the same area in the presence of adsorbed protein. Presence of additional solute molecule, as well as of its surrounding solvation layer, will produce overlaps with some water molecules (or, in terms of Langevin dipole model, number of solvent dipoles) in surface solvation layers. Overlapping dipoles of the surface solvation layer are removed from the system, thus reducing the magnitude of solvation free energy of the solid surface. A simplified graphical representation of the volume of solvent above the solid surface in the absence and presence of solute molecule is shown in Figure 7.3. As indicated in the figure, solvation energy of the solid surface area of interest is calculated as $\Delta G_{s(1)}^{surf}$, while the solvation of the same

area in the presence of solute molecule is $\Delta G_{s(2)}^{surf}$, and is usually smaller in magnitude owing to the reduction of number of water molecules due to overlaps with solute atoms. The difference between two solvation energies, $\Delta\Delta G_{s}^{surf} = \Delta G_{s(2)}^{surf} - \Delta G_{s(1)}^{surf}$, is considered as an individual term in the sum that forms overall free energy of the system.

The area of the solid surface for calculation of ΔG_{s}^{surf} contribution is constructed as a disc with the diameter 40 Å larger than maximum length of the adsorbed protein. A large disc is necessary to ensure that all dipoles from the solute solvation layers are above the area of interest even when molecule is in its most elongated conformation. The distribution of Langevin dipoles over the solid surface is performed in a way analogous to their distribution around the solute – 3 layers of inner grid dipoles with node distance of 1 Å are followed by an additional 3 layers of outer grid dipoles with node distance equal to 3.1043 Å. Since graphite surface bears no atomic charges, it does not generate any electrostatic field and surface solvation free energy comprises only of hydrophobic and van der Waals terms. Hydrophobic term is relevant only to inner grid dipoles (those closest to the surface), while it was shown that extending number of outer grid dipoles to more than 3 layers had negligible effect on van der Waals term in the solvation energy due to the distance of additional layers from the surface.

Although smoothing of surface planes for calculation of their interactions with adsorbed molecules is already described elsewhere (Steele, 1974), the procedure has been developed for 12-6 Lennard-Jones potential. The interaction between solid surface and point dipoles of the LD method is, however, modelled through 9-6 potential and the original equations for smooth surface had to be adjusted. Starting from equation for interaction between an individual atom in graphite lattice and point dipole and applying principles outlined in Steele procedure, an equation is derived for van der Waals contribution to solvation energy of graphite (derivation shown in Appendix D)

$$\Delta G_{vdW}^{surf} = 2\pi\rho k_{vdW} C \sum_{l=0}^{L-1} \sum_{j=1}^{N_d} N_j \left[\frac{2}{7} \left(\frac{r^*}{z_j + l\Delta} \right)^7 - \frac{3}{4} \left(\frac{r^*}{z_j + l\Delta} \right)^4 \right] \tag{7.5}$$

where k_{vdW} is van der Waals parameter defined by Florián and Warshel (Florián and Warshel, 1997) and equal to 0.84 kcal/mol, while ρ and Δ have already been introduced in equation (7.4). L and N_d represent the number of surface layers and Langevin dipoles, respectively. Analogously to the calculation of protein-surface interactions, only the first two surface layers are accounted for in equation (7.5) since the contribution of lower layers is too small. As in the original expression for van der Waals interactions involving dipoles, proposed by Florián and Warshel (Florián and Warshel, 1997), N_j is the normalisation factor introduced to balance the increase of dipole density of the inner solvation layer. z_j is the height of dipole j, or its vertical distance from the nearest surface layer. Values of r^* and C for carbon atoms in graphite planes have already been assigned above.

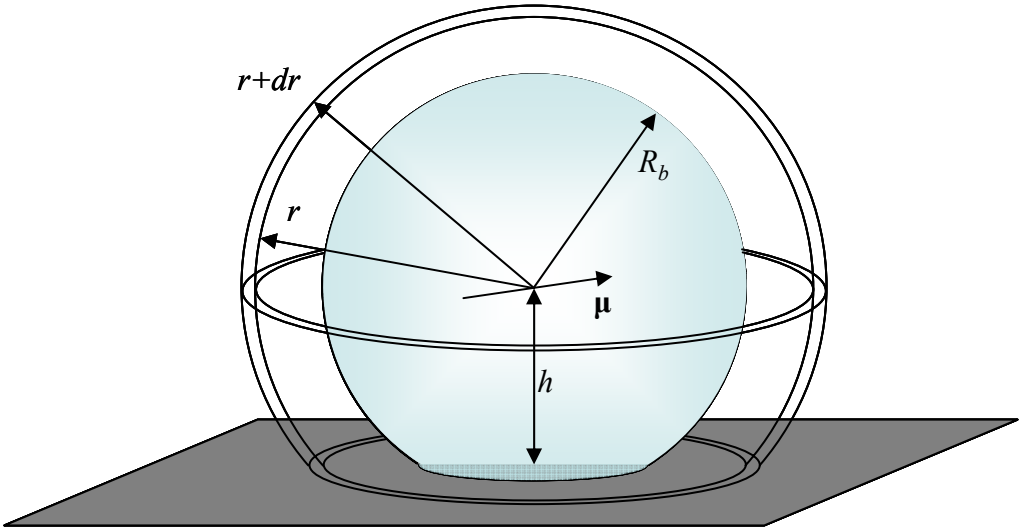


Figure 7.4 Bulk contribution to solvation free energy extends from sphere of radius R_b to infinity, but only in the domain above the solid surface.

Finally, presence of solid surface also affects calculation of solvation free energy of the solute itself. Any dipoles from the protein solvation layer that overlap with the solid surface are removed from the system, while hydrophobic and van der Waals energies of the remaining dipoles are calculated by summing up their corresponding interactions with both the solute and solid surface. Bulk contribution to solvation free energy is also adjusted due to the fact that presence of solid surface prevents integration of bulk solvent to infinity in all directions, as shown in Figure 7.4. Following approach of Born (Born, 1920b) for ionic solutes and Bell (Bell, 1931) for solvated dipoles, the integration is performed by summing individual contributions of infinitesimally thin spherical shells that surround the solute starting

from radius R_b from its geometric center to infinity, where R_b is radius of spherical volume whose interior is modelled using Langevin dipoles. However, unlike in the original approach of Born and Bell, solvation energy of an infinitesimal layer is multiplied by fraction of a shell that lies above the solid surface. For an ionised molecule whose geometry center is on a distance h from the solid surface, contribution of continuous solvent to solvation free energy can be calculated as

$$\Delta G_{\mathrm{es(c)}}^{ion} = \begin{Bmatrix} -\frac{1}{4}\frac{1}{4\pi\varepsilon_0}\left(1-\frac{1}{\varepsilon_r}\right)\frac{q^2\left(2R_b+h\right)}{2R_b^2} & ,h \le R_b \\ -\frac{1}{4}\frac{1}{4\pi\varepsilon_0}\left(1-\frac{1}{\varepsilon_r}\right)\frac{q^2\left(4h-R_b\right)}{2R_b h} & ,h > R_b \end{Bmatrix} \tag{7.6}$$

where q is the net electrostatic charge of the molecule, while ε_0 and ε_r are vacuum and relative dielectric permittivity, respectively. If the net charge of the solute molecule is 0, bulk solvent contribution to solvation free energy is calculated using molecule's total dipole moment, μ

$$\Delta G_{\mathrm{es(c)}}^{\mu} = \begin{Bmatrix} \frac{\mu^2}{96\pi\varepsilon_0 R_b^6 h^3}\frac{\varepsilon_r-1}{\left(2\varepsilon_r+1\right)^2}\Big[\left(\varepsilon_r-1\right)\left(16R_b^2-20h^6\right)- & \\ -4\left(2\varepsilon_r+1\right)R_b^3h^3-4\left(2\varepsilon_r+1\right)R_b^3h^3-9R_b^2h^4\Big] & ,h \le R_b \\ \frac{\mu^2}{96\pi\varepsilon_0 R_b^3 h^3}\frac{\varepsilon_r-1}{\left(2\varepsilon_r+1\right)^2}\Big[\left(4\varepsilon_r-1\right)R_b^3-8\left(2\varepsilon_r+1\right)h^3\Big] & ,h > R_b \end{Bmatrix} \tag{7.7}$$

Equations (7.6) and (7.7) are derived in the appendix.

Combining all new energy terms with those that have already been defined for the simple dissolved systems, total free energy of a protein solution in vicinity of a solid surface can be calculated using the following equation

$$G^s = U^g + E_{surf} + \Delta G_{\mathrm{s}} \tag{7.8}$$

in which ΔG_{s} is now calculated as the following sum

$$\Delta G_{\mathrm{s}} = \Delta G_{\mathrm{es(LD)}} + \Delta G_{\mathrm{es(c)}} + \Delta G_{\mathrm{vdW}} + \Delta G_{\mathrm{phob}} + \Delta\Delta G_{\mathrm{s}}^{surf} \tag{7.9}$$

where all the contributions have been described above.

7.3. Study Details

The basic outline of the evolutionary algorithm remains the same as the one used in EA performance study on met-enkephalin in gas phase – steady state, real encoding, multipoint crossover and uniform parent selection, $\mathrm{SRM_U}$ (Djurdjevic and

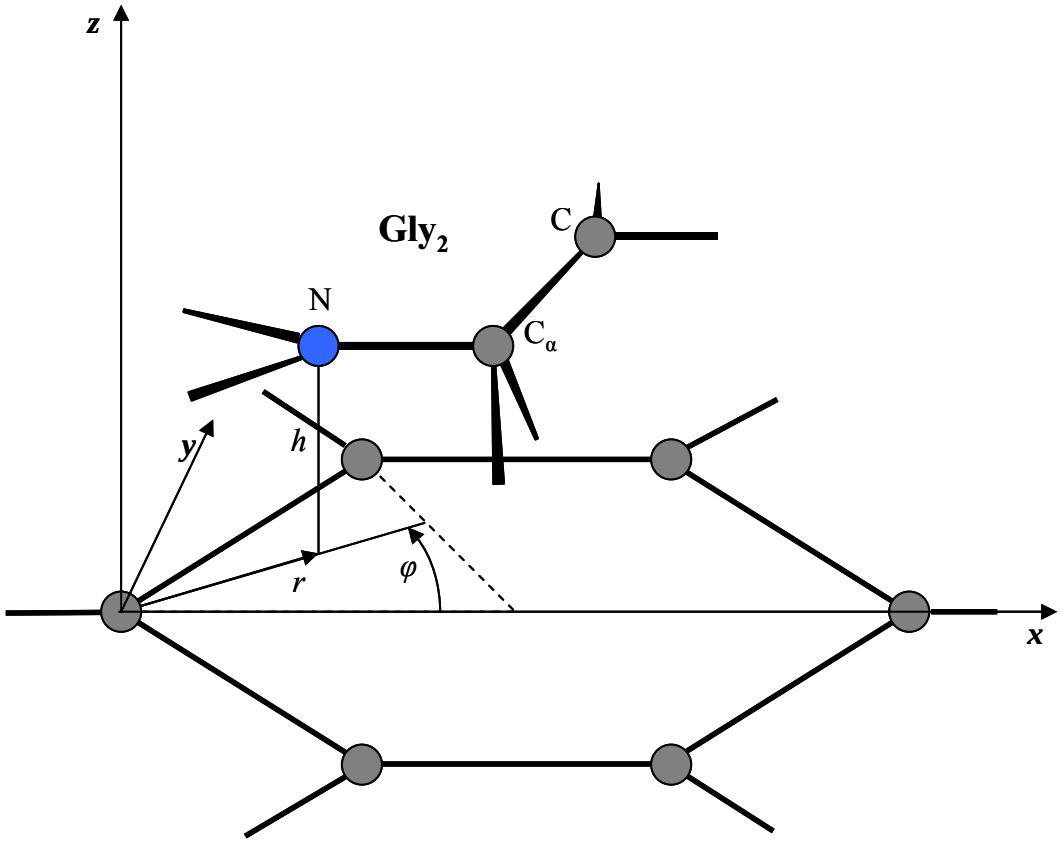


Figure 7.5 Illustration of additional degrees of freedom for simulation of met-enkephalin molecule adsorbed on graphite surface. Only the central residue, second glycine, is shown. Added degrees of freedom describe the position of N atom of the second glycine residue inside the hexagon of carbon atoms: distance of perpendicular projection of N atom from the origin of coordinate system, r, angle between the x-axis and N atom perpendicular projection vector, φ, and the distance of N atom from the surface, h.

Biggs, 2006). Degrees of freedom used here are derived from our study of polyalanine adsorption on smooth surfaces. Introduction of structural details of the solid surface has led to extension of the set of variables by position of the pivotal atom inside the graphene hexagon, as well as the change of central rotation atom from N atom of the first residue to N of the residue from the middle of the sequence (Figure 7.5). The latter modification is introduced only for the purpose of achieving higher efficiency and does not otherwise affect calculation of adsorption energies. A significant difference in comparison to our previous applications of EA appears in the central step of evaluation of fitness function. As described above, fitness function in LD-EA method is expressed as total free energy of a solute molecule in a solvated environment, rather than its intramolecular potential energy.

The parametric study of EA performance in prediction of met-enkephalin 3D structure in gas phase has been used as a guide for choosing optimal EA control parameters for energy minimisation in water solution. Although the optimal set of control parameters for gas phase minimisation has been obtained using capped molecule, lack of corresponding parameterisation procedure for met-enkephalin in zwitterionic form has led us to use the same set of parameters for both met-enkephalin structures investigated in this study.

Table 7.1 Design and control parameters for evolutionary algorithm

Design parameters	
Evolutionary algorithm type	Steady-state
Encoding type	Real
Crossover type	Multipoint
Parent selection strategy	Uniform
Control parameters	
Convergence criterion	0.0001
Generational gap	1
Exponential replacement factor	0.1
Stop range	5000
Population size	500
Mutation probability	0.1
Truncation selection parameter	0.1
Number of crossover points	4
Crossover probability	0.1

The gas phase parameterisation study has indicated that there is no single set of universally applicable optimal control parameters, but a range of the optimal sets which should be applied based on the success criterion and expected precision of the algorithm. Thus, if one aims for structural matching of lower accuracy between an EA outcome and the structure that corresponds to presumed global energy minimum, high values of mutation probability should be utilised. On the other hand, if a structural matching of very high precision is required, or if the EA is expected to find structures with as low energy as possible, mutation probabilities should be small. Being our first study of LD-EA method and not knowing the relationship between RMSD and energy difference for solvated structures, our decision was to pursue a rigorous energy minimisation procedure as it should exploit the full potential of EA approach. Accordingly, the set of control parameters was adjusted for energy minimisation with strict definition of successful outcome. EA parameterisation in gas phase has shown that the optimal parameter set for such a demand is the one shown

in Table 7.1 and Table 7.2 provides the parameters used for Langevin dipole part of the algorithm.

Table 7.2 Langevin dipole method parameters used in the study

Parameters of Langevin dipole part of the LD-EA method	
Inner grid node distance	1 Å
Outer grid node distance	3.1043 Å
Outer grid dipole moment	0.26 eÅ
Temperature	298.15 K
Inner grid thickness	2.0 Å
Neighbouring dipoles exclusion distance	2.5 Å
Lower neighbour inclusion cutoff distance	6.0 Å
Upper neighbour inclusion cutoff distance	18 Å
Electrostatic field threshold for inclusion of dipoles	0.0015 e/Å^2
Electrostatic field threshold for iterating dipoles	0.0021 e/Å^2
Convergence criterion	0.001
Number of iterating points for averaging	10
Number of random positions for a single structure	10

7.4. Results and Discussion

7.4.1. Capped Met-enkephalin in Gas Phase and Water Solution

Capped met-enkephalin in gas phase or vacuum has already been investigated in great detail in our study of EA performance with different force fields. However, since application of Langevin dipole method requires utilisation of Amber atomic charges, only result obtained for Amber PE model are of immediate interest for comparisons. Optimal met-enkephalin conformation in vacuum, as calculated by Amber set of equations, had total intramolecular energy of -76.096 kcal/mol. The structure associated with this energy is shown in Figure 7.6. The right hand side of the figure clearly shows that the backbone is double folded into a β-turn structure

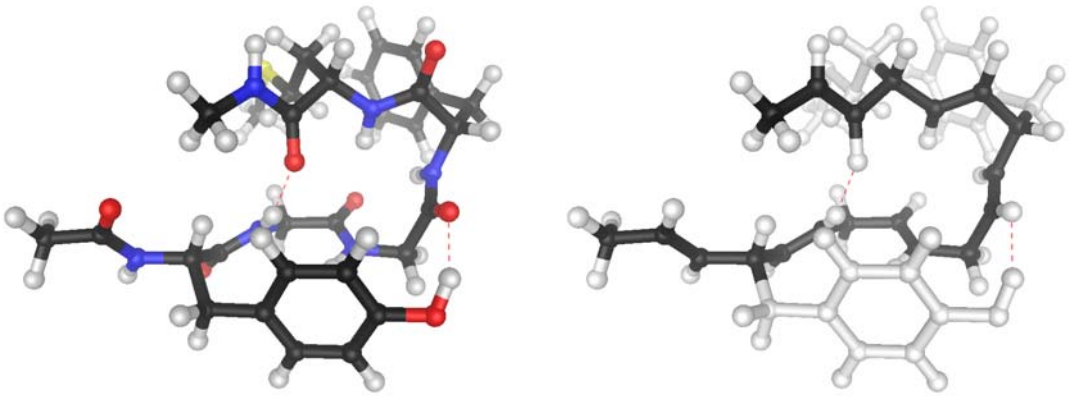


Figure 7.6 Met-enkephalin structure with the lowest intramolecular energy coloured by element (left) and with the emphasised backbone (right).

stabilised by hydrogen bonds between NH-group of the first glycine residue and CO group of methionine, as well as by an additional hydrogen bond between CO group of the second glycine residue and OH-group from tyrosine side chain. Decomposition of energy terms for the structure, as well as for the other structures found to be optimal in other environmental conditions, is given in Table 7.3. The main contribution to overall intramolecular energy in the gas phase appears to be that of electrostatic interactions. Since Amber PE model does not include hydrogen bonds explicitly, their stabilising effect is captured through electrostatic and van der Waals interactions of involved atoms.

Introduction of solvent has major impact on conformation of capped met-enkephalin molecule, as shown in Figure 7.7. The minimal energy structure no longer exhibits turns. Instead of two parallel extended legs, backbone is now folded into a helical structure with hydrogen bonds established between CO-group from acetyl cap and NH-group of phenylalanine, as well as between CO of tyrosine and

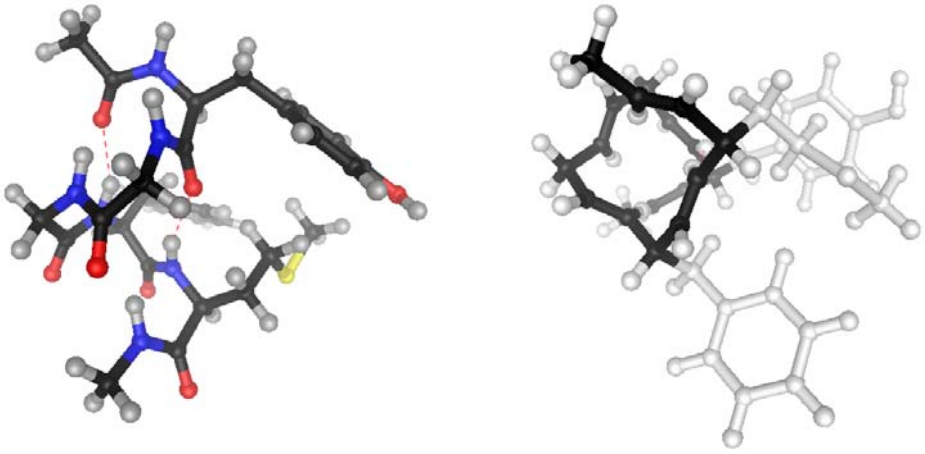


Figure 7.7 Capped met-enkephalin structure in water solution – side view of structure coloured by element (left) and frontal view of the emphasised backbone (right).

NH-group of methionine. Although CO and NH groups from the first glycine residue and amino-methyl cap, respectively, do not satisfy geometric requirements for hydrogen bond, their O and H atoms are separated by only 2.503 Å. Distance between O atom from the second glycine and H atom from NH group of amino-

methyl cap is somewhat higher (3.104 Å), thus placing NH group in between CO groups from the two residues. Since amino-methyl cap is at the end of the sequence, it is not surrounded by residues from both of its ends, thus being more mobile than regular inner residues. This increased flexibility and attraction of NH by CO-group from the second glycine is a probable reason for deviation of hydrogen bond between the first glycine and amino-methyl cap. Nevertheless, the two existing hydrogen bonds are sufficient for establishing a pattern of bonds between CO-group of residue *i* and NH-group of residue *i*+4 – a pattern that determines α-helix. The transition from β-turn to α-helix is further illustrated in Ramachandran plot of gas phase and solution conformations of capped met-enkephalin, shown in Figure 7.8. While dihedral angle pairs for vacuum structure are dispersed along a broad region, solvated structure is characterised with their concentration in a much narrower area of α-helix.

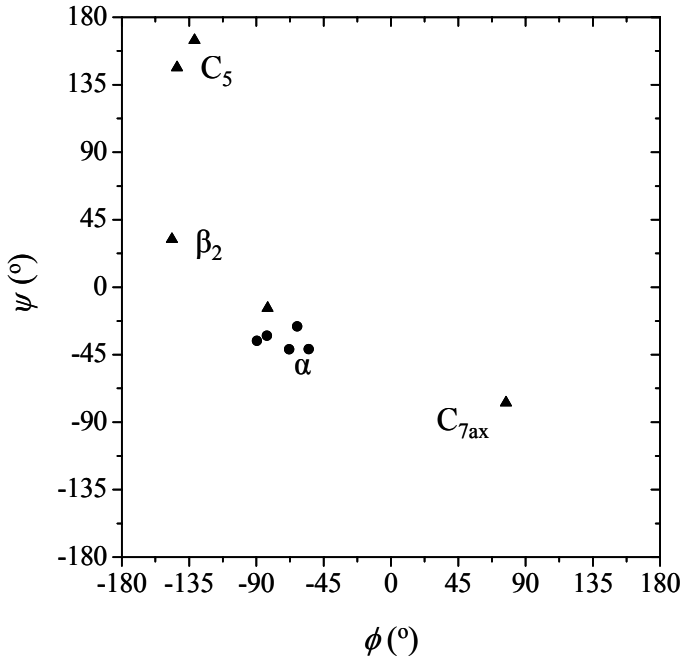


Figure 7.8 Concentration of dihedral angles in the region of α-helix with the change of environment from vacuum (triangles) to solvent (circles).

Table 7.3 Decomposition of energies of two forms of met-enkephalin in three environments considered in this study

	Energy terms											
Case[a]	U^{tor}	U^{es}	U^{vdW}	U^{g}	ΔG_{phob}	$\Delta G_{es(c)}$	$\Delta G_{es(LD)}$	ΔG_{vdW}	$\Delta\Delta G_{s}^{surf}$	ΔG_{s}	E^{surf}	G
CV	15.090	-87.272	-3.913	-76.095	-	-	-	-	-	-	-	-76.095
ZV	9.299	-170.297	10.584	-150.414	-	-	-	-	-	-	-	-150.414
ZVG	9.387	-163.593	9.913	-144.293	-	-	-	-	-	-	-54.738	-199.031
CS	13.101	-69.521	-3.791	-60.211	3.784	-0.185	-30.135	-16.185	-	-42.721	-	-102.932
ZS	11.265	-64.586	1.143	-52.178	2.199	-0.353	-117.006	-16.837	-	-131.997	-	-184.175
ZSG	9.236	-65.799	6.275	-50.288	2.451	0.032	-108.558	-168.809	81.041	-193.843	-0.888	-245.019

a **CV** – capped molecule in vacuum; **ZV** – zwitterion in vacuum; **ZVG** – zwitterion adsorbed on graphite in gas phase (vacuum); **CS** – capped molecule in solvent; **ZS** – zwitterion in solvent; **ZSG** – zwitterions adsorbed from solvent

There is a remarkable similarity in solvation induced conformational changes between capped met-enkephalin and alanine dipeptide, studied with our LD-Amber model. As a reminder, water solvation of alanine dipeptide promotes switch from equatorial C_7 conformation to right-handed α-helix. Although none of met-enkephalin residues exhibits $C_{7,eq}$ conformation angles, most of dihedral angle pairs of gas phase molecule are found in the same Ramachandran plot quadrant as $C_{7,eq}$. Upon introduction of water, however, the most stable conformation appears to be closest to right handed α-helix, α_R. An explanation for this phenomenon should be sought for in exposing partially charged protein atoms to favourable interactions with surrounding water molecules. An illustrative example is solvation induced breaking of hydrogen bond between tyrosine OH-group and CO-group from the second glycine residue. In the absence of water, favourable electrostatic interactions are established between partially positively charged H atom and partially negative O atom from CO-group. However, introduction of water enables exposure of these atoms to oppositely charged atoms from solvent molecules, thus compensating for the decrease of stability caused by intramolecular hydrogen bond breaking. View from bottom in Figure 7.7 (right hand side) clearly shows that all side chain groups are stretched away from the backbone and into the solvent, which facilitates their solvation.

Hydrogen bonds in the backbone seem to suffer only rearrangements, which should not affect intramolecular energy substantially. Overall effect of transition from optimal structure in vacuum to α_R-helix in water solution is increase in intramolecular energy for more than 15 kcal/mol (Table 7.3). The bulk of the effect is achieved through electrostatic component, which increases from -87.272 to -69.521 kcal/mol, (i.e. ~20%). This is, however, balanced by a large negative value of the free energy of solvation. The largest contribution to the free energy of solvation (~70%), as calculated by Langevin dipole model, originates in interactions of dipoles from inner solvation layers with permanent electrostatic field of the solute molecule. Apart from solvation of individual partial charges, LD model may reproduce formation of water-mediated hydrogen bridges established between two oxygen or two hydrogen atoms, respectively. This phenomenon has been described by Beglov and Roux (Beglov and Roux, 1995) who used molecular dynamics and atomistic

water model to show formation of water bridges between the two O atoms of the CO-groups and between the two H atoms of the NH-groups of alanine dipeptide in left handed α-helix conformation, α_L. Our study of the LD-Amber method applied to alanine dipeptide in α_L conformation has shown that even if the level of description of water is reduced to Langevin dipoles, the model can still predict establishing of water-mediated hydrogen bridges between like charged sites in the molecule (Mijajlovic and Biggs, 2007b). The same analysis can be applied for larger molecules, such as met-enkephalin. Figure 7.7 shows several structural elements in which consecutive CO and NH-groups are oriented in the same direction, thus exposing their O and H atoms to serve as a base for formation of water bridges. An example of such a water bridge, formed between O atoms from aligned CO groups of the second glycine and phenylalanine residues can be seen in Figure 7.9. The figure is a 3-dimensional view of electrostatic field formed exclusively by Langevin dipoles distributed around met-enkephalin molecule, i.e. the effect of point charges from the solute molecule is extracted from the overall electrostatic field. The 3D space is cut by a plane that passes through the oxygen atoms of interest (designated by O in the figure). The blue colour in the cutting plane indicates positive electrostatic potential. Since this potential is generated only by solvent, it indicates increased concentration of solvent originating positive charges, i.e. hydrogen atoms. Although Langevin dipole method does not operate with explicit water molecules, we have shown earlier that analysis of electrostatic field generated by dipoles allows indirect derivation of positions of water molecules in the first solvation layer. Analogous to the analysis of water bridges in α_L-conformation of alanine dipeptide, existence of two distinct curved regions with positive electrostatic potential leads to conclusion that these belong to hydrogen atoms from two water molecules involved in hydrogen bridging between the two oxygen atoms from solute. Faint red areas in the upper part of the figure belong to domains of the space below the cutting plane and correspond to increased concentration of water oxygen atoms around hydrogen atoms from NH-groups on the back side of the solute. Although not clearly visible from this perspective, these regions also indicate formation of water-mediated hydrogen bridges between NH-groups.

The ability of the Langevin dipole method to represent solvent structuring is probably a more general phenomenon, i.e. it should not be restricted to water alone. Whilst in water solutions, it is reasonable to relate the structural changes to the formation of new hydrogen bonds, the LD method is able to operate with any solvent whose molecules have finite dipole moment, irrespective of its ability to form hydrogen bonds. It should be noted that dipoles themselves do not have the ability to engage into hydrogen bridging. The bridging concept is invoked here only because it has been shown in Chapter 6 that dipole restructuring obtained in the LD model corresponds to establishing of water bridges as seen by the MD simulation. In general case, the dipole restructuring does not have to be constrained to water bridges and is probably common to all dipolar solvents.

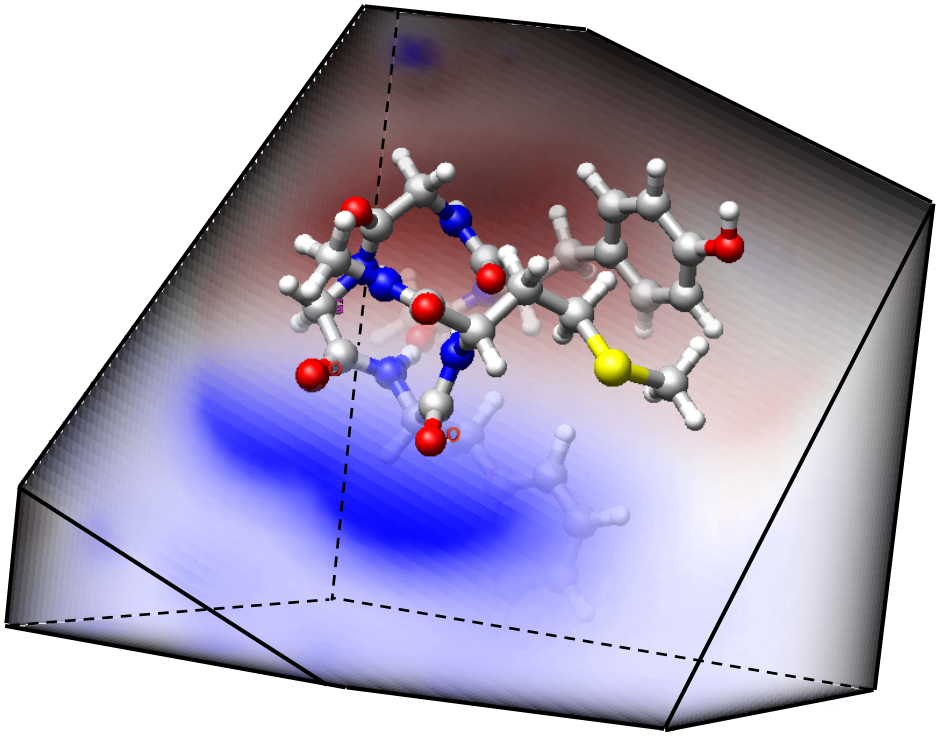


Figure 7.9 Electrostatic field of water dipoles around capped met-enkephalin molecule in solution.

Figure 7.7 shows that apart from CO-groups of the second glycine and phenylalanine residues, there are several more combinations of both CO and NH-groups in an orientation suitable for formation of water bridges. For example, this kind of stabilisation can also be established between CO-groups from the first and the second glycine residues, as well as between corresponding groups from

phenylalanine and methionine. Although some other CO-groups from consecutive residues are also in a favourable mutual position, formation of water bridges between them is hampered by side chains or other parts of the backbone. Similarly, NH-groups of the acetyl cap and the tyrosine residue, as well as of tyrosine and the first glycine are in a suitable position for formation of water bridges with O atoms from water molecules facing the solute. This is a clear contrast to the conformation of the molecule in vacuum, where, as Figure 7.6 shows, CO or NH-groups from consecutive residues are usually turned in opposite directions, thus disabling formation of water bridges if such a structure were introduced into water solution. The ability to form numerous water bridges and thus compensate the increase in intramolecular energy is, hence, the most plausible explanation of conformational change that occurs with solvation of met-enkephalin molecule by water.

7.4.2. Met-enkephalin Zwitterion in Gas Phase and Water Solution

Zwitterionic form of met-enkephalin is characterised by replacement of acetyl and amino-methyl caps by NH_3^+ and COO^- ionised caps, respectively. While still remaining in an electroneutral state, with net zero charge, this form of the molecule features NH_3^+-group on its N-terminus combined with COO^--group on C-terminus. This separation of charges has a profound effect on met-enkephalin conformation, especially in vacuum, where the environment does not provide any screening between the charged ends. Since the ends are oppositely charged, electrostatic forces cause strong attraction between them. On the other hand, van der Waals repulsive forces limit the number of possible configurations in which the two termini can be found on the small distance. The balance between the attractive electrostatic forces and repulsive steric interactions is accomplished by folding the backbone of the molecule into a loop, shown in Figure 7.10. Visual comparison between this and the structure shown in Figure 7.6 shows a remarkable similarity between the optimal vacuum conformations of capped and zwitterionic met-enkephalin form. Strong electrostatic attraction between the charged ends results in higher degree of folding towards the termini in the zwitterionic structure, as well as modification of positions of some of the side chains (most obviously expressed for tyrosine side chain), but the rest of the backbone appears to be folded in a conformation very similar to the one in vacuum. A relatively small RMSD of 0.541 Å between the two structures further

confirms this observation. It should be noted that, since our RMSD calculation is based on determining the distances between corresponding C_α atoms, this procedure can be applied to the two forms of the molecule since removing end groups does not strip the molecule from its C_α atoms. A very small RMSD (lower than the threshold value used to define similar structures in our study of EA performance with different force fields), indicates that, although important for bending of protein termini, electrostatic interactions between the ends may not be crucial for folding the rest of the backbone. As a comparison, the RMSD between two capped structures, in vacuum and in solvent, is much higher, with a value of 2.353 Å, while the deviation between optimal zwitterionic forms in vacuum and in water is even higher at 3.212 Å. Apparently, the environment exhibits much stronger influence on conformation than removal of end groups and ionisation of termini. This can be explained by effect of solvent to electrostatic interactions throughout the whole length of the molecule, while end ionisation affects only small parts of it.

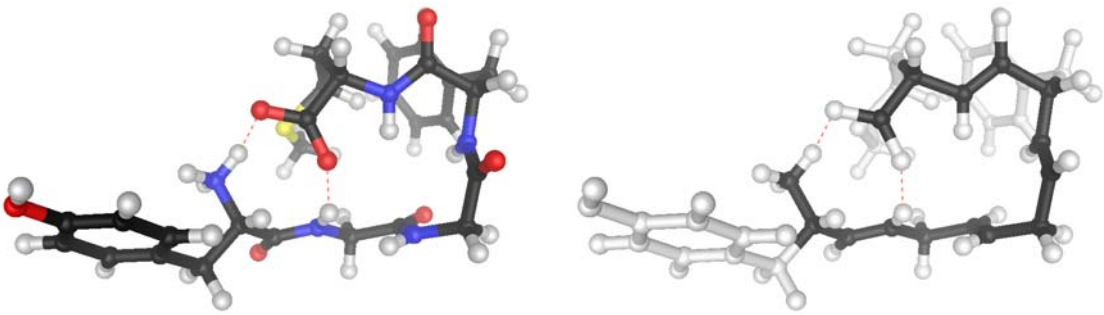


Figure 7.10 Zwitterionic met-enkephalin conformation in vacuum coloured by element (left) and with emphasised backbone (right).

While there is a strong degree of similarity between capped and zwitterionic structures in vacuum, solvated conformations of the two forms are substantially different, with RMSD of 2.625 Å. The conformation of energetically most favourable zwitterionic met-enkephalin molecule in water is shown in Figure 7.11. Although completely different than the capped form, this result is in a good agreement with other simulation studies which suggest that met-enkephalin in water solutions is found in highly flexible conformation with extended backbone (Kinoshita et al., 1998). On the other hand, the structure proposed here is not

completely identical to any of the experimentally found solvated conformations (Roques et al., 1976; Jones et al., 1977; Khaled et al., 1977; Spirtes et al., 1978; Graham et al., 1992). However, experimental studies themselves produced results in a wide range of conformations and most of them agree that met-enkephalin in dilute solutions is found in an unfolded and very flexible conformation. The unfolding aspect is in a very good agreement with our results.

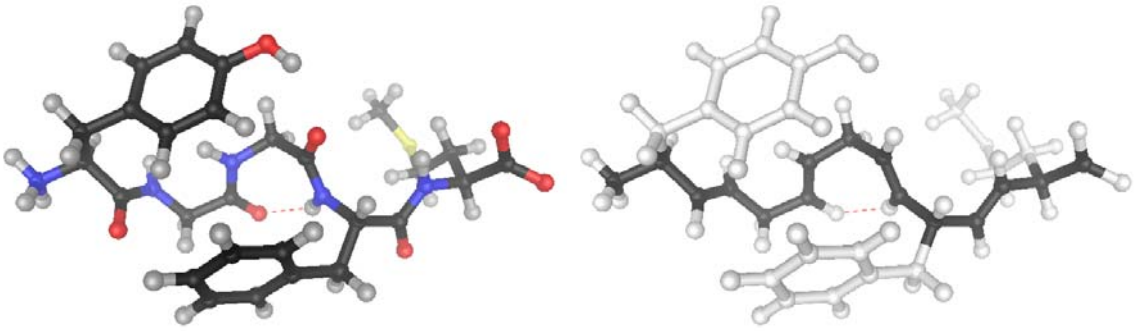


Figure 7.11 Zwitterionic met-enkephalin in water solution coloured by element (left) and with emphasised backbone (right).

Evolutionary algorithms do not, of course, offer insight into the flexibility of the molecule as they only allow identification of the global minimum. However, an indirect indicator of conformational flexibility is an apparent lack of intramolecular hydrogen bonds as well as water-mediated hydrogen bridges. The only intramolecular hydrogen bond is established between the CO-group of the first glycine and NH-group of the phenylalanine residue. Further to that, number of pairs of CO- and NH-groups that are in orientation suitable for formation of water bridges is substantially lower than in optimal conformation of capped molecule in solution. The only CO-pair that can serve as a template for water bridge consists of CO-group from phenylalanine and one part of the carboxyl end group, while the only two suitable NH-groups are one part of the ammonium group at N-term and NH-group of the first glycine residue. If the number of hydrogen bonds and water bridges was

higher, the flexibility of the molecule would be severely reduced, which would create disagreement with experimental findings.

As expected, Table 7.3 shows that the highest contribution to overall energy of zwitterion in gas phase is that of electrostatic interactions, which is explained by small distance between the charged termini of the molecule. Intramolecular electrostatic interactions, however, diminish in the presence of solvent due to the significant degree of separation between the charges (found on the opposite ends of extended molecule). The level of reduction in magnitude of charge-charge interactions is so great that electrostatic energy contribution to overall potential energy is even lower than in capped form of the molecule, which has much lower atomic charges. Nevertheless, this energetic loss is balanced by increase in electrostatic interactions with surrounding solvent. The magnitude of interactions between Langevin dipoles in inner solvation layers and charges from the solute is so strong that it is more than two times higher than overall intramolecular potential energy. Although contribution of bulk solvent, $\Delta G_{es(c)}$, to the overall solvation free energy is very small, the table shows that it is about two times higher in magnitude for solvated zwitterion than for capped molecule. This can, again, be explained by the higher degree of separation of charges and formation of a stronger solute dipole. Elongation of the zwitterionic form also contributes to increase in magnitude of the solute dipole. Comparison of hydrophobic terms for zwitterionic and capped form shows that zwitterionic met-enkephalin is more hydrophilic, again probably due to existence of strongly charged groups at its ends.

7.4.3. Met-enkephalin Zwitterion Adsorption on Graphite

As discussed above, backbone conformation of met-enkephalin in vacuum is very similar for both capped and zwitterionic forms. It is, therefore, expected that when the adsorption of the molecule is conducted from the gas phase, the resulting structures will be similar for both forms. Although presence of solvent introduces significant difference in folding pattern of capped and zwitterionic met-enkephalin, computational constraints have limited our choice to zwitterionic form as it has higher biological significance and is more often used in experimental studies.

Met-enkephalin structure adsorbed on graphite in vacuum is shown in Figure 7.12 in top and side view. Although visually very similar to vacuum conformation

illustrated in Figure 7.10, the RMSD between the two structures is 1.093 Å. The reason for the high RMSD despite high visual similarity between the two conformations is a stronger effect of surface interactions on positions of C_α atoms, which are used for RMSD calculation. C_α atoms are expected to be more susceptible to surface induced deviation since they are anchoring points for side chain groups. Side chains have high degree of flexibility in vacuum, but, as Figure 7.12 shows, are constrained to positions parallel to the surface in adsorbed molecule. The translocation of side chains causes distortion of the backbone that is stronger in C_α positions than in positions of neighbouring N and carboxyl C atoms. Consequently, the overall shape of the backbone remains similar to that in vacuum, but RMSD is high due to changed C_α positions.

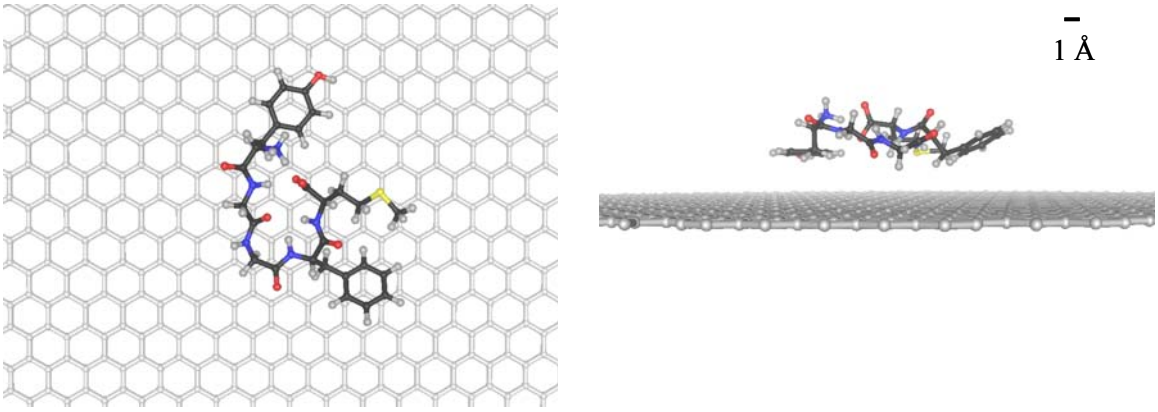


Figure 7.12 Zwitterionic met-enkephalin molecule adsorbed on graphite from gas phase: view from top (left) and side view (right).

Energy decomposition, shown in Table 7.3, reveals that overall intramolecular potential energy is increased for only about 4% compared to the optimal vacuum structure. The main source of energy change is in electrostatic interactions which increase from -170.297 kcal/mol to -163.593 kcal/mol (i.e. ~3.9%). Small variation in electrostatic energy is consistent with the analysis of conformational changes discussed above. Most of the backbone transformation stems from changes in C_α positions, while positions of its backbone neighbours, N and C atoms of peptide bond, undergo smaller variations. Since magnitude of point charge assigned to C_α atoms is much lower than for N and C atoms, modification of electrostatic interactions is also lower for shifting C_α than it would be for N and C backbone atoms. Majority of atoms in the side chain groups carry charges of low intensity and have small effect on electrostatic energy of the molecule. The most notable

exceptions are O and H atoms from tyrosine side chain and S atom in side chain of methionine. However, these side chains are separated in a vacuum and, despite conformational change, remain separated upon adsorption. Adsorption, therefore, does not introduce any significant changes in electrostatic interactions between these two residues. Distance between O and H atoms of tyrosine, due to fixed bond length, remains constant during the adsorption process. Interactions between all side chains and the backbone do not suffer significant variations as in both free and adsorbed molecule, side chains are stretched away from the backbone.

Van der Waals interactions with the surface are very favourable and their magnitude is more than sufficient to offset the decrease of stability caused by increase in intramolecular potential energy. A notable feature of the adsorbed structure is, as expected, alignment of aromatic rings of tyrosine and phenylalanine with the surface, as can be seen in Figure 7.12. This is clearly visible for tyrosine ring which is virtually parallel to the surface, thus substantially increasing the

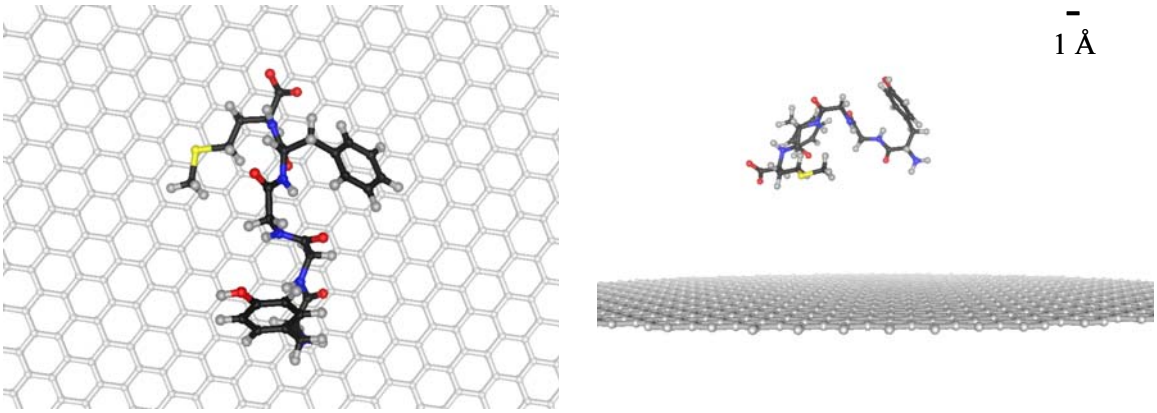


Figure 7.13 Zwitterionic met-enkephalin molecule adsorbed on graphite from dilute water solution: view from top (left) and side view (right).

magnitude of surface interactions.

Adsorption of zwitterionic met-enkephalin in the presence of solvent is substantially different from its adsorption in vacuum. Position of the molecule above the graphite surface is shown in Figure 7.13. A striking difference in comparison with vacuum adsorption is that the molecule is no longer attached to the surface. Minimal and average distances of met-enkephalin atoms from the surface in vacuum are 2.327 Å and 4.120 Å, respectively. However, corresponding distances in the presence of water are 8.662 Å and 12.836 Å, which indicates distribution of several

solvation layers between the surface and the solute. Assuming that the distance between layers corresponds to the position of the first peak in O-O radial distribution function of liquid water, the solvation layer distance is estimated to be about 3 Å (Narten et al., 1967; Narten, 1972; Jorgensen, 1981), which indicates that three to four layers of water molecules can be placed between met-enkephalin and graphite. This is somewhat surprising outcome considering the hydrophobic character of graphite surface and favourable interactions between graphite and met-enkephalin, especially between the aromatic rings of the two.

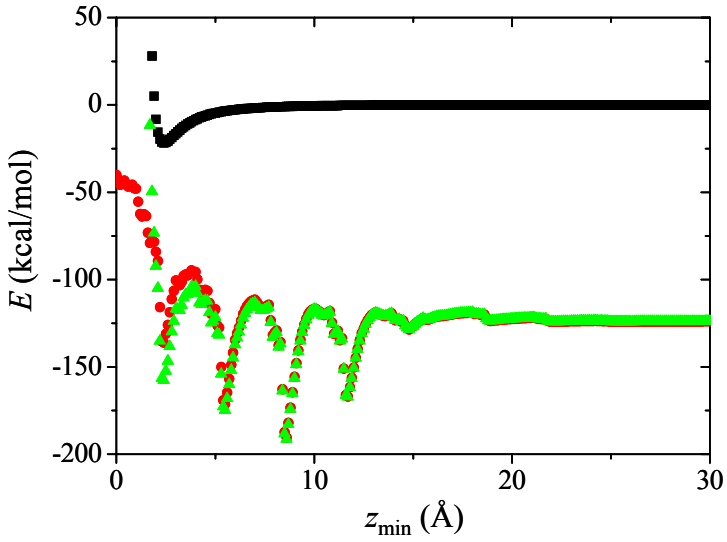


Figure 7.14 Change of surface interaction (■) and free energy of solvation (●) with distance between met-enkephalin and graphite surface. The sum of the two (▲) shows that solvation effects are dominant in this coupling.

In order to verify this result and elucidate the behaviour of solvent in the space between the solid surface and the solute molecule, we have systematically varied the distance between the surface and the peptide, keeping met-enkephalin in rigid conformation and fixed orientation with respect to graphite planes. The distance between the surface and the closest met-enkephalin atom has been gradually increased from 0 to 30 Å in steps of 0.1 Å. Since the conformation of the molecule is fixed, the only terms that remain susceptible to change during distance variation are adsorption energy and energy of solvation (including the term for change in surface solvation energy, $\Delta\Delta G_s^{surf}$. Functional relationship between energies and distance between the surface and protein's closest atom is shown in Figure 7.14. Adsorption energy, or sum of van der Waals interactions between the protein and the surface,

changes in a way similar to classical Lennard-Jones potential, i.e. with a steep increase in strength of repulsive term with small distances and slower increase in attractive term for increasing distance of met-enkephalin from the surface. Combination of the two terms creates a function with a single minimum, at a distance of about 2.4 Å. Value of surface energy at this distance is -21.7 kcal/mol.

Free energy of solvation, on the other hand, has far more complex behaviour as its change with distance from the surface is characterised with multiple local minima. Four of the local minima dominate in this function's landscape: (2.4 Å, -136.0 kcal/mol), (5.5 Å, -171.3 kcal/mol), (8.6 Å, -190.4 kcal/mol), and (11.7 Å, -167.0 kcal/mol). The distance between the minima is 3.1 Å, which closely corresponds to the separation between the solvation layers. Thus, each minimum corresponds to insertion of a single solvation layer. The third local minimum is characterised with the lowest solvation energy, which leads to conclusion that optimal position of met-enkephalin above the graphite surface is the one which leaves average number of three solvation layers between them. Since the magnitude of surface energy is considerably lower than that of the free energy of solvation, the latter term dominates the sum and optimal position of the molecule corresponds to global minimum of solvation free energy. This ordered insertion of solvation layers is phenomenologically very similar to structuring of water layers during water adsorption in graphite pores (Ulberg and Gubbins, 1995), which may suggest that met-enkephalin molecule plays a role analogous to that of a pore wall in this system.

It is interesting to note that solvation energy decreases almost steadily after the fourth minimum, i.e. after four solvation layers have been inserted between the molecule and the surface. Any new solvation layers do not contribute significantly to energy of solvation. In order to get a better understanding of this phenomenon, total solvation energy is decomposed for each position of met-enkephalin above the surface. Figure 7.15 shows how each of the energy terms changes with the distance. While hydrophobic and electrostatic contributions change almost continuously, van der Waals interactions and changes in surface solvation energy show strong local extrema with addition of each new solvation layer. When protein and solid surface are on a small distance from each other, inserted layer engages in van der Waals interactions with both of them, which substantially increases magnitude of van der

Waals energy. At the same time, each new inserted layer will disturb surface solvation layers distributed over graphite when met-enkephalin is not present. This disturbance of surface solvation layers causes decrease in magnitude of surface solvation energy, which manifests as jumps in $\Delta\Delta G_s^{surf}$.

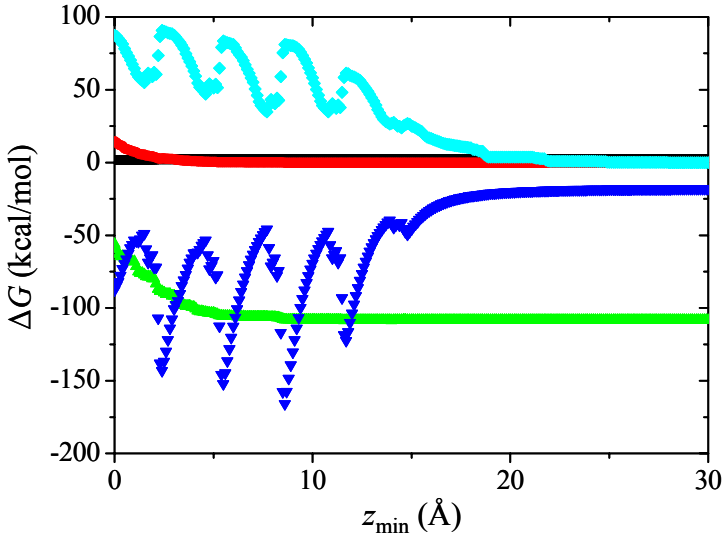


Figure 7.15 Change of individual terms of solvation energy with distance of met-enkephalin from graphite surface: ΔG_{phob} (■), $\Delta G_{es(c)}$ (●), $\Delta G_{es(LD)}$ (▲), ΔG_{vdW} (▼), and $\Delta\Delta G_s^{surf}$ (♦).

Comparison of the separation between the minima of the solvation free energy (Figures 7.14 and 7.15) and the parameters of the LD-EA model shows that the distance between the minima is very similar to the distance between nodes of the coarse grid of the LD model. The results collected in this study are not sufficient to make a decisive conclusion whether the observed minima separation is an artefact of the chosen grid representation. Further studies, in which different grid geometries (e.g. tetrahedral) and node distances will be examined, are expected to help in rectifying the situation.

7.4.4. Computational Cost of the LD-EA Method

Met-enkephalin simulations, irrespective of the form of the molecule used (capped or zwitterionic), require on average between 48 and 72 hours of wall time for a single EA run. Since all calculations are performed with 11 CPUs, this translates to 3-4 weeks of CPU time. Although this computational cost does not appear to be so small, it should be noted that utilisation of full atomistic solvent

models for calculation of EA fitness function would be at least an order of magnitude more expensive (Mijajlovic and Biggs, 2007b). Furthermore, we expect that further optimisation of evolutionary algorithm (such as implementation of adaptive control parameters) will significantly reduce CPU times needed for a single simulation.

7.5. Conclusions

The LD-Amber method developed previously has proven as a very fast and reliable technique for calculation of solvation free energies of amino acid residues and small proteins in different conformations. This part of our work focuses on extending its application to development of an evolutionary algorithm based global minimisation method that uses individual LD-Amber calculated energies of solvation for evaluation of fitness function of specific protein conformations. Being a combination of LD-Amber and EA techniques, we have designated the new method as LD-EA. To our knowledge, no similar techniques that combine Langevin dipole with evolutionary algorithms have been developed.

Apart from designing a completely novel method, we have also utilised it in a much more complex system than the systems we used in our previous LD-Amber study. The new method has been applied to evaluate solvated conformation of met-enkephalin molecule in its zwitterionic, as well as form capped with acetyl and amino-methyl groups on N- and C-terms, respectively. The results obtained show significant degree of conformational change in the process of solvation and are in a good qualitative agreement with experimental and other simulation studies.

Further development of LD-EA approach has been accomplished by expanding it into a new environment – system consisting of protein, water and solid surface as a substrate for protein adsorption. Met-enkephalin molecule in its zwitterionic form has been simulated in contact with graphite surface both in vacuum and in water solution. Vacuum based adsorption results in a conformation whose backbone and side chains are aligned with the surface and on a small distance from it. This can be explained by increase of magnitude of favourable protein-surface interactions. Adsorption from water solution, however, produces somewhat unexpected result. Although graphite surface is supposed to be hydrophobic and, therefore, attract met-enkephalin more favourably than it attracts water, the protein is not attached to the surface as we expected. Rather than that, simulations show that the optimal position

is accomplished with an average of three solvation layers between the surface and the molecule. One of the explanations is that solvation of met-enkephalin has a dominant effect over protein-surface interactions. There is, however, a possibility that Langevin dipole parameters for sp^2 hybridised carbon atoms have to be readjusted for their application in smooth solid surface.

Chapter 8. Conclusions and Future Work

8.1. Summary of Major Findings

Chapter 4 describes the study of the influence of EA fitness function on EA performance and the choice of optimal control parameters. Different fitness functions have been represented with four PE models commonly used in protein conformation studies. It has been shown that the choice of a PE model can profoundly affect the performance of the EA, changing the number of potential energy evaluations for up to two times. It has also been discovered that different PE models are associated with different sets of control parameters that provide optimal performance. An important finding of the study indicates that the set of optimal control parameters is not only bound to the fitness function being optimised, but also to the required level of accuracy. A detailed investigation using the Amber PE model (Cornell et al., 1995) has shown that increasing the required level of accuracy for an order of magnitude causes optimal mutation rate to decrease from values close to 1 to almost 0. The same change also causes the number of necessary PE evaluations to increases for about 30 times.

In Chapter 5, an EA approach has been applied in predicting the 3D structure of polyalanine molecules of different length adsorbed on smooth surface modelled with the Steele potential (Steele, 1974). The adsorption is studied in the gas phase, i.e. on solid-gas interface. It was concluded that, despite expected gradual change of conformation with continuous increase in strength of protein-surface interaction, the polyalanine molecules switch from one conformation to the other when a surface interaction threshold is reached. It was found that polyalanine adsorbs in one of the three conformations – right-handed α-helix, 3_{10}-, and 2_7-helix – which are all characterised with a specific hydrogen bond pattern established between CO- and

NH-groups of the backbone. Investigation of the behaviour of polyalanine molecules with different numbers of residues has shown that the switching point for each molecule depends on its size. This has implications in potential industrial application of the switching phenomenon as it allows design of molecules that will undergo conformational changes at prespecified values of protein-surface interaction strength. The effect of length on the value of switching point can also be utilised in industrial separation of molecules based on their lengths.

The Langevin dipole model (Florián and Warshel, 1997) has been shown to predict the solvation free energies accurately and with low computational cost. However, the original model is based on solute atomic charges calculated from quantum mechanical (QM) methods. QM calculation of charges is very time consuming and necessity to recalculate the charge distribution for every conformation of the solute makes it inapplicable in evolutionary algorithm based protein structure prediction. The study described in Chapter 6 shows that the LD model coupled with atomic charges adopted from the Amber PE model (LD-Amber) does not suffer from any deterioration in accuracy. It was found that the free energies of solvation of amino acid side chain analogues calculated by the LD-Amber method are, in most cases, very close to experimentally calculated values, and, in general, no worse than the results obtained using more sophisticated, explicit solvent model. Application of the LD-Amber method on a small alanine-dipeptide molecule in a range of its conformations has shown that, in addition to being able to operate with different amino acids, the method is capable of providing good results on a single molecule in different 3D structures. It has also been shown that, whilst being up to two orders of magnitude faster than explicit solvent models, the LD-Amber method can still predict solvent restructuring – something that would be impossible with implicit solvent models.

Chapter 7 describes implementation of an EA based approach with the LD-Amber calculated solvation free energy. The method obtained by coupling EA with the LD-Amber was termed LD-EA. LD-EA has been tested on prediction of solvated 3D structure of met-enkephalin molecule in its zwitterionic and capped forms. The results obtained with the capped molecule have shown substantial differences compared to the results collected for the same molecule in vacuum. Zwitterionic

form has been observed in an extended conformation in water solution – a good qualitative agreement with experimental results for the 3D structure of solvated met-enkephalin in zwitterionic form. It should be noted, though, that experimental studies do not offer a single conformation for met-enkephalin in water solutions, but a set of structures similar to that obtained in our study. The LD-EA method has also been used to investigate the 3D structure of zwitterionic met-enkephalin molecule on the graphite-water interface. The implementation of the new method has shown that, rather than closely adsorbing to the graphite surface, met-enkephalin molecule is found in its vicinity, but with three solvation layers between the surface and the molecule – a phenomenon similar to water adsorption in pores of microporous graphitic carbons (Ulberg and Gubbins, 1995).

8.2. Overview of the Contribution to the Body of Knowledge

- It was shown for the first time that the choice of the PE model can profoundly influence the EA performance and location of optimal EA control parameters in an EA based prediction of protein 3D structure. This finding is important, as many past EA based protein studies have used an arbitrary set of control parameters without clear understanding of their effect on the EA performance.
- The studies of polyalanine at solid surfaces and met-enkephalin at the graphite-water interface represent the first applications of an evolutionary algorithm in the context of prediction of the 3D structure of proteins at a solid-fluid interface.
- The study of polyalanine at the solid surface has also shown a phenomenon that has never been reported before – conformational switching of the polyalanine molecule induced by the changes in surface interaction energy. The phenomenon can potentially be exploited in emerging technologies, such as nanocomputing and construction of nanomotors.
- We have shown that coupling of the Langevin dipoles (LD) model with solute atomic charges adopted from the Amber PE model creates a very fast computational method with the level of accuracy comparable to explicit solvent representations. The LD-Amber model eliminates the need to conduct expensive QM calculations for evaluation of atomic charges for different conformations of the same molecule. Thus, the LD-Amber model extends the applicability of the original LD model into

numerical methods that otherwise would not be able to cope with its embedded QM charge calculation.

- Contrary to some previous arguments from the scientific community, it has been shown that Langevin dipoles are capable of giving a high level of insight into the restructuring of solvation layers around the solute molecule. We have demonstrated the ability of the LD-Amber method to capture solvent restructuring and formation of water bridges around solvated alanine dipeptide molecule. This phenomenon has previously been observed using molecular dynamic methods (Beglov and Roux, 1995), but with the computational cost almost two orders of magnitude higher than that of the LD-Amber approach.
- For the first time, an LD based calculation of solvation free energies has been used to facilitate calculation of the fitness function in an EA determination of protein 3D structure in solution.
- A novel model for interaction of proteins with a solid-fluid interface has been developed. The new model encompasses evaluation of protein intramolecular potential energy, energy of interaction between the surface and the protein and solvation of both the surface and the protein molecule.

8.3. Future Work

8.3.1. Adaptive Evolutionary Algorithm

Our study of the relationship between control parameters and EA performance has revealed that the mutation probability has stronger effect on performance than any other parameter. It has also been demonstrated that the optimal mutation probability depends on the required level of accuracy of the EA outcome. If an evolutionary algorithm is run with one value of mutation probability, P_M, in the initial stage and with another value or range of values in latter stages, then the initial P_M value will direct the EA to a broad proximity of the global optimum, while latter P_M values will narrow down the search to a very accurate solution.

Following the same principle, it is possible to construct a “self adaptive” evolutionary algorithm, which will autonomously modify the mutation probability during the course of the simulation. The same principle can be applied to the other control parameters. Development of an adaptive EA will significantly reduce

computational time, thus allowing improvement in statistics of the method and enabling the EA approach to be applied to larger molecules.

8.3.2. Calculation of Protein Conformational Entropy and Free Energy

Free energy of a protein in solution is a sum of free energy of the protein and solvation free energy. However, free energy of the protein is currently simplified and approximated with its potential energy. In order to obtain a more accurate fitness function for EA minimisation, protein free energy calculation should be augmented by contribution of conformational entropy. It has also been demonstrated that protein conformational entropy may play an even more significant role during protein adsorption (Liu and Haynes, 2004).

Due to its nature, an EA based approach requires the entropy and free energy to be associated with individual conformations. It is, therefore, necessary to apply an empirically based method for evaluation of conformational entropy based on a single 3D structure (Karplus and Kushick, 1981; Sternberg and Chickos, 1994; Cole and Warwicker, 2002). One of the ways in which the entropic contribution can be calculated is by using a Hessian or the second derivative of the potential energy for a given conformation (Klepeis et al., 2002). Knowing the PE model, the second derivative at a local minimum associated with the conformation can easily be obtained numerically.

8.3.3. Implementation of Protein Ionisation and Polarisation

The current implementation of the EA based protein 3D structure prediction operates with proteins in a single ionised state. Proteins that include ionisable amino acid residues, such as aspartic acid or arginine, change their state of ionisation by protonation and deprotonation of acidic and basic groups. The protonation state is a function of the pH value of the solution. It is, however, also a function of protein conformation. The conformation is, in turn, strongly influenced by the distribution of atomic charges, which depends on protonation state. Consequently, the protonation state and conformation are mutually dependent and an *ab initio* method for conformation prediction will couple optimisation of conformation with the optimisation of the protonation state (Antosiewicz and Porschke, 1989; Mehler, 1996).

In addition to ionisation, protein solvation may also be accompanied by a significant degree of electronic polarisation. Polarisation of the solute has been deliberately neglected in our LD-Amber studies, but it can be included using one of the PE models for biomolecules that explicitly include electronic polarisability, such as those developed by Cieplak and co-workers (Cieplak et al., 2001; Wang et al., 2006).

8.3.4. Development of Simplified Protein Models

Atomistic protein models provide very high accuracy and the best insight into events on atomic levels. However, they are extremely computationally demanding. Being a method that relies on generation of random structures, evolutionary algorithm is bound to operate with many conformations that are characterised with high potential energies, especially in the early stages of the algorithm execution. In such a situation, an EA based method spends a significant amount of time doing detailed energy calculations for structures that will quickly be rejected. It may be advantageous to utilise other, simplified models of protein structure, such as united-residue model (Zhou et al., 2003) for primary, approximate evaluation of the potential energy associated with a structure. Structures that show high fitness (i.e. low potential energy) would then be subjected to a detailed atomistic PE calculation. Although some of the structures would have their potential energies calculated twice, overall, the number of expensive all-atom calculations would significantly reduce.

Implementation of united-residue or other bead representations is not straightforward as physical parameters in the Langevin dipole model are based on all-atom representation of the solute (Florián and Warshel, 1997). It may, therefore, be necessary to reparameterise the LD model in accordance with the simplified protein representation, or adopt a new solvation model.

8.3.5. Development of an Evolutionary Algorithm Approach for Prediction of Amino Acid Sequences with Optimal Adsorbing Properties

The EA approach discussed so far determines the conformation and associated adsorption energy for a protein with the known primary structure or amino acid sequence. The method may, however, be embedded into a more complex evolutionary algorithm that will be utilised to find the optimal sequence of amino

acids for adsorption on a given surface. In such a case, amino acid sequence would be generated by an "outer level EA", whilst the adsorption energies of each generated peptide would be determined by an "inner EA" described in this work. The fitness function for an outer EA would be adsorption energy of an optimal conformation produced by the inner algorithm. Alternative fitness functions could also be designed to satisfy other applications (e.g. finding a peptide that optimally binds to two different solid surfaces). A method for determining optimally binding peptide can, for example, find application in nanotechnology, where the peptide could be used to bind two solid nanoparticles (as indicated in the Introduction).

References

Abe H., Gō N. (1981). "Noninteracting local-structure model of folding and unfolding transition in globular proteins. II. Application to two-dimensional lattice proteins". *Biopolymers*, **20**, 1013-1031.

Albrecht A., Mundlos S. (2005). "The other trinucleotide repeat: polyalanine expansion disorders". *Curr. Opin. Gen. Dev.*, **15**, 285-293.

Allen M.P., Tildesley D.J. (1989). Computer Simulation of Liquids. Oxford University Press, Oxford.

Anderson A.G., Hermans J. (1988). "Microfolding: Conformational probability map for the alanine dipeptide in water from molecular dynamics simulations". *Proteins: Struct., Funct., Genet.*, **3**, 262-265.

Androulakis I.P., Maranas C.D., Floudas C.A. (1997). "Prediction of Oligopeptide Conformations via Deterministic Global Optimization". *J. Global Opt.*, **11**, 1-34.

Anfinsen C.B. (1973). "Principles that Govern the Folding of Protein Chains". *Science*, **181**, 223-230.

Antosiewicz J., Porschke D. (1989). "The nature of protein dipole moments: experimental and calculated permanent dipole of α-chymotrypsin". *Biochemistry*, **28**, 10072-10078.

Apostolakis J., Ferrara P., Caflisch A. (1999). "Calculation of conformational transitions and barriers in solvated systems: Application to the alanine dipeptide in water". *J. Chem. Phys.*, **110**, 2099-2108.

Arnautova Y.A., Jagielska A., Scheraga H.A. (2006). "A New Force Field (ECEPP-05) for Peptides, Proteins, and Organic Molecules". *J. Phys. Chem. B*, **110**, 5025-5044.

Asthagiri D., Lenhoff A.M. (1997). "Influence of Structural Details in Modeling Electrostatically Driven Protein Adsorption". *Langmuir*, **13**, 6761-6768.

Avbelj F. (2000). "Amino acid conformational preferences and solvation of polar backbone atoms in peptides and proteins". *J. Mol. Biol.*, **300**, 1335-1359.

Bäck T., Hoffmeister F. (1991). "Extended selection mechanisms in genetic algorithms", in *Proceedings of the Fourth International Conference on Genetic Algorithms*, R.K. Belew, L.B. Booker, eds. Morgan Kaufmann Publishers Inc., San Mateo, CA, 92-99.

Baker D., Sali A. (2001). "Protein Structure Prediction and Structural Genomics". *Science*, **294**, 93-96.

Bandosz T.J., Biggs M.J., Gubbins K.E., Hattori Y., Iiyama T., Kaneko K., Pikunic J., Thomson K.T. (2003). "Molecular Models of Porous Carbons", in *Chemistry and Physics of Carbon*, L.R. Radovic, ed Marcel Dekker, Inc., New York, 41-228.

Barron L.D., Hecht L., Wilson G. (1997). "The Lubricant of Life: A Proposal That Solvent Water Promotes Extremely Fast Conformational Fluctuations in Mobile Heteropolypeptide Structure". *Biochemistry*, **36**, 13143-13147.

Bartels C., Karplus M. (1998). "Probability Distributions for Complex Systems: Adaptive Umbrella Sampling of the Potential Energy". *J. Phys. Chem. B*, **102**, 865-880.

Beglov D., Roux B. (1995). "Dominant solvation effects from the primary shell of hydration: Approximation for molecular dynamics simulations". *Biopolymers*, **35**, 171-178.

Bell R.P. (1931). "The electrostatic energy of dipole molecules in different media". *Trans. Farad. Soc.*, **27**, 797-802.

Ben-Naim A. (1978). "Standard thermodynamics of transfer. Uses and misuses". *J. Phys. Chem.*, **82**, 792-803.

Ben-Naim A. (1990). "Solvent effects on protein association and protein folding". *Biopolymers*, **29**, 567-596.

Berendsen H.J.C., Postma J.P.M., van Gunsteren W.F., Hermans J. (1981). "Interaction models for water in relation to protein hydration", in *Intermolecular Forces*, B. Pullman, ed Reidel, Dordrecht, 331-342.

Berendsen H.J.C., Postma J.P.M., van Gunsteren W.F., DiNola A., Haak J.R. (1984). "Molecular dynamics with coupling to an external bath". *J. Chem. Phys.*, **81**, 3684-3690.

Berg B.A., Neuhaus T. (1992). "Multicanonical ensemble: A new approach to simulate first-order phase transitions". *Phys. Rev. Lett.*, **68**, 9-12.

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. (2000). "The Protein Data Bank". *Nucl. Acids Res.*, **28**, 235-242.

Bernal J.D., Fowler R.H. (1933). "A Theory of Water and Ionic Solution, with Particular Reference to Hydrogen and Hydroxyl Ions". *J. Chem. Phys.*, **1**, 515-548.

Biggs M.J., Buts A., Williamson D. (2004). "Absolute Assessment of Adsorption-Based Porous Solid Characterization Methods: Comparison Methods". *Langmuir*, **20**, 7123-7138.

Birge R.R. (1992). "Protein-Based Optical Computing and Memories". *Computer*, **25**, 56-67.

Bizzarri A.R. (2006). "Topological and dynamical properties of Azurin anchored to a gold substrate as investigated by molecular dynamics simulation". *Biophys. Chem.*, **122**, 206-214.

Bizzarri A.R., Costantini G., Cannistraro S. (2003). "MD simulation of a plastocyanin mutant adsorbed onto a gold surface". *Biophys. Chem.*, **106**, 111-123.

Blaney J.M., Weiner P.K., Dearing A., Kollman P.A., Jorgensen E.C., Oatley S.J., Burridge J.M., Blake C.C.F. (1982). "Molecular mechanics simulation of protein-ligand interactions: binding of thyroid hormone analogs to prealbumin". *J. Am. Chem. Soc.*, **104**, 6424-6434.

Bleich H.E., Cutnell J.D., Day A.R., Freer R.J., Glasel J.A., McKelvy J.F. (1976). "Preliminary Analysis of ^{1}H and ^{13}C Spectral and Relaxation Behavior in Methionine-Enkephalin". *Proc. Natl. Acad. Sci. U. S. A.*, **73**, 2589-2593.

Bogen H.J. (1968). Modern Biology. Weidenfeld & Nicolson, London.

Bojan M.J., Steele W.A. (1987). "Virial coefficients for N_2 and CO adsorbed on the graphite basal plane". *Langmuir*, **3**, 116-120.
Bolhuis P.G. (2003). "Transition-path sampling of β-hairpin folding". *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 12129-12134.
Born M. (1920a). "Volumen und Hydratationswärme der Ionen". *Z. Phys. A: Hadrons Nucl.*, **1**, 45-48.
Born M. (1920b). "Volumen und Hydratationswärme der Ionen". *Zeitschrift für Physik*, **1**, 45-48.
Bragg L., Kendrew J.C., Perutz M.F. (1950). "Polypeptide Chain Configurations in Crystalline Proteins". *Proc. R. Soc. London, A*, **203**, 321-357.
Braun R., Sarikaya M., Schulten K. (2002). "Genetically Engineered Gold-Binding Polypeptides: Structure Prediction and Molecular Dynamics". *Journal of Biomaterials Science - Polymer Edition*, **13**, 747-757.
Brünger A.T., Kuriyan J., Karplus M. (1987). "Crystallographic R Factor Refinement by Molecular Dynamics". *Science*, **235**, 458-460.
Buesnel R., Hillier I.H., Masters A.J. (1997). "A molecular dynamics study of the conformation of the alanine dipeptide in aqueous solution using a quantum mechanical potential". *Mol. Phys.*, **90**, 787-792.
Bujnowski A.M., Pitt W.G. (1998). "Water Structure around Enkephalin near a PE Surface: A Molecular Dynamics Study". *J. Colloid Interface Sci.*, **203**, 47-58.
Bursulaya B.D., Brooks C.L. (1999). "Folding Free Energy Surface of a Three-Stranded β-Sheet Protein". *J. Am. Chem. Soc.*, **121**, 9947-9951.
Cantor C.R., Schimmel P.R. (1980). Part I: The Conformation of Biological Macromolecules. W. H. Freeman and Company, San Francisco.
Carlsson F., Hyltner E., Arnebrant T., Malmsten M., Linse P. (2004). "Lysozyme Adsorption to Charged Surfaces. A Monte Carlo Study". *J. Phys. Chem. B*, **108**, 9871-9881.
Carravetta V., Monti S. (2006). "Peptide-TiO_2 Surface Interaction in Solution by Ab Initio and Molecular Dynamics Simulations". *J. Phys. Chem. B*, **110**, 6160-6169.
Castells V., Yang S., Van Tassel P.R. (2002). "Surface-induced conformational changes in lattice model proteins by Monte Carlo simulation". *Phys. Rev. E*, **65**, 031912.
Castillo J., Gáspár S., Leth S., Niculescu M., Mortari A., Bontidean I., Soukharev V., Dorneanu S.A., Ryabov A.D., Csöregi E. (2004). "Biosensors for life quality: Design, development and applications". *Sens. Actuators, B*, **102**, 179-194.
Chan H.S., Dill K.A. (1994). "Transition states and folding dynamics of proteins and heteropolymers". *J. Chem. Phys.*, **100**, 9238-9257.
Chandler D., Andersen H.C. (1972). "Optimized Cluster Expansions for Classical Fluids. II. Theory of Molecular Liquids". *J. Chem. Phys.*, **57**, 1930-1937.
Cheng X., Cui G., Hornak V., Simmerling C. (2005). "Modified Replica Exchange Simulation Methods for Local Structure Refinement". *J. Phys. Chem. B*, **109**, 8220-8230.
Chipot C., Pohorille A. (1998). "Conformational Equilibria of Terminally Blocked Single Amino Acids at the Water-Hexane Interface. A Molecular Dynamics Study". *J. Phys. Chem. B*, **102**, 281-290.
Chothia C. (1974). "Hydrophobic bonding and accessible surface area in proteins". *Nature*, **248**, 338-339.

Cieplak P., Caldwell J., Kollman P. (2001). "Molecular mechanical models for organic and biological systems going beyond the atom centered two body additive approximation: aqueous solution free energies of methanol and N-methyl acetamide, nucleic acid base, and amide hydrogen bonding and chloroform/water partition coefficients of the nucleic acid bases". *J. Comput. Chem.*, **22**, 1048-1057.

Clark J.D., Hodgkin E.E., Marshall G.R. (1991). "Helical Transitions in Peptides", in *Molecular Conformation and Biological Interactions*, P. Balaram, S. Ramaseshan, eds. Indian Academy of Sciences, Bangalore, 503-510.

Clement-Jones V., Corder R., Smith R., Medbak S., Lowry P.J., Rees L.H., Besser G.M. (1982). "Met-enkephalin and related peptides in Man", in *Regulatory Peptides: From Molecular Biology to Function*, E. Costa, M. Trabucchi, eds. Raven Press, New York, 379-386.

Clore G.M., Gronenborn A.M. (1987). "Determination of three-dimensional structures of proteins in solution by nuclear magnetic resonance spectroscopy". *Protein Eng.*, **1**, 275-288.

Cole C., Warwicker J. (2002). "Side-chain conformational entropy at protein-protein interfaces". *Protein Sci.*, **11**, 2860-2870.

Connolly M.L. (1983). "Solvent-accessible surfaces of proteins and nucleic acids". *Science*, **221**, 709-713.

Cormack A.N., Lewis R.J., Goldstein A.H. (2004). "Computer Simulation of Protein Adsorption to a Material Surface in Aqueous Solution: Biomaterials Modeling of a Ternary System". *J. Phys. Chem. B*, **108**, 20408-20418.

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

Cornette J.L., Cease K.B., Margalit H., Spouge J.L., Berzofsky J.A., DeLisi C. (1987). "Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins". *J. Mol. Biol.*, **195**, 659-685.

Cracknell R.F., Gubbins K.E., Maddox M., Nicholson D. (1995). "Modeling Fluid Behavior in Well-Characterized Porous Materials". *Acc. Chem. Res.*, **28**, 281-288.

D'Amelio N., Gaggelli E., Gaggelli N., Mancini F., Molteni E., Valensin D., Valensin G. (2003). "The structure of the Ce(III)-Angiotensin II complex as obtained from NMR data and molecular dynamics calculations". *J. Inorg. Biochem.*, **95**, 225-229.

Darden T., York D., Pedersen L. (1993). "Particle mesh Ewald: An $N\cdot\log(N)$ method for Ewald sums in large systems". *J. Chem. Phys.*, **98**, 10089-10092.

Dauber-Osguthorpe P., Roberts V.A., Osguthorpe D.J., Wolff J., Genest M., Hagler A.T. (1988). "Structure and energetics of ligand binding to proteins: *Escherichia coli* dihydrofolate reductase-trimethoprim, a drug-receptor system". *Proteins: Struct., Funct., Genet.*, **4**, 31-47.

Davidon W.C. (1975). "Optimally conditioned optimization algorithms without line searches". *Mathematical Programming*, **9**, 1-30.

De Jong K.A. (1975) An Analysis of the Behavior of a Class of Genetic Adaptive Systems. In: Logic of Computers Group. Ann Arbor, MI: University of Michigan.

Djurdjević D. (2006) *Ab initio* Protein Fold Prediction Using Evolutionary Algorithms. In: School of Engineering and Electronics. Edinburgh: The University of Edinburgh.

Djurdjevic D.P., Biggs M.J. (2006). "*Ab initio* protein fold prediction using evolutionary algorithms: Influence of design and control parameters on performance". *J. Comput. Chem.*, **27**, 1177-1195.

Djurdjevic D.P., Biggs M.J., Mijajlovic M. In.

Doucet N., Pelletier J.N. (2007). "Simulated annealing exploration of an active-site tyrosine in TEM-1β-lactamase suggests the existence of alternate conformations". *Proteins: Struct., Funct., Bioinf.*, **69**, 340-348.

Drozdov A.N., Grossfield A., Pappu R.V. (2004). "Role of Solvent in Determining Conformational Preferences of Alanine Dipeptide in Water". *J. Am. Chem. Soc.*, **126**, 2574-2581.

Edsall J.T., McKenzie H.A. (1983). "Water and proteins. II. The location and dynamics of water in protein systems and its relation to their stability and properties". *Advances in Biophysics*, **16**, 53-183.

Esteve V., Blondelle S., Celda B., Pérez-Payá E. (2001). "Stabilization of an α-helical conformation in an isolated hexapeptide inhibitor of calmodulin". *Biopolymers*, **59**, 467-476.

Feringa B.L. (2001). "In Control of Motion: From Molecular Switches to Molecular Motors". *Acc. Chem. Res.*, **34**, 504-513.

Flemming H.-C. (2002). "Biofouling in water systems – cases, causes and countermeasures". *Appl. Microbiol. Biotechnol.*, **59**, 629-640.

Florián J., Warshel A. (1997). "Langevin Dipoles Model for ab Initio Calculations of Chemical Processes in Solution: Parametrization and Application to Hydration Free Energies of Neutral and Ionic Solutes and Conformational Analysis in Aqueous Solution". *J. Phys. Chem. B*, **101**, 5583-5595.

Forrest L.R., Sansom M.S.P. (2000). "Membrane simulations: bigger and better?" *Curr. Opin. Struct. Biol.*, **10**, 174-181.

Frecer V., Ho B., Ding J.L. (2004). "De Novo Design of Potent Antimicrobial Peptides". *Antimicrob. Agents Chemother.*, **48**, 3349-3357.

Freedman H., Truong T.N. (2004). "An application of coupled reference interaction site model/molecular dynamics to the conformational analysis of the alanine dipeptide". *J. Chem. Phys.*, **121**, 12447-12456.

Frenkel D., Smit B. (1996). Understanding Molecular Simulation. Academic Press, San Diego.

Friedel M., Baumketner A., Shea J.-E. (2006). "Effects of surface tethering on protein folding mechanisms". *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 8396-8401.

Frimand K., Bohr H., Jalkanen K.J., Suhai S. (2000). "Structures, vibrational absorption and vibrational circular dichroism spectra of L-alanine in aqueous solution: a density functional theory and RHF study". *Chem. Phys.*, **255**, 165-194.

Giacomelli C.E., Bremer M.G.E.G., Norde W. (1999). "ATR-FTIR Study of IgG Adsorbed on Different Silica Surfaces". *J. Colloid Interface Sci.*, **220**, 13-23.

Gnanakaran S., Hochstrasser R.M. (2001). "Conformational Preferences and Vibrational Frequency Distributions of Short Peptides in Relation to Multidimensional Infrared Spectroscopy". *J. Am. Chem. Soc.*, **123**, 12886-12898.

Gō N., Abe H. (1981). "Noninteracting local-structure model of folding and unfolding transition in globular proteins. I. Formulation". *Biopolymers*, **20**, 991-1011.
Goldberg D.E. (1989). Genetic Algorithms in Search, Optimization, and Machine Learning. Addison Wesley Longman, Reading, MA.
Goldberg D.E., Korb B., Deb K. (1989). "Messy Genetic Algorithms: Motivation, Analysis, and First Results". *Complex Systems*, **3**, 493-530.
Goodsell D.S. (1996). Our Molecular Nature: The Body's Motors. Machines and Messages. Copernicus Springer-Verlag New York, Inc., New York.
Gorba C., Geyer T., Helms V. (2004). "Brownian dynamics simulations of simplified cytochrome c molecules in the presence of a charged surface". *J. Chem. Phys.*, **121**, 457-464.
Gordon H.L., Kwan W.K., Gong C., Larrass S., Rothstein S.M. (2003). "Efficient generation of low-energy folded states of a model protein". *J. Chem. Phys.*, **118**, 1533-1540.
Gould I.R., Cornell W.D., Hillier I.H. (1994). "A quantum Mechanical Investigation of the Conformational Energetics of the Alanine and Glycine Dipeptides in the Gas Phase and in Aqueous Solution". *J. Am. Chem. Soc.*, **116**, 9250-9256.
Graham W.H., Carter E.S.I., Hicks R.P. (1992). "Conformational analysis of met-enkephalin in both aqueous solution and in the presence of sodium dodecyl sulfate micelles using multidimensional NMR and molecular modeling". *Biopolymers*, **32**, 1755-1764.
Griffin M.A., Friedel M., Shea J.-E. (2005). "Effects of frustration, confinement, and surface interactions on the dimerization of an off-lattice β-barrel protein". *J. Chem. Phys.*, **123**, 174707.
Guru B.S., Hiziroğlu H.R. (2004). Electromagnetic Field Theory Fundamentals. 2nd Edition, Cambridge University Press, Cambridge.
Han W.-G., Jalkanen K.J., Elstner M., Suhai S. (1998). "Theoretical Study of Aqueous *N*-Acetyl-L-alanine *N*′-Methylamide: Structures and Raman, VCD, and ROA Spectra". *J. Phys. Chem. B*, **102**, 2587-2602.
Hancock P.J.B. (1994). "An empirical comparison of selection methods in evolutionary algorithms", in *Evolutionary Computing*, T.C. Fogarty, ed Springer, Berlin, 80-94.
Hansmann U.H.E. (1997). "Parallel tempering algorithm for conformational studies of biological molecules". *Chem. Phys. Lett.*, **281**, 140-150.
Hansmann U.H.E., Okamoto Y. (1999). "New Monte Carlo algorithms for protein folding". *Curr. Opin. Struct. Biol.*, **9**, 177-183.
Hansmann U.H.E., Okamoto Y., Eisenmenger F. (1996). "Molecular dynamics, Langevin and hydrid Monte Carlo simulations in a multicanonical ensemble". *Chem. Phys. Lett.*, **259**, 321-330.
Härtl A., Schmich E., Garrido J.A., Hernando J., Catharino S.C.R., Walter S., Feulner P., Kromka A., Steinmüller D., Stutzmann M. (2004). "Protein-modified nanocrystalline diamond thin films for biosensor applications". *Nature Materials*, **3**, 736-742.
Haupt R.L., Haupt S.E. (1998). Practical Genetic Algorithms. John Wiley & Sons, New York.

Head-Gordon T., Head-Gordon M., Frisch M.J., Brooks C.L., Pople J.A. (1991). "Theoretical study of blocked glycine and alanine peptide analogs". *J. Am. Chem. Soc.*, **113**, 5989-5997.

Hénin J., Pohorille A., Chipot C. (2005). "Insights into the Recognition and Association of Transmembrane α-Helices. The Free Energy of α-Helix Dimerization in Glycophorin A". *J. Am. Chem. Soc.*, **127**, 8478-8484.

Hermans J., Berendsen H.J.C., van Gunsteren W.F., Postma J.P.M. (1984). "A consistent empirical potential for water-protein interactions". *Biopolymers*, **23**, 1513-1518.

Hille B., Catterall W.A. (2006). "Electrical Excitability and Ion Channels", in *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*, 7th Edition, G.J. Siegel, R.W. Albers, S.T. Brady, D.L. Price, eds. Elsevier, Amsterdam,

Hoang T.X., Cieplak M. (2000). "Molecular dynamics of folding of secondary structures in Go-type models of proteins". *J. Chem. Phys.*, **112**, 6851-6862.

Holland J.H. (1975). Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI.

Honig B., Nicholls A. (1995). "Classical electrostatics in biology and chemistry". *Science*, **268**, 1144-1149.

Hu H., Elstner M., Hermans J. (2003). "Comparison of a QM/MM force field and molecular mechanics force fields in simulations of alanine and glycine "dipeptides" (Ace-Ala-Nme and Ace-Gly-Nme) in water in relation to the problem of modeling the unfolded peptide backbone in solution". *Proteins: Struct., Funct., Genet.*, **50**, 451-463.

Huggins M.L. (1943). "The Structure of Fibrous Proteins". *Chem. Rev.*, **32**, 195-218.

Hughes J., Smith T.W., Kosterlitz H.W., Fothergill L.A., Morgan B.A., Morris H.R. (1975). "Identification of two related pentapeptides from the brain with potent opiate agonist activity". *Nature*, **258**, 577-579.

Hultschig C., Kreutzberger J., Seitz H., Konthur Z., Büssow K., Lehrach H. (2006). "Recent advances of protein microarrays". *Curr. Opin. Chem. Biol.*, **10**, 4-10.

Huston S.E., Marshall G.R. (1994). "$\alpha/3_{10}$-Helix transitions in α-methylalanine homopeptides: Conformational transition pathway and potential of mean force". *Biopolymers*, **34**, 75-90.

Imamura K., Kawasaki Y., Awadzu T., Sakiyama T., Nakanishi K. (2003). "Contribution of acidic amino residues to the adsorption of peptides onto a stainless steel surface". *J. Colloid Interface Sci.*, **267**, 294-301.

Isogai Y., Nemethy G., Scheraga H.A. (1977). "Enkephalin: Conformational Analysis by means of Empirical Energy Calculations". *Proc. Natl. Acad. Sci. U. S. A.*, **74**, 414-418.

Jin A.Y., Leung F.Y., Weaver D.F. (1999). "Three variations of genetic algorithm for searching biomolecular conformation space: Comparison of GAP 1.0, 2.0, and 3.0". *J. Comput. Chem.*, **20**, 1329-1342.

Jones C.R., Gibbons W.A., Garsky V. (1976). "Proton magnetic resonance studies of conformation and flexibility of enkephalin peptides". *Nature*, **262**, 779-782.

Jones C.R., Garsky V., Gibbons W.A. (1977). "Molecular conformations of met-enkephalin: Comparison of the zwitterionic and cationic forms". *Biochem. Biophys. Res. Commun.*, **76**, 619-625.

Jorgensen W.L. (1981). “Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water”. *J. Am. Chem. Soc.*, **103**, 335-340.

Jorgensen W.L. (1982). “Revised TIPS for simulations of liquid water and aqueous solutions”. *J. Chem. Phys.*, **77**, 4156-4163.

Jorgensen W.L., Maxwell D.S., Tirado-Rives J. (1996). “Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids”. *J. Am. Chem. Soc.*, **118**, 11225-11236.

Jorgensen W.L., Chandrasekhar J., Madura J.D., Impey R.W., Klein M.L. (1983). “Comparison of simple potential functions for simulating liquid water”. *J. Chem. Phys.*, **79**, 926-935.

Juffer A.H., Argos P., de Vlieg J. (1996). “Adsorption of proteins onto charged surfaces: A Monte Carlo approach with explicit ions”. *J. Comput. Chem.*, **17**, 1783-1803.

Kabsch W., Sander C. (1983). “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features”. *Biopolymers*, **22**, 2577-2637.

Kantarci N., Tamerler C., Sarikaya M., Haliloglu T., Doruker P. (2005). “Molecular dynamics simulations on constraint metal binding peptides”. *Polymer*, **46**, 4307-4313.

Kaptein R., Boelens R., Scheek R.M., Van Gunsteren W.F. (1988). “Protein structures from NMR”. *Biochemistry*, **27**, 5389-5395.

Karplus M., Kushick J.N. (1981). “Method for estimating the configurational entropy of macromolecules”. *Macromolecules*, **14**, 325-332.

Karplus M., Kuriyan J. (2005). “Chemical Theory and Computation Special Feature: Molecular dynamics and protein function”. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 6679-6685.

Karplus M., Ichiye T., Pettitt B.M. (1987). “Configurational entropy of native proteins”. *Biophys. J.*, **52**, 1083-1085.

Kasemo B. (2002). “Biological surface science”. *Surf. Sci.*, **500**, 656-677.

Katz E., Willner I. (2004). “Biomolecule-Functionalized Carbon Nanotubes: Applications in Nanobioelectronics”. *ChemPhysChem*, **5**, 1084-1104.

Kawai H., Kikuchi T., Okamoto Y. (1989). “A prediction of tertiary structures of peptide by the Monte Carlo simulated annealing method”. *Protein Eng.*, **3**, 85-94.

Kelly T.R. (2001). “Progress toward a Rationally Designed Molecular Motor”. *Acc. Chem. Res.*, **34**, 514-522.

Khaled M.A., Long M.M., Thompson W.D., Bradley R.J., Brown G.B., Urry D.W. (1977). “Conformational states of enkephalins in solution”. *Biochem. Biophys. Res. Commun.*, **76**, 224-231.

Kim Y.S., Wang J., Hochstrasser R.M. (2005). “Two-Dimensional Infrared Spectroscopy of the Alanine Dipeptide in Aqueous Solution”. *J. Phys. Chem. B*, **109**, 7511-7521.

Kinoshita M., Okamoto Y., Hirata F. (1997). “Solvation structure and stability of peptides in aqueous solutions analyzed by the reference interaction site model theory”. *J. Chem. Phys.*, **107**, 1586-1599.

Kinoshita M., Okamoto Y., Hirata F. (1998). "First-Principle Determination of Peptide Conformations in Solvents: Combination of Monte Carlo Simulated Annealing and RISM Theory". *J. Am. Chem. Soc.*, **120**, 1855-1863.

Kirkpatrick S., Gelatt C.D., Jr., Vecchi M.P. (1983). "Optimization by Simulated Annealing". *Science*, **220**, 671-680.

Kitano M., Kuchitsu K. (1973). "Molecular Structure of Acetamide as Studied by Gas Electron Diffraction". *Bull. Chem. Soc. Jpn.*, **46**, 3048-3051.

Kitano M., Fukuyama T., Kuchitsu K. (1973). "Molecular Structure of N-Methylacetamide as Studied by Gas Electron Diffraction". *Bull. Chem. Soc. Jpn.*, **46**, 384-387.

Klepeis J.L., Floudas C.A. (1999). "Free energy calculations for peptides via deterministic global optimization". *J. Chem. Phys.*, **110**, 7491-7512.

Klepeis J.L., Schafroth H.D., Westerberg K.M., Floudas C.A. (2002). "Deterministic global optimization and *ab initio* approaches for the structure prediction of polypeptides, dynamics of protein folding, and protein-protein interactions", in *Computational Methods for Protein Folding: A Special Volume of Advances in Chemical Physics*, R.A. Friesner, ed John Wiley & Sons, 265-457.

Knotts IV T.A., Rathore N., de Pablo J.J. (2005). "Structure and stability of a model three-helix-bundle protein on tailored surfaces". *Proteins: Struct., Funct., Bioinf.*, **61**, 385-397.

Koča J., Carlsen P.H.J. (1995). "Conformational behavior and flexibility of met-enkephalin". *J. Mol. Struct. THEOCHEM*, **337**, 17-24.

Kolinski A., Skolnick J., Yaris R. (1986). "Monte Carlo Simulations on an Equilibrium Globular Protein Folding Model". *Proc. Natl. Acad. Sci. U. S. A.*, **83**, 7267-7271.

Kondo A., Urabe T., Yoshinaga K. (1996). "Adsorption activity and conformation of α-amylase on various ultrafine silica particles modified with polymer silane coupling agents". *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, **109**, 129-136.

Krigbaum W.R., Lin S.F. (1982). "Monte Carlo simulation of protein folding using a lattice model". *Macromolecules*, **15**, 1135-1145.

Lau K.F., Dill K.A. (1990). "Theory for Protein Mutability and Biogenesis". *Proc. Natl. Acad. Sci. U. S. A.*, **87**, 638-642.

Le Grand S.M., Merz K.M. (1993). "The application of the genetic algorithm to the minimization of potential energy functions". *J. Global Opt.*, **3**, 49-66.

Le Grand S.M., Merz K.M. (1994). "The Genetic Algorithm and the Conformational Search of Polypeptides and Proteins". *Mol. Sim.*, **13**, 299 - 320.

Lee J. (1993). "New Monte Carlo algorithm: Entropic sampling". *Phys. Rev. Lett.*, **71**, 211-214.

Lee J., Scheraga H.A., Rackovsky S. (1997). "New optimization method for conformational energy calculations on polypeptides: Conformational space annealing". *J. Comput. Chem.*, **18**, 1222-1232.

Lee M.S., Olson M.A. (2005). "Evaluation of Poisson Solvation Models Using a Hybrid Explicit/Implicit Solvent Method". *J. Phys. Chem. B*, **109**, 5223-5236.

Leo A., Hansch C., Elkins D. (1971). "Partition coefficients and their uses". *Chem. Rev.*, **71**, 525-616.

Li Z., Scheraga H.A. (1987). "Monte Carlo-Minimization Approach to the Multiple-Minima Problem in Protein Folding". *Proc. Natl. Acad. Sci. U. S. A.*, **84**, 6611-6615.

Lin J.H., Baumgaertner A. (2000). "Molecular dynamics simulations of hydrophobic and amphiphatic proteins interacting with a lipid bilayer membrane". *Computational and Theoretical Polymer Science*, **10**, 97-102.

Liou Y.-C., Tocilj A., Davies P.L., Jia Z. (2000). "Mimicry of ice structure by surface hydroxyls and water of a β-helix antifreeze protein". *Nature*, **406**, 322-324.

Liu S.M., Haynes C.A. (2004). "Mesoscopic analysis of conformational and entropic contributions to nonspecific adsorption of HP copolymer chains using dynamic Monte Carlo simulations". *J. Colloid Interface Sci.*, **275**, 458-469.

Liu Z., Mao F., Li W., Han Y., Lai L. (2000). "Calculation of Protein Surface Loops Using Monte-Carlo Simulated Annealing Simulation". *Journal of Molecular Modeling*, **6**, 1-8.

Losonczi J.A., Olejniczak E.T., Betz S.F., Harlan J.E., Mack J., Fesik S.W. (2000). "NMR Studies of the Anti-Apoptotic Protein Bcl-xL in Micelles". *Biochemistry*, **39**, 11024-11033.

Lu D.R. (1993). "Glucagon adsorption on polymer surfaces with -helical and extended -strand conformations: A computational approach". *Journal of Biomaterials Science, Polymer Edition*, **4**, 323-335.

Lu D.R., Park K. (1989). "Protein adsorption on polymer surfaces: calculation of adsorption energies". *Journal of Biomaterials Science, Polymer Edition*, **1**, 243-260.

Lu D.R., Lee S.J., Park K. (1992). "Calculation of solvation interaction energies for protein adsorption on polymer surfaces". *Journal of Biomaterials Science, Polymer Edition*, **3**, 127-147.

MacKerell A.D., Bashford D., Bellott M., Dunbrack R.L., Evanseck J.D., Field M.J., Fischer S., Gao J., Guo H., Ha S., Joseph-McCarthy D., Kuchnir L., Kuczera K., Lau F.T.K., Mattos C., Michnick S., Ngo T., Nguyen D.T., Prodhom B., Reiher W.E., Roux B., Schlenkrich M., Smith J.C., Stote R., Straub J., Watanabe M., Wiorkiewicz-Kuczera J., Yin D., Karplus M. (1998). "All-Atom Empirical Potential for Molecular Modeling and Dynamics Studies of Proteins". *J. Phys. Chem. B*, **102**, 3586-3616.

Mackerell A.D., Jr. (2004). "Empirical force fields for biological macromolecules: Overview and issues". *J. Comput. Chem.*, **25**, 1584-1604.

Madison V., Kopple K.D. (1980). "Solvent-dependent conformational distributions of some dipeptides". *J. Am. Chem. Soc.*, **102**, 4855-4863.

Mahaffy R., Bhatia R., Garrison B.J. (1997). "Diffusion of a Butanethiolate Molecule on a Au{111} Surface". *J. Phys. Chem. B*, **101**, 771-773.

Mahoney M.W., Jorgensen W.L. (2000). "A five-site model for liquid water and the reproduction of the density anomaly by rigid, nonpolarizable potential functions". *J. Chem. Phys.*, **112**, 8910-8922.

Mantero S., Piuri D., Montevecchi F.M., Vesentini S., Ganazzoli F., Raffaini G. (2002). "Albumin adsorption onto pyrolytic carbon: A molecular mechanics approach". *J. Biomed. Mat. Res.*, **59**, 329-339.

Marcotte I., Separovic F., Auger M., Gagne S.M. (2004). "A Multidimensional 1H NMR Investigation of the Conformation of Methionine-Enkephalin in Fast-Tumbling Bicelles". *Biophys. J.*, **86**, 1587-1600.

Martin J., Letellier G., Marin A., Taly J.-F., de Brevern A.G., Gibrat J.-F. (2005). "Protein secondary structure assignment revisited: a detailed analysis of different assignment methods". *BMC Struct. Biol.*, **5**, 17.

McCammon J.A., Gelin B.R., Karplus M. (1977). "Dynamics of folded proteins". *Nature*, **267**, 585-590.

McQuarrie D.A. (1976). Statistical Mechanics. Harper Collins, New York.

Mehler E.L. (1996). "Self-Consistent, Free Energy Based Approximation To Calculate pH Dependent Electrostatic Effects in Proteins". *J. Phys. Chem.*, **100**, 16006-16018.

Mehler E.L., Solmajer T. (1991). "Electrostatic effects in proteins: comparison of dielectric and charge models". *Protein Eng.*, **4**, 903-910.

Mehta M.A., Fry E.A., Eddy M.T., Dedeo M.T., Anagnost A.E., Long J.R. (2004). "Structure of the Alanine Dipeptide in Condensed Phases Determined by ^{13}C NMR". *J. Phys. Chem. B*, **108**, 2777-2780.

Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.H., Teller E. (1953). "Equation of State Calculations by Fast Computing Machines". *J. Chem. Phys.*, **21**, 1087-1092.

Mezei M., Mehrotra P.K., Beveridge D.L. (1985). "Monte Carlo determination of the free energy and internal energy of hydration for the Ala dipeptide at 25 ^{o}C". *J. Am. Chem. Soc.*, **107**, 2239-2245.

Mijajlovic M., Biggs M.J. (2007a). "Ab initio protein fold prediction using evolutionary algorithms: Influence of potential energy models and desired accuracy on performance characteristics."

Mijajlovic M., Biggs M.J. (2007b). "On Use of the Amber Potential with the Langevin Dipole Method". *J. Phys. Chem. B*, **111**, 7591-7602.

Mijajlovic M., Biggs M.J. (2007c). "Study of Conformational Switching in Polyalanine at Solid Surfaces Using Molecular Simulation". *J. Phys. Chem. C*, **111**, 15839-15847.

Mitchell M. (1996). An Introduction to Genetic Algorithms. Massachusetts Institute of Technology, Cambridge, MA.

Mitsutake A., Sugita Y., Okamoto Y. (2001). "Generalized-ensemble algorithms for molecular simulations of biopolymers". *Peptide Science*, **60**, 96-123.

Mitsutake A., Kinoshita M., Okamoto Y., Hirata F. (2000). "Multicanonical algorithm combined with the RISM theory for simulating peptides in aqueous solution". *Chem. Phys. Lett.*, **329**, 295-303.

Mizushima S.-i., Simanouti T., Nagakura S., Kuratani K., Tsuboi M., Baba H., Fujioka O. (1950). "The Molecular Structure of N-Methylacetamide". *J. Am. Chem. Soc.*, **72**, 3490-3494.

Momany F.A., McGuire R.F., Burgess A.W., Scheraga H.A. (1975). "Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids". *J. Phys. Chem.*, **79**, 2361-2381.

Morris G.M., Goodsell D.S., Halliday R.S., Huey R., Hart W.E., Belew R.K., Olson A.J. (1998). "Automated docking using a Lamarckian genetic algorithm and

an empirical binding free energy function". *J. Comput. Chem.*, **19**, 1639-1662.

Mungikar A.A., Forciniti D. (2004). "Conformational Changes of Peptides at Solid/Liquid Interfaces: A Monte Carlo Study". *Biomacromolecules*, **5**, 2147-2159.

Mungikar A.A., Forciniti D. (2006). "Effect of Cosolvents on the Adsorption of Peptides at the Solid-Liquid Interface". *Biomacromolecules*, **7**, 239-251.

Nachman J., Pai E.F., Pomès R. (2002). "Structure and Energy Calculations for Low-Affinity Peptide-Protein Complexes". *Biophys. J. (Annual Meeting Abstracts)*, **82**, 134a-134a.

Nakajima N., Nakamura H., Kidera A. (1997). "Multicanonical Ensemble Generated by Molecular Dynamics Simulation for Enhanced Conformational Sampling of Peptides". *J. Phys. Chem. B*, **101**, 817-824.

Narten A.H. (1972). "Liquid Water: Atom Pair Correlation Functions from Neutron and X-Ray Diffraction". *J. Chem. Phys.*, **56**, 5681-5687.

Narten A.H., Danford M.D., Levy H.A. (1967). "X-ray diffraction study of liquid water in the temperature range 4–200°C". *Disc. Farad. Soc.*, **43**, 97-107.

Nayeem A., Vila J., Scheraga H.A. (1991). "A comparative study of the simulated-annealing and Monte Carlo-with-minimization approaches to the minimum-energy structures of polypeptides: [Met]-enkephalin". *J. Comput. Chem.*, **12**, 594-605.

Nemethy G., Pottle M.S., Scheraga H.A. (1983). "Energy parameters in polypeptides. 9. Updating of geometrical parameters, nonbonded interactions, and hydrogen bond interactions for the naturally occurring amino acids". *J. Phys. Chem.*, **87**, 1883-1887.

Nemethy G., Gibson K.D., Palmer K.A., Yoon C.N., Paterlini G., Zagari A., Rumsey S., Scheraga H.A. (1992). "Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides". *J. Phys. Chem.*, **96**, 6472-6484.

Nguyen H.D., Hall C.K. (2006). "Spontaneous Fibril Formation by Polyalanines; Discontinuous Molecular Dynamics Simulations". *J. Am. Chem. Soc.*, **128**, 1890-1901.

Nicholson D. (1996). "Using computer simulation to study the properties of molecules in micropores". *J. Chem. Soc., Faraday Trans.*, **92**, 1-9.

Noinville S., Bruston F., El Amri C., Baron D., Nicolas P. (2003). "Conformation, Orientation, and Adsorption Kinetics of Dermaseptin B2 onto Synthetic Supports at Aqueous/Solid Interface". *Biophys. J.*, **85**, 1196-1206.

Noinville V., Vidal-Madjar C., Sébille B. (1995). "Modeling of Protein Adsorption on Polymer Surfaces. Computation of Adsorption Potential". *J. Phys. Chem.*, **99**, 1516-1522.

Olszewski K.A., Piela L., Scheraga H.A. (1992). "Mean field theory as a tool for intramolecular conformational optimization. 1. Tests on terminally-blocked alanine and met-enkephalin". *J. Phys. Chem.*, **96**, 4672-4676.

Oren E.E., Tamerler C., Sarikaya M. (2005). "Metal Recognition of Septapeptides via Polypod Molecular Architecture". *Nano Lett.*, **5**, 415-419.

Orozco M., Luque F.J. (2000). "Theoretical Methods for the Description of the Solvent Effect in Biomolecular Systems". *Chem. Rev.*, **100**, 4187-4226.

Otero-Cruz J.D., Báez-Pagán C.A., Caraballo-González I.M., Lasalde-Dominicci J.A. (2007). "Tryptophan-scanning Mutagenesis in the αM3 Transmembrane Domain of the Muscle-type Acetylcholine Receptor: A SPRING MODEL REVEALED". *J. Biol. Chem.*, **282**, 9162-9171.

Pal S.K., Peon J., Zewail A.H. (2002). "Biological water at the protein surface: Dynamical solvation probed directly with femtosecond resolution". *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 1763-1768.

Park C., Goddard W.A., III (2000). "Stabilization of α-Helices by Dipole-Dipole Interactions within α-Helices". *J. Phys. Chem. B*, **104**, 7784-7789.

Pauling L. (1940). The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry. 2nd Edition, Cornell University Press, Ithaka, NY.

Pervushin K.V., Arseniev A.S. (1992). "Three-dimensional structure of (1-36)bacterioopsin in methanol--chloroform mixture and SDS micelles determined by 2D ^{1}H-NMR spectroscopy". *FEBS Lett.*, **308**, 190-196.

Pettitt B.M., Rossky P.J. (1982). "Integral equation predictions of liquid state structure for waterlike intermolecular potentials". *J. Chem. Phys.*, **77**, 1451-1457.

Pettitt B.M., Karplus M. (1988). "Conformational free energy of hydration for the alanine dipeptide: thermodynamic analysis". *J. Phys. Chem.*, **92**, 3994-3997.

Podtelezhnikov A.A., Wild D.L. (2005). "Exhaustive Metropolis Monte Carlo sampling and analysis of polyalanine conformations adopted under the influence of hydrogen bonds". *Proteins: Struct., Funct., Bioinf.*, **61**, 94-104.

Ponder J.W. (2004) TINKER – Software Tools for Molecular Design (V4.2). In.

Ponder J.W., Case D.A. (2003). "Force Fields for Protein Simulations", in *Protein Simulations*, V. Daggett, ed Academic Press, 27-85.

Poon C.-D., Samulski E.T., Weise C.F., Weisshaar J.C. (2000). "Do Bridging Water Molecules Dictate the Structure of a Model Dipeptide in Aqueous Solution?" *J. Am. Chem. Soc.*, **122**, 5642-5643.

Press W.H., Teukolsky S.A., Vetterling W.T., Flannery B.P. (1992). Numerical Recipes in C: The Art of Scientific Computing. 2nd Edition, Cambridge University Press, Cambridge.

Przybycien T.M., Pujar N.S., Steele L.M. (2004). "Alternative bioseparation operations: life beyond packed-bed chromatography". *Curr. Opin. Biotechnol.*, **15**, 469-478.

Quine J.R. (1999). "Helix parameters and protein structure using quaternions". *J. Mol. Struct. THEOCHEM*, **460**, 53-66.

Raffaini G., Ganazzoli F. (2003). "Simulation Study of the Interaction of Some Albumin Subdomains with a Flat Graphite Surface". *Langmuir*, **19**, 3403-3412.

Raffaini G., Ganazzoli F. (2004a). "Molecular Dynamics Simulation of the Adsorption of a Fibronectin Module on a Graphite Surface". *Langmuir*, **20**, 3371-3378.

Raffaini G., Ganazzoli F. (2004b). "Surface Ordering of Proteins Adsorbed on Graphite". *J. Phys. Chem. B*, **108**, 13850-13854.

Raffaini G., Ganazzoli F. (2006). "Protein adsorption on the hydrophilic surface of a glassy polymer: a computer simulation study". *Phys. Chem. Chem. Phys.*, **8**, 2765-2772.

Ramachandran G.N., Ramakrishnan C., Sasisekharan V. (1963). "Stereochemistry Of Polypeptide Chain Configurations". *J. Mol. Biol.*, **7**, 95-99.

Rambidi N.G. (2003). "Lure of molecular electronics--from molecular switches to distributed molecular information processing media". *Microelectron. Eng.*, **69**, 485-500.

Rappé A.K., Casewit C.J. (1997). Molecular Mechanics Across Chemistry. University Science Books, Sausalito, Calif.

Rathore N., de Pablo J.J. (2002). "Monte Carlo simulation of proteins through a random walk in energy space". *J. Chem. Phys.*, **116**, 7225-7230.

Rathore N., Chopra M., de Pablo J.J. (2005). "Optimal allocation of replicas in parallel tempering simulations". *J. Chem. Phys.*, **122**, 024111.

Ratner B.D., Bryant S.J. (2004). "BIOMATERIALS: Where We Have Been and Where We are Going". *Annu. Rev. Biomed. Eng.*, **6**, 41-75.

Ravichandran S., Madura J.D., Talbot J. (2001). "A Brownian Dynamics Study of the Initial Stages of Hen Egg-White Lysozyme Adsorption at a Solid Interface". *J. Phys. Chem. B*, **105**, 3610-3613.

Resat H., Maye P.V., Mezei M. (1997). "The sensitivity of conformational free energies of the alanine dipeptide to atomic site charges". *Biopolymers*, **41**, 73-81.

Ripoll D.R., Scheraga H.A. (1988). "On the multiple-minima problem in the conformational analysis of polypeptides. II. An electrostatically driven Monte Carlo method - tests on poly(L-alanine)". *Biopolymers*, **27**, 1283-1303.

Ripoll D.R., Scheraga H.A. (1989). "The multiple-minima problem in the conformational analysis of polypeptides. III. An Electrostatically Driven Monte Carlo Method: Tests on enkephalin". *J. Protein Chem.*, **8**, 263-287.

Ripoll D.R., Scheraga H.A., Pottle M.S., Gibson K.D., Liwo A., Li Z. (1995) ECEPPAK. In.

Roques B.P., Garbay-Jaureguiberry C., Oberlin R., Anteunis M., Lala A.K. (1976). "Conformation of Met^5-enkephalin determined by high field PMR spectroscopy". *Nature*, **262**, 778-779.

Rossky P.J., Karplus M., Rahman A. (1979). "A model for the simulation of an aqueous dipeptide solution". *Biopolymers*, **18**, 825-854.

Rosso L., Abrams J.B., Tuckerman M.E. (2005). "Mapping the Backbone Dihedral Free-Energy Surfaces in Small Peptides in Solution Using Adiabatic Free-Energy Dynamics". *J. Phys. Chem. B*, **109**, 4162-4167.

Roush D.J., Gill D.S., Willson R.C. (1994). "Electrostatic potentials and electrostatic interaction energies of rat cytochrome b_5 and a simulated anion-exchange adsorbent surface". *Biophys. J.*, **66**, 1290-1300.

Sanbonmatsu K.Y., García A.E. (2002). "Structure of Met-enkephalin in explicit aqueous solution using replica exchange molecular dynamics". *Proteins: Struct., Funct., Genet.*, **46**, 225-234.

Sapsford K.E., Shubin Y.S., Delehanty J.B., Golden J.P., Taitt C.R., Shriver-Lake L.C., Ligler F.S. (2004). "Fluorescence-based array biosensors for detection of biohazards". *J. Appl. Microbiol.*, **96**, 47-58.

Sarikaya M., Tamerler C., Jen A.K.-Y., Schulten K., Baneyx F. (2003). "Molecular biomimetics: nanotechnology through biology". *Nature Materials*, **2**, 577-585.

Scharnagl C., Reif M., Friedrich J. (2005). "Stability of proteins: Temperature, pressure and the role of the solvent". *Biochimica et Biophysica Acta (BBA) - Proteins & Proteomics*, **1749**, 187-213.

Schmidt A.B., Fine R.M. (1994). "A CFF 91-based Continuum Solvation Model: Solvation Free Energies of Small Organic Molecules and Conformations of the Alanine Dipeptide in Solution". *Mol. Sim.*, **13**, 347-365.

Schwefel H.-P. (1981). Numerical Optimization of Computer Models. Wiley, Chichester, UK.

Seeman N.C., Belcher A.M. (2002). "Emulating biology: Building nanostructures from the bottom up". *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 6451-6455.

Shang J., Geva E. (2005). "A Computational Study of the Correlations between Structure and Dynamics in Free and Surface-Immobilized Single Polymer Chains". *J. Phys. Chem. B*, **109**, 16340-16349.

Sheng Y.-J., Tsao H.-K., Zhou J., Jiang S. (2002). "Orientation of a Y-shaped biomolecule adsorbed on a charged surface". *Phys. Rev. E*, **66**, 011911.

Shin H., Jo S., Mikos A.G. (2003). "Biomimetic materials for tissue engineering". *Biomaterials*, **24**, 4353-4364.

Shirts M.R., Pitera J.W., Swope W.C., Pande V.S. (2003). "Extremely precise free energy calculations of amino acid side chain analogs: Comparison of common molecular mechanics force fields for proteins". *J. Chem. Phys.*, **119**, 5740-5761.

Shulze-Kremer S. (1992). "Genetic algorithms for protein tertiary structure prediction", in *Parallel Problem Solving From Nature 2*, R. Männer, B. Manderick, eds. Elsevier Science Publishers B.V., Amsterdam, 391-400.

Siegel G.J., Albers R.W., Brady S.T., Price D.L. (2006). Basic Neurochemistry: Molecular, Cellular and Medical Aspects. 7th Edition, Elsevier, Amsterdam.

Sippl M.J., Nemethy G., Scheraga H.A. (1984). "Intermolecular potentials from crystal data. 6. Determination of empirical potentials for O-H.O = C hydrogen bonds from packing configurations". *J. Phys. Chem.*, **88**, 6231-6233.

Skepö M., Linse P., Arnebrant T. (2006). "Coarse-Grained Modeling of Proline Rich Protein 1 (PRP-1) in Bulk Solution and Adsorbed to a Negatively Charged Surface". *J. Phys. Chem. B*, **110**, 12141-12148.

Smart J.L., Marrone T.J., McCammon J.A. (1997). "Conformational sampling with Poisson-Boltzmann forces and a stochastic dynamics/Monte Carlo method: Application to alanine dipeptide". *J. Comput. Chem.*, **18**, 1750-1759.

Smith B.J. (1999a). "Solvation parameters for amino acids". *J. Comput. Chem.*, **20**, 428-442.

Smith P.E. (1999b). "The alanine dipeptide free energy surface in solution". *J. Chem. Phys.*, **111**, 5568-5579.

Smythe M.L., Huston S.E., Marshall G.R. (1993). "Free energy profile of a 3_{10}- to α-helical transition of an oligopeptide in various solvents". *J. Am. Chem. Soc.*, **115**, 11594-11595.

Solov'yov I.A., Yakubovich A.V., Solov'yov A.V., Greiner W. (2006). "*Ab initio* study of alanine polypeptide chain twisting". *Phys. Rev. E*, **73**, 021916.

Song D., Forciniti D. (2001). "Monte Carlo simulations of peptide adsorption on solid surfaces (Monte Carlo simulations of peptide adsorption)". *J. Chem. Phys.*, **115**, 8089-8100.

Spadaccini R., Temussi P.A. (2001). “Natural peptide analgesics: the role of solution conformation”. *Cell. Mol. Life Sci.*, **58**, 1572-1582.

Spirtes M.A., Schwartz R.W., Mattice W.L., Coy D.H. (1978). “Circular dichroism and absorption study of the structure of methionine-enkephalin in solution”. *Biochem. Biophys. Res. Commun.*, **81**, 602-609.

Steele W. (1993). “Molecular interactions for physical adsorption”. *Chem. Rev.*, **93**, 2355-2378.

Steele W.A. (1974). The Interaction of Gases with Solid Surfaces. Pergamon Press, Oxford.

Stern H.A., Rittner F., Berne B.J., Friesner R.A. (2001). “Combined fluctuating charge and polarizable dipole models: Application to a five-site water potential function”. *J. Chem. Phys.*, **115**, 2237-2251.

Sternberg M.J.E., Chickos J.S. (1994). “Protien side-chain conformational entropy derived from fusion data-comparison with other empirical scales”. *Protein Eng.*, **7**, 149-155.

Still W.C., Tempczyk A., Hawley R.C., Hendrickson T. (1990). “Semianalytical treatment of solvation for molecular mechanics and dynamics”. *J. Am. Chem. Soc.*, **112**, 6127-6129.

Stillinger F.H., Rahman A. (1974). “Improved simulation of liquid water by molecular dynamics”. *J. Chem. Phys.*, **60**, 1545-1557.

Sugita Y., Okamoto Y. (1999). “Replica-exchange molecular dynamics method for protein folding”. *Chem. Phys. Lett.*, **314**, 141-151.

Sun Y., Welsh W.J., Latour R.A. (2005). “Prediction of the Orientations of Adsorbed Protein Using an Empirical Energy Function with Implicit Solvation”. *Langmuir*, **21**, 5616-5626.

Syswerda G. (1991). “A Study of Reproduction in Generational and Steady State Genetic Algorithms”, in *Foundations of Genetic Algorithms*, G.J.E. Rawlins, ed Morgan Kaufmann Publishers, San Mateo, CA, 94-101.

Takekiyo T., Imai T., Kato M., Taniguchi Y. (2004). “Temperature and pressure effects on conformational equilibria of alanine dipeptide in aqueous solution”. *Biopolymers*, **73**, 283-290.

Tirado-Rives J., Maxwell D.S., Jorgensen W.L. (1993). “Molecular dynamics and Monte Carlo simulations favor the α-helical form for alanine-based peptides in water”. *J. Am. Chem. Soc.*, **115**, 11590-11593.

Tobias D.J., Brooks C.L. (1992). “Conformational equilibrium in the alanine dipeptide in the gas phase and aqueous solution: a comparison of theoretical results”. *J. Phys. Chem.*, **96**, 3864-3870.

Torrie G.M., Valleau J.P. (1977). “Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling”. *J. Comput. Phys.*, **23**, 187-199.

Trucano P., Chen R. (1975). “Structure of graphite by neutron diffraction”. *Nature*, **258**, 136-137.

Ulberg D.E., Gubbins K.E. (1995). “Water adsorption in microporous graphitic carbons”. *Mol. Phys.*, **84**, 1139 - 1153.

van der Spoel D., Berendsen H.J.C. (1997). “Molecular dynamics simulations of Leu-enkephalin in water and DMSO”. *Biophys. J.*, **72**, 2032-2041.

van Gunsteren W.F., Berendsen H.J.C. (1990). “Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry”. *Angewandte Chemie International Edition in English*, **29**, 992-1023.

Vengadesan K., Gautham N. (2004). “Energy Landscape of Met-Enkephalin and Leu-Enkephalin Drawn Using Mutually Orthogonal Latin Squares Sampling”. *J. Phys. Chem. B*, **108**, 11196-11205.

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

Vermeer A.W.P., Norde W. (2000). “CD Spectroscopy of Proteins Adsorbed at Flat Hydrophilic Quartz and Hydrophobic Teflon Surfaces”. *J. Colloid Interface Sci.*, **225**, 394-397.

Vorobjev Y.N., Scheraga H.A. (1997). “A fast adaptive multigrid boundary element method for macromolecular electrostatic computations in a solvent”. *J. Comput. Chem.*, **18**, 569-583.

Wang Z.-X., Zhang W., Wu C., Lei H., Cieplak P., Duan Y. (2006). “Strike a balance: Optimization of backbone torsion parameters of AMBER polarizable force field for simulations of proteins and peptides”. *J. Comput. Chem.*, **27**, 781-790.

Warshel A., Levitt M. (1976). “Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme”. *J. Mol. Biol.*, **103**, 227-249.

Warshel A., Russell S.T. (1984). “Calculation of Electrostatic Interactions in Biological Systems and in Solutions”. *Q. Rev. Biophys.*, **17**, 283-422.

Warshel A., Russell S.T., Churg A.K. (1984). “Macroscopic Models for Studies of Electrostatic Interactions in Proteins: Limitations and Applicability”. *Proc. Natl. Acad. Sci. U. S. A.*, **81**, 4785-4789.

Warwicker J., Watson H.C. (1982). “Calculation of the electric potential in the active site cleft due to α-helix dipoles”. *J. Mol. Biol.*, **157**, 671-679.

Weiner S., Addadi L. (1997). “Design strategies in mineralized biological materials”. *J. Mater. Chem.*, **7**, 689-702.

Wilson K., Stuart S.J., Garcia A., Latour R.A., Jr. (2004). “A molecular modeling study of the effect of surface chemistry on the adsorption of a fibronectin fragment spanning the 7-10th type III repeats”. *J. Biomed. Mat. Res. A*, **69A**, 686-698.

Wolfenden R., Andersson L., Cullis P.M., Southgate C.C.B. (1981). “Affinities of amino acid side chains for solvent water”. *Biochemistry*, **20**, 849-855.

Wu Y., Tepper H.L., Voth G.A. (2006). “Flexible simple point-charge water model with improved liquid-state properties”. *J. Chem. Phys.*, **124**, 024503.

Xu F., Cross T.A. (1999). “Water: Foldase activity in catalyzing polypeptide conformational rearrangements”. *Proc. Natl. Acad. Sci. U. S. A.*, **96**, 9057-9061.

Yang Z., Zhao Y.-P. (2007). “Adsorption of His-tagged peptide to Ni, Cu and Au (1 0 0) surfaces: Molecular dynamics simulation”. *Engineering Analysis with Boundary Elements*, **31**, 402-409.

Zahn H. (1947). “Über die Struktur des α-Keratins”. *Z. Naturforsch.*, **2b**, 427.

Zauhar R.J., Morgan R.S. (1988). “The rigorous computation of the molecular electric potential”. *J. Comput. Chem.*, **9**, 171-187.

Zhang L., Hermans J. (1994). "3_{10} Helix Versus α-Helix: A Molecular Dynamics Study of Conformational Preferences of Aib and Alanine". *J. Am. Chem. Soc.*, **116**, 11915-11921.

Zhdanov V.P., Kasemo B. (1997). "Simulations of denaturation of adsorbed proteins". *Phys. Rev. E*, **56**, 2306-2309.

Zhdanov V.P., Kasemo B. (1998a). "Kinetics of irreversible adsorption of deformable proteins". *J. Chem. Phys.*, **109**, 6497-6501.

Zhdanov V.P., Kasemo B. (1998b). "Monte Carlo simulation of denaturation of adsorbed proteins". *Proteins: Struct., Funct., Genet.*, **30**, 168-176.

Zhdanov V.P., Kasemo B. (2000). "Ordering of adsorbed proteins". *Proteins: Struct., Funct., Genet.*, **40**, 539-542.

Zhdanov V.P., Kasemo B. (2001). "Folding of bundles of α-helices in solution, membranes, and adsorbed overlayers". *Proteins: Struct., Funct., Genet.*, **42**, 481-494.

Zheng M., Jagota A., Semke E.D., Diner B.A., McLean R.S., Lustig S.R., Richardson R.E., Tassi N.G. (2003). "DNA-assisted dispersion and separation of carbon nanotubes". *Nature Materials*, **2**, 338-342.

Zhou J., Chen S., Jiang S. (2003). "Orientation of Adsorbed Antibodies on Charged Surfaces by Computer Simulation Based on a United-Residue Model". *Langmuir*, **19**, 3472-3478.

Zhou J., Tsao H.-K., Sheng Y.-J., Jiang S. (2004). "Monte Carlo simulations of antibody adsorption and orientation on charged surfaces". *J. Chem. Phys.*, **121**, 1050-1057.

Appendix A. Protein Structure Definition

Proteins are linear combinations of amino acid residues. Table A.1 shows the structures of all 20 naturally occurring amino acids. Although they form a limited set, the number of combinations in which they can be arranged is vast.

A common feature of all amino acids is that they can be divided into two structural parts: a backbone and a side chain (Figure A.1). The backbone is built of an amino group, $-NH_2$, the so called α-carbon and its associated hydrogen atom, $-C_\alpha H-$, and a carboxyl group, -COOH. The side chain (-R in Figure A.1) defines an amino acid. It should be noted that, as Table A.1 shows, glycine and proline are somewhat special compared to other amino acids. Glycine is the simplest amino acid and its side chain consists of a single hydrogen atom (R=H), while the backbone of

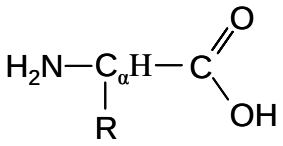
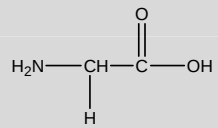
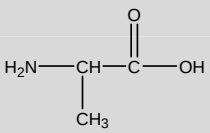
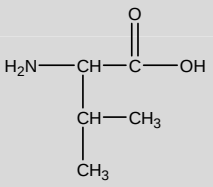
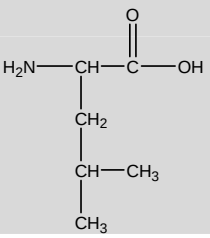
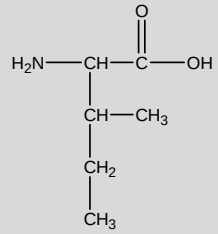
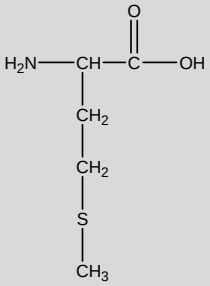
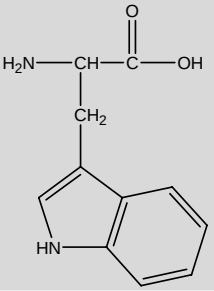
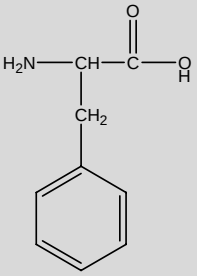
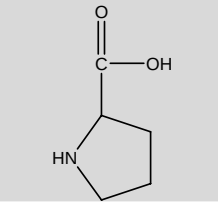
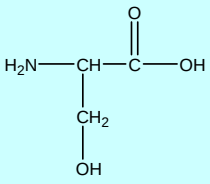
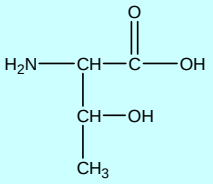
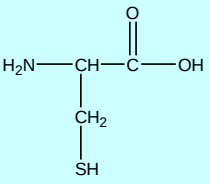
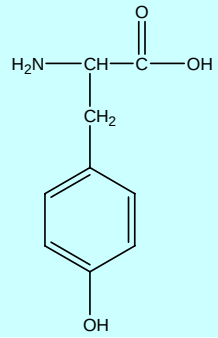
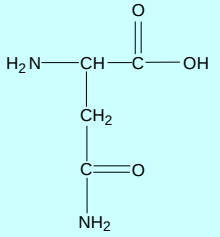
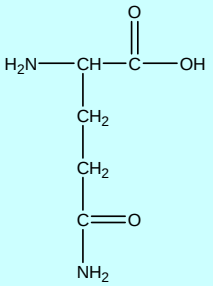
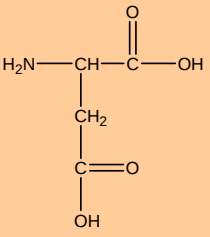
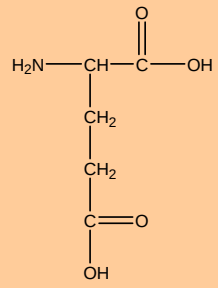
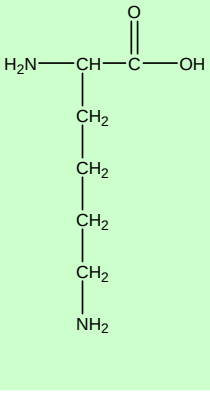
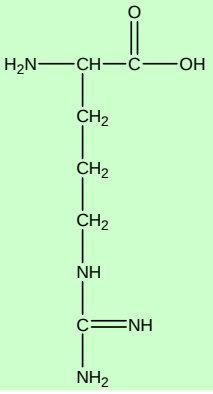
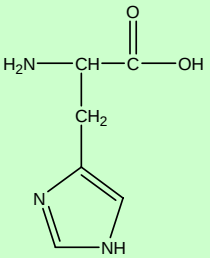


Figure A.1 General structural formula of all amino acids.

proline is looped and connected to the N atom from amino group, i.e. its backbone and side chain are fused into a cyclic structure. These odd features give glycine and proline some characteristics that other amino acids do not possess. Due to the small side chain, glycine is much more flexible than other amino acids (Rappé and Casewit, 1997), which has significant implications in conformational analysis of proteins that contain glycine, such as met-enkephalin. Proline, on the other hand, is much more rigid than other amino acids as the movement of its backbone is constrained by chemical bonds with the side chain. When found in the middle of an amino acid sequence, proline, due to this rigidity, has a tendency to interrupt canonical spatial arrangement of its surrounding residues.

Amino acid residues are connected to each other through peptide bonds. A peptide bond is formed in a dehydration process in which the carboxyl group of the first amino acid reacts with the amino group of the following residue, as shown in Figure A.2 (Rappé and Casewit, 1997). Since the new molecule possesses an amino

Table A.1 Natural amino acids[a]

			
Glycine (Gly, G)	Alanine (Ala, A)	Valine (Val, V)	Leucine (Leu, L)
			
Isoleucine (Ile, I)	Methionine (Met, M)	Tryptophan (Trp, W)	Phenylalanine (Phe, F)
			
Proline (Pro, P)	Serine (Ser, S)	Threonine (Thr, T)	Cysteine (Cys, C)
			
Tyrosine (Tyr, Y)	Asparagine (Asn, N)	Glutamine (Gln, Q)	Aspartic acid (Asp, D)
			
Glutamic acid (Glu, E)	Lysine (Lys, K)	Arginine (Arg, R)	Histidine (His, H)

a. Colour code: grey – hydrophobic; blue – polar; orange – acidic; green – basic;

group on its left hand side and carboxyl group on the right hand side, further amino acids can be added *ad infinitum*.

Macromolecules obtained by the polymerisation process depicted in Figure A.2 are called polypeptides if the number of amino acid residues is less than 50. Proteins are chains of amino acids that contain more than 50 residues (Rappé and Casewit, 1997). Some of the naturally occurring proteins can have thousands of amino acid residues, whilst others, such as met-enkephalin used in our studies, may contain only a few. Small polypeptides (several amino acid residues long) are commonly referred to as oligopeptides or simply peptides.

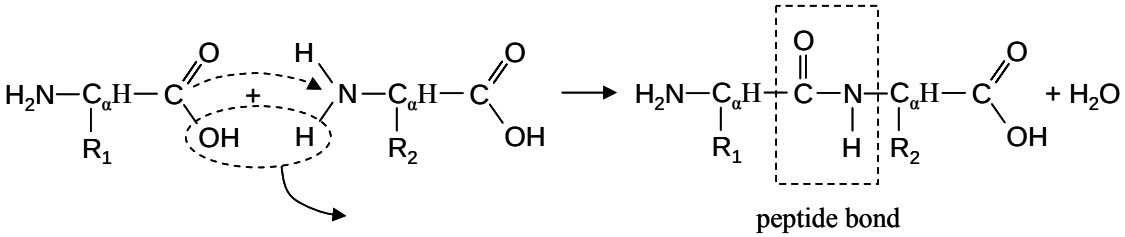


Figure A.2 Formation of the peptide bond.

The end of the molecule on which the amino group is located (left hand side in Figure A.2) is called N-terminus, while the other end is C-terminus. By convention, proteins are defined using a sequence of amino acids that starts from the N-terminus and finishes at the C-terminus (Rappé and Casewit, 1997).

The linear sequence of amino acids is what is known as the *primary structure*. Thus, protein primary structure is completely defined by knowing its constituent amino acids and their order from N- to C-terminus. Primary structure is, however, not sufficient to describe the 3D structure (also known as *conformation* (Cantor and Schimmel, 1980)) and biochemical characteristics of a protein. In order to describe these, one should also know the secondary, tertiary and, for protein aggregates, quaternary structure.

Secondary structure represents the configuration of continuous regions of a protein, in which amino acid backbones form locally symmetric 3D structure (Cantor and Schimmel, 1980). The most common elements of the secondary structure include helices, sheets and turns (Rappé and Casewit, 1997). α-helix is, for instance, very widely distributed in many biologically relevant proteins. It is characterised with a screw axis of symmetry and completion of a full helical turn after every 3.6 residues

on average (Cantor and Schimmel, 1980). The formation of a secondary structure is usually facilitated by establishing hydrogen bonds between different amino acid residues. Thus, in an α-helix, hydrogen bonds are formed between CO-group from the peptide bond of residue *i* and NH-group from the peptide bond of residue $i+4$.

Tertiary structure of a protein is the 3D structure or conformation of a complete chain of residues. It is a product of further rearrangements of secondary structure units (Rappé and Casewit, 1997). The tertiary structure is, effectively, a corollary of the secondary structure since the overall shape of a molecule is dictated by the structure of all of its individual units. On the other hand, the tertiary structure is tightly related to the biochemical function of a protein as the latter is a consequence of 3D positions of specific active groups or properties of the electric field formed by protein atomic charges.

Quaternary structure defines binding of different protein chains into a single biochemical unit. Single protein chains can be coupled by chemical bonds or van der Waals and electrostatic interactions (Rappé and Casewit, 1997). Ionic channels in cell membranes (Hille and Catterall, 2006) and virus shells or capsids (Cantor and Schimmel, 1980) are typical examples of heteromers formed by several individual amino acid chains. The quaternary structure is not an objective of our work since, at this stage, we are interested in prediction of 3D structures of isolated protein chains.

A.1 Ramachandran Plot

3D structure of a protein can be defined in terms of all the dihedral angles in its backbone. The backbone is built by consecutively adding N, C_α and C atoms of individual amino acids. This repetitive sequence of three atoms allows definition of three backbone dihedral angles (Figure A.3). ϕ is the angle that defines the torsion around N–C_α bond. ψ defines the torsion around C_α–C bond, while the torsion around the peptide bond, C–N, is defined by ω dihedral angle. The peptide bond, CO–NH, is normally represented as a single bond between carbon and nitrogen atoms. However, in reality it possesses some characteristics of a double bond due to hybridisation with the double C=O bond (Pauling, 1940), as shown in Figure A.4. From the perspective of conformational analysis of proteins, the most important double bond feature that the peptide bond possesses is its strong rigidity. While single bonds have relatively

low rotation energy barriers, the double bond is usually kept planar in either *cis* or *trans*-conformation (Pauling, 1940; Mizushima et al., 1950). It has been shown that *trans*-conformation, in which α-carbon atoms are on the opposite sides of the C–N bond (i.e. $\omega = 180^{o}$), is by far the dominant conformation of the peptide bond (Kitano et al., 1973; Kitano and Kuchitsu, 1973; Momany et al., 1975). The *trans*-conformation of the peptide bond will, accordingly, be used throughout this work, i.e. ω dihedral angle has been fixed at 180^{o} in all our simulations.

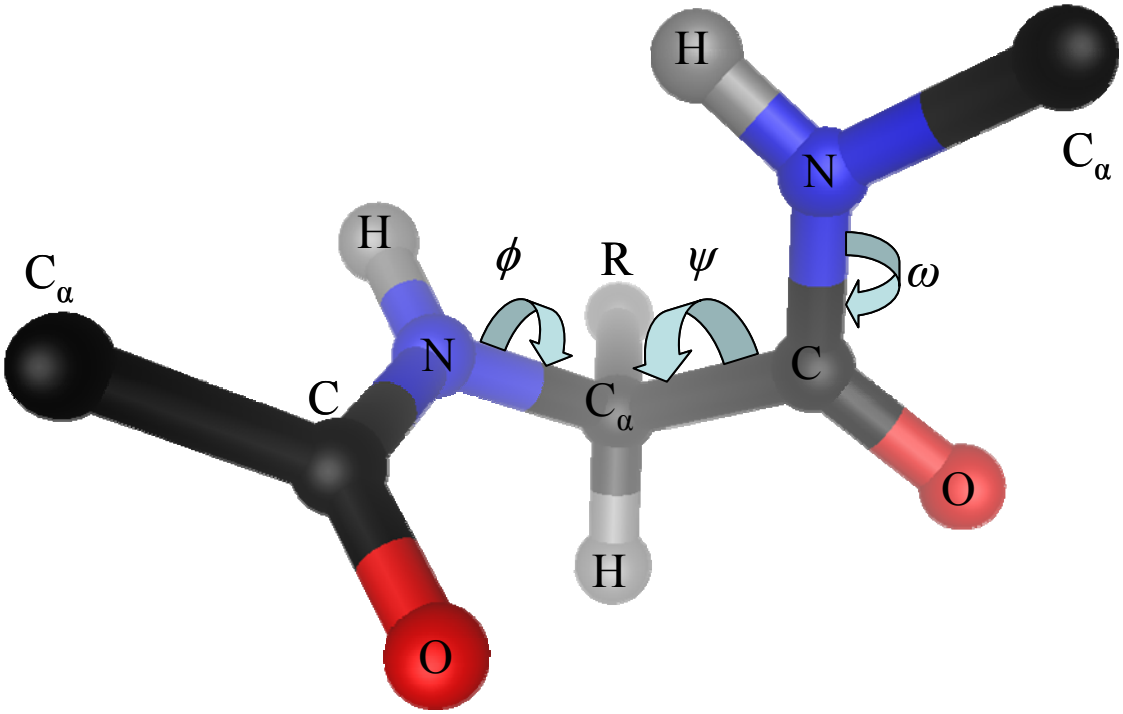


Figure A.3 Definition of the backbone dihedral angles in a protein: $\phi = \sphericalangle(\mathrm{C-N-C_\alpha-C})$; $\psi = \sphericalangle(\mathrm{N-C_\alpha-C-N})$; $\omega = \sphericalangle(\mathrm{C_\alpha-C-N-C_\alpha})$. In a similar way, side chain dihedral angles (not shown in the figure) describe torsion around bonds in side chains. For known values of bond lengths and angles, dihedral angles provide complete description of protein 3D structure.

With the peptide bond fixed in its *trans*-conformation, the 3D structure of the backbone can now be fully described with pairs of ϕ and ψ dihedral angles for each amino acid residue. A very convenient way to represent pairs of ϕ and ψ angles

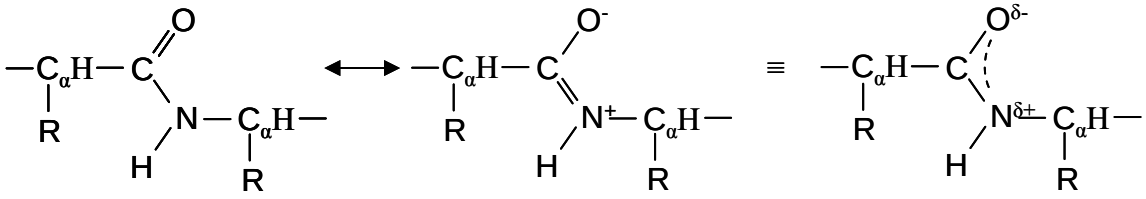


Figure A.4 Delocalisation (hybridisation) of the peptide bond in proteins.

graphically is the Ramachandran plot (Ramachandran et al., 1963). Ramachandran plot with sterically allowed regions and areas in which protein dihedral angles are most commonly found is shown in Figure A.5. Regions of the plot outside of the dashed lines are normally inaccessible to dihedral angles. However, as noted earlier, there are exceptions, especially with glycine and proline (Rappé and Casewit, 1997).

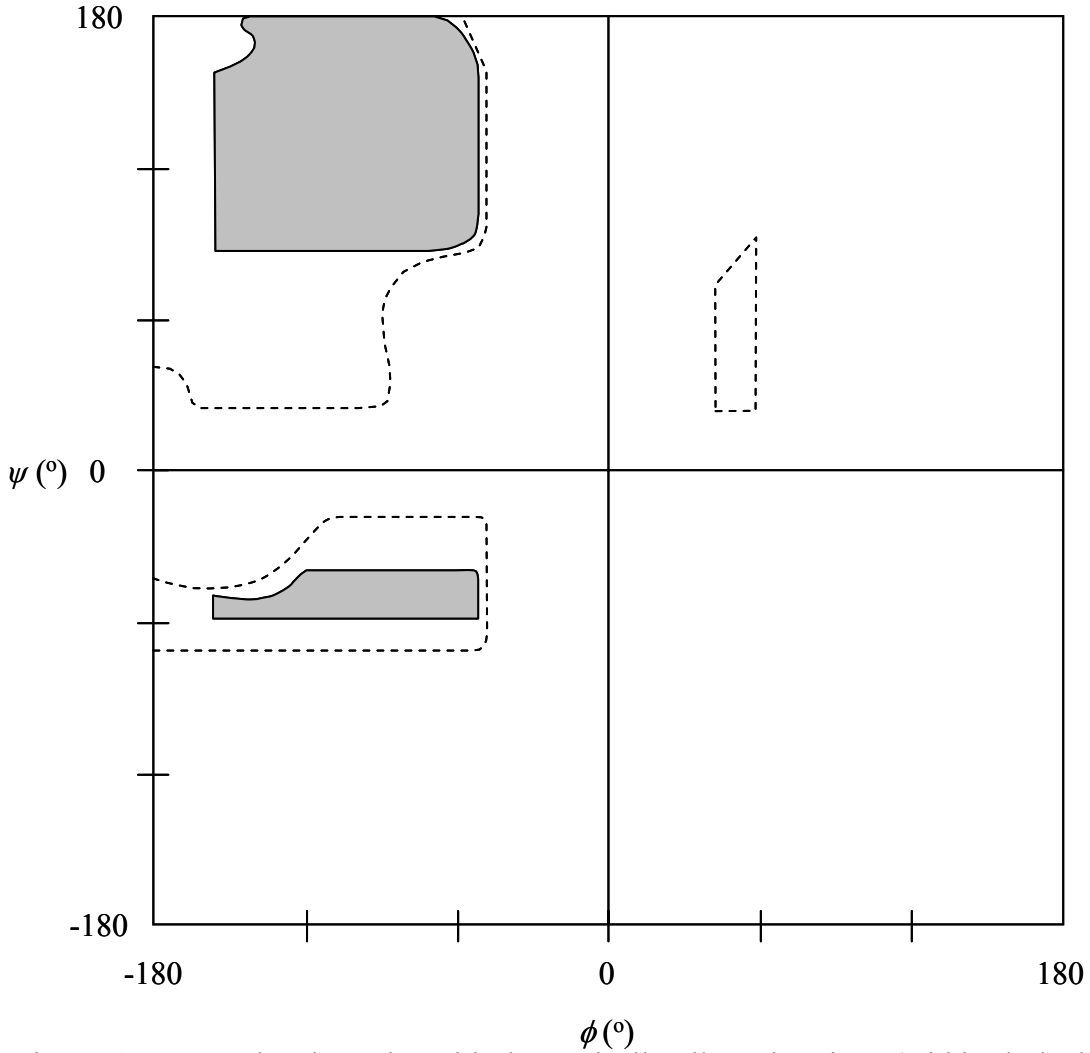


Figure A.5 Ramachandran plot with the sterically allowed regions (within dashed lines) and regions in which dihedral angle pairs are most commonly located (shaded).

Appendix B. Potential Energy of a Protein Conformation

Potential energy of a protein, as of any other molecule, is a function of its 3D structure. For example, change of atomic coordinates leads to changes in distances between different atoms, the ultimate consequence of which is the change in van der Waals and electrostatic energy between pairs of atoms. Overall intramolecular potential energy of proteins is, however, far more complex and includes several more contributions besides van der Waals and Coulomb interactions.

B.1 Decomposition of a Protein Potential Energy

The most accurate way to calculate potential energy of a molecule is by applying quantum mechanical methods. This approach is, however, computationally extremely expensive even for smaller molecules, and impractical for proteins that can include thousands of atoms. Potential energy of proteins and other large biomolecules is, therefore, approximated using a set of empirical equations, commonly referred to as force fields or potential energy models.

Force fields represent overall potential energy, U, of an isolated molecule in a specified conformation as a sum of several terms

$$U = U_{\mathrm{b}} + U_{\mathrm{a}} + U_{\mathrm{d}} + U_{\chi} + U_{\mathrm{es}} + U_{\mathrm{di}} + U_{\mathrm{eo}} + U_{\mathrm{hb}} \qquad \text{(B.1)}$$

where U_{b} is the energy of bond stretching or contraction, U_{a} is the angle bending energy, U_{d} is the energy associated with torsion around chemical bonds, U_{χ} represents the inversion term, and U_{es} , U_{di}, U_{eo} and U_{hb} are the non-bonded terms, electrostatic, dispersion, electron cloud overlap and explicit hydrogen bond energies, respectively. Bond and angle energies have self-explanatory names. They are a product of deformation of chemical bonds and angles between bonds from their equilibrium values. The torsion or dihedral angle energy, U_{d}, originates in torsion around a chemical bond. It is calculated for the rotation around the central bond for every set of four consecutive atoms. Figure B.1 shows graphical definitions of bond lengths, angles and dihedral angles used in calculation of the corresponding energy terms. Inversion energy relates to deformations of planar structures. Non-bonded energy terms are calculated for pairs of atoms that are separated by three or more

chemical bonds, e.g. atoms A and D in Figure B.1. Van der Waals interactions (dispersion and electron overlap terms) are calculated using Lennard-Jones potential or similar expression. Electrostatic energy term is derived from Coulombic expression for the force between two charged particles. It should be noted that many force fields do not include explicitly defined hydrogen bonds. Hydrogen bonds are, instead, often modelled through electrostatic and van der Waals interactions of relevant atoms. On the other hand, some force fields may include additional terms, such as the so called cross terms, which describe mutual influence of different deformations. These terms are, however, not very common and are not considered in detail here.

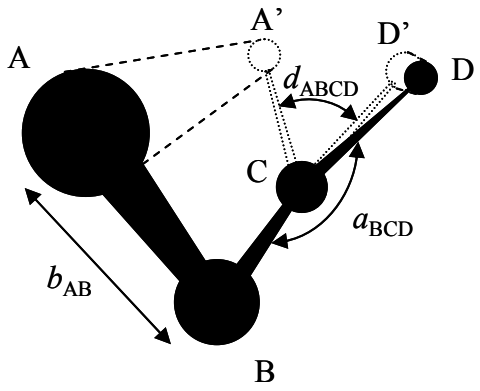


Figure B.1 Definition of chemical bonds, angles between them and dihedral angles. Bond stretching (or contraction) energy is calculated for all three chemical bonds: AB, BC and CD. Analogously, angle bending energy is calculated for both angles: ABC and BCD. The torsion energy is evaluated for the dihedral angle ABCD. The value of the dihedral angle, d_{ABCD}, is calculated by projecting bonds AB and CD into a single plane perpendicular to the central bond, BC, and evaluating the angle between the projections.

Some of the most commonly used force fields in the studies of protein 3D structure are Amber94 (Cornell et al., 1995), OPLS (Jorgensen et al., 1996), CVFF (Dauber-Osguthorpe et al., 1988) and ECEPP/3 (Nemethy et al., 1992), also used in Chapter 4 of this thesis. Numerical details of these force fields are summarised in Table B.1.

Table B.1 Terms of the force fields used in this thesis

Term[a]	Amber	OPLS	CVFF	ECEPP
U_{b}	$\sum_b K_b^r \left(r_b - \overline{r}_b\right)^2$	$\sum_b K_b^r \left(r_b - \overline{r}_b\right)^2$	$\sum_b K_b^r \left[1 - e^{-\alpha_b (r_b - \overline{r}_b)}\right]$	0
U_{a}	$\sum_a K_a^\theta \left(\theta_a - \overline{\theta}_a\right)^2$	$\sum_a K_a^\theta \left(\theta_a - \overline{\theta}_a\right)^2$	$\sum_a K_a^\theta \left(\theta_a - \overline{\theta}_a\right)^2$	0
U_{d}	$\sum_d \frac{K_d^\phi}{2}\left[1 + \cos\left(n_d\phi_d - \gamma_d\right)\right]$	$\sum_d \left\{ \frac{K_d^{\phi_1}}{2}\left[1 + \cos\left(\phi_d + \gamma_d^{\phi_1}\right)\right] + \frac{K_d^{\phi_2}}{2}\left[1 - \cos\left(2\phi_d + \gamma_d^{\phi_2}\right)\right] + \frac{K_d^{\phi_3}}{2}\left[1 + \cos\left(3\phi_d + \gamma_d^{\phi_3}\right)\right] \right\}$	$\sum_d K_d^\phi \left[1 + s_d \cos\left(n_d\phi_d\right)\right]$	$\sum_d K_d^\phi \left[1 + \cos\left(n_d\phi_d + \gamma_d\right)\right]$
U_{es}	$\sum_{i<j} f_{ij}^{es} \frac{q_i q_j}{\varepsilon r_{ij}}$	$\sum_{i<j} f_{ij} \frac{q_i q_j}{\varepsilon r_{ij}}$	$\sum_{i<j} \frac{q_i q_j}{\varepsilon r_{ij}}$	$\sum_{i<j} \frac{q_i q_j}{\varepsilon r_{ij}}$
U_{eo}	$\sum_{i<j} f_{ij}^{vdW} \frac{A_{ij}}{r_{ij}^{12}}$	$\sum_{i<j} f_{ij} \frac{A_{ij}}{r_{ij}^{12}}$	$\sum_{i<j} \frac{A_{ij}}{r_{ij}^{12}}$	$\sum_{i<j} f_{ij} \frac{A_{ij}}{r_{ij}^{12}}$
U_{di}	$-\sum_{i<j} f_{ij}^{vdW} \frac{B_{ij}}{r_{ij}^{6}}$	$-\sum_{i<j} f_{ij} \frac{B_{ij}}{r_{ij}^{6}}$	$-\sum_{i<j} \frac{B_{ij}}{r_{ij}^{6}}$	$-\sum_{i<j} \frac{B_{ij}}{r_{ij}^{6}}$
U_{hb}	0	0	0	$\sum_{hb} \left(\frac{A_{hb}}{r_{hb}^{12}} - \frac{B_{hb}}{r_{hb}^{10}} \right)$

a. Symbols are defined in equation (B.1).

b. The remaining symbols take the following meaning: r_b and $\overline{r}_b$ are the length and equilibrium length of a bond, *b*, respectively, of stiffness, K_b^r, θ_a and $\overline{\theta}_a$ are the angle, *a*, between two bonds having a common atom and its equilibrium value respectively, K_b^θ is the associated stiffness, ϕ_d, is the dihedral angle with the associated barrier height parameters, K_d^ϕ and K_d^ϕ, periodicity, n_d, and phase, γ_d, r_{ij} is the distance between two non-bonded atoms *i* and *j*, A_{ij} and B_{ij} are the associated van der Waals repulsive and attractive parameters respectively, q_i is the partial charge associated with atom *i*, ε is the dielectric constant of the surrounding medium, r_{hb} is the distance between a hydrogen atom and either a non-bonded oxygen or nitrogen atoms defining a hydrogen bond, A_{hb} and B_{hb} are the associated hydrogen bond repulsive and attractive parameters respectively, and f_{ij} are the scaling factors for 1-4 interactions.

c. The shaded terms are directly affected by the degrees of freedom varied during the EA simulations. The remainder are used to determine the bond lengths and angles as described in the text, which are then fixed during the EA – these terms will, therefore, have only an indirect impact on the performance of the EA. Improper torsion is not allowed and cross terms are neglected in PE calculation.

Appendix C. Determination of Switching Points in Polyalanine Adsorbtion on Smooth Surfaces

As shown in Chapter 5, polyalanine molecules switch conformation from α- to 3_{10}-helix and from 3_{10}- to 2_7-helix when the energy of the protein-surface interaction is gradually increased. The switching point has been defined as the protein-surface energy for which overall energies of two conformations of the same molecule are equal.

This analysis considers two conformations A and B either side of a switching point. The total potential energy of these conformations, E_t, is the sum of the intra-peptide potential energy, E_p, and the peptide-surface potential energy, E_{ps}

$$\begin{aligned} E_t(\mathrm{A}) &= U(\mathrm{A}) + E_{ps}(\mathrm{A}) \\ E_t(\mathrm{B}) &= U(\mathrm{B}) + E_{ps}(\mathrm{B}) \end{aligned} \tag{C.1}$$

Figure 5.7(a) shows that the intra-peptide potential energies are, to a good approximation, independent of the surface energy, E_s. This figure also shows, on the other hand, that the peptide-surface potential energy is very much dependent on the surface energy. Ignoring the small changes in conformations between the switching points, this dependency for the conformations can be well modeled by a straight line passing through the origin with a slope equal to

$$S_{ps} = dE_{ps} / dE_s \tag{C.2}$$

The variations of the total energy of the two conformations with surface energy now become

$$\begin{aligned} E_t(\mathrm{A}) &= U(\mathrm{A}) + S_{ps}(\mathrm{A})E_s \\ E_t(\mathrm{B}) &= U(\mathrm{B}) + S_{ps}(\mathrm{B})E_s \end{aligned} \tag{C.3}$$

The switching point is the value of E_s for which the total potential energies of the structures are equal (shown in insert of Figure 5.7(a)); i.e. $E_t(\mathrm{A}) = E_t(\mathrm{B})$. Thus, equating the expressions for the total energy of the two conformations, the switching point is calculated as

$$E_{sw} = \left[U(\mathrm{A}) - U(\mathrm{B})\right] \Big/ \left[S_{ps}(\mathrm{B}) - S_{ps}(\mathrm{A})\right] = -\Delta U / \Delta S_{ps} \tag{C.4}$$

as shown in equation (5.4).

Appendix D. Derivation of van der Waals Energy Between Langevin Dipoles and Smooth Surface

The derivation uses the same principles as those used by Steele (Steele, 1974) for the derivation of energies of van der Waals interactions between the molecules of gases and solid surfaces. The major difference is that, instead of 12-6 Lennard-Jones potential used in the initial step by Steele, a softer 9-6 potential is used here, as suggested by Florián and Warshel (2007).

Figure D.1 shows the interaction of a dipole j with one smooth plane. The overall energy of interaction is derived by integrating the energies of van der Waals interactions between the dipole and planar rings, E_r, from 0 to ∞.

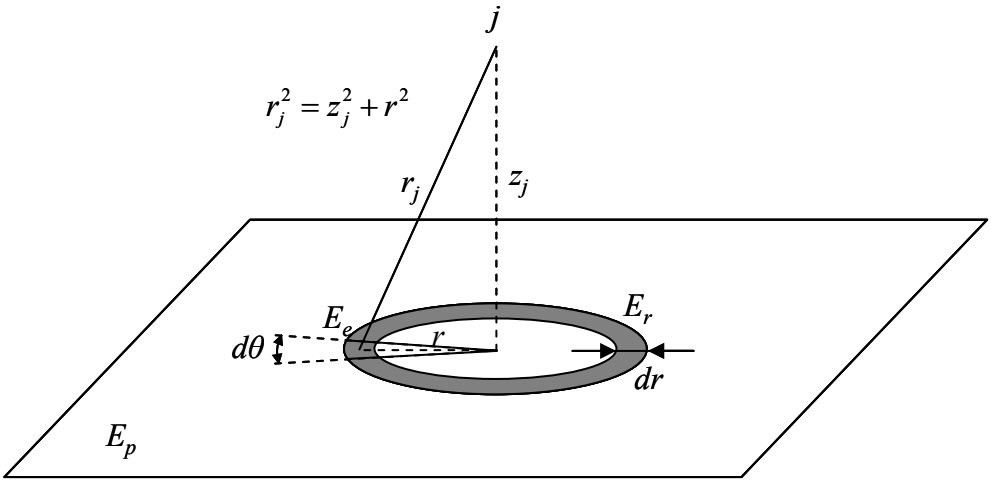


Figure D.1 Energies of interaction of dipole j with a ring element E_e, ring E_r, and plane E_p.

The energy of interaction between the dipole j and surface atom a on a distance r_j from each other is (Florián and Warshel, 1997)

$$E_a = k_{\mathrm{vdW}} \frac{C}{(r^*)^2} N_j \left[2\left(\frac{r^*}{r_j}\right)^9 - 3\left(\frac{r^*}{r_j}\right)^6 \right] \tag{D.1}$$

where k_{vdW} is the van der Waals coefficient $k_{\mathrm{vdW}} = 0.84\,\mathrm{kcal/mol}$, r^* and C are the dimension and interaction strength van der Waals parameters of carbon atoms in graphite, while N_j is the normalisation factor for balancing interactions in the coarse and fine grids of the LD model (Florián and Warshel, 1997).

In a smooth surface, however, all atoms are fused into a plane. The plane can be represented as a set of concentric infinitesimally thin rings, *r*. Each ring, in turn, can be divided into infinitesimal elements, *e*. The energy of interaction between an element *e* of a ring and the dipole *j* depends on the distance between the two and the number of atoms that can fit into the element, n_a. The number of atoms in an element, analogous to the derivation of Steele, can be calculated using surface density of atoms in the plane, ρ, and the area of the element, A_e: $n_a = \rho A_e$. The surface area of an element of a ring can be expressed in terms of the ring radius, *r*, and the angle of the circular segment covered by the element, $d\theta$: $A_e = \left(d\theta / 2\pi\right) 2\pi r dr = r dr d\theta$. The van der Waals energy of interaction between the dipole *j* and an element *e* can, therefore, be expressed as

$$E_e = \rho E_a r dr d\theta \tag{D.2}$$

The energy of interaction between the dipole and an infinitesimally thin ring which contains element *e* is calculated by integrating over all the elements

$$E_r = \int_0^{2\pi} E_e = \int_0^{2\pi} \rho E_a r dr d\theta = 2\pi\rho E_a r dr \tag{D.3}$$

The energy of interaction between the dipole and the plane is, then, obtained through integration of all the rings, whose radii range from 0 to ∞ (Figure D.1)

$$E_p = \int_0^{\infty} E_r = \int_0^{\infty} 2\pi\rho E_a r dr = 2\pi\rho \int_0^{\infty} E_a r dr \tag{D.4}$$

The distance between the ring at radius *r* from the vertical projection of the dipole *j* to the plane is (from Figure D.1) $r_j^2 = z_j^2 + r^2$, whilst the energy E_a is substituted from equation (D.1). Thus, equation (D.4) can be transformed into

$$E_p = 2\pi\rho k_{\mathrm{vdW}} \frac{C}{\left(r^*\right)^2} N_j \left[2\left(r^*\right)^9 \int_0^{\infty} \frac{r dr}{\left(z_j^2 + r^2\right)^{\frac{9}{2}}} - 3\left(r^*\right)^6 \int_0^{\infty} \frac{r dr}{\left(z_j^2 + r^2\right)^3} \right] \tag{D.5}$$

Using the integral solution

$$\int \frac{x dx}{\left(a^2 + x^2\right)^n} = -\frac{1}{2\left(n-1\right)\left(a^2 + x^2\right)^{n-1}} \tag{D.6}$$

for the limits indicated in equation (D.5) we obtain

$$\int_0^\infty \frac{x dx}{\left(a^2 + x^2\right)^n} = -\frac{1}{2(n-1)a^{2n-2}} \tag{D.7}$$

Replacing a with z_j and substituting into equation (D.5), the following expression is produced

$$E_p = 2\pi\rho k_{\text{vdW}} \frac{C}{(r^*)^2} N_j \left[2(r^*)^9 \frac{1}{7z_j^7} - 3(r^*)^6 \frac{1}{4z_j^4} \right] \tag{D.7}$$

$$E_p = 2\pi\rho k_{\text{vdW}} C N_j \left[\frac{2}{7}\left(\frac{r^*}{z_j}\right)^7 - \frac{3}{4}\left(\frac{r^*}{z_j}\right)^4 \right] \tag{D.8}$$

Finally, adding lower surface planes produces

$$E_{surf}^j = 2\pi\rho k_{\text{vdW}} C N_j \sum_{l=0}^{L-1} \left[\frac{2}{7}\left(\frac{r^*}{z_j + l\Delta}\right)^7 - \frac{3}{4}\left(\frac{r^*}{z_j + l\Delta}\right)^4 \right] \tag{D.9}$$

Summing these interactions for all the dipoles produces equation (7.4).

Appendix E. Bulk Contribution to Solvation Free Energy in the LD-EA Method for Molecules above a Solid Surface

Two cases should be considered – a charged molecule (or ion) with the net charge of q, and a neutral molecule with dipole moment $\boldsymbol{\mu}$. Each will be considered in turn.

E.1 Solvation of an Ion at the Water-Solid Interface

Following the derivation of the solvation free energy of an ion in an implicit solvent, proposed by Born (Born, 1920a), the same principle can be applied to obtain the free energy of solvation for an ion above the solid surface, Figure E.1. It should be noted, though, that while Born's integration extends from the ion radius to infinity in all directions, the integration conducted here has to start from R_b (which is the radius of the sphere whose internal volume is modelled by Langevin dipoles) and has to take into account the spherical dome immersed into the solid surface.

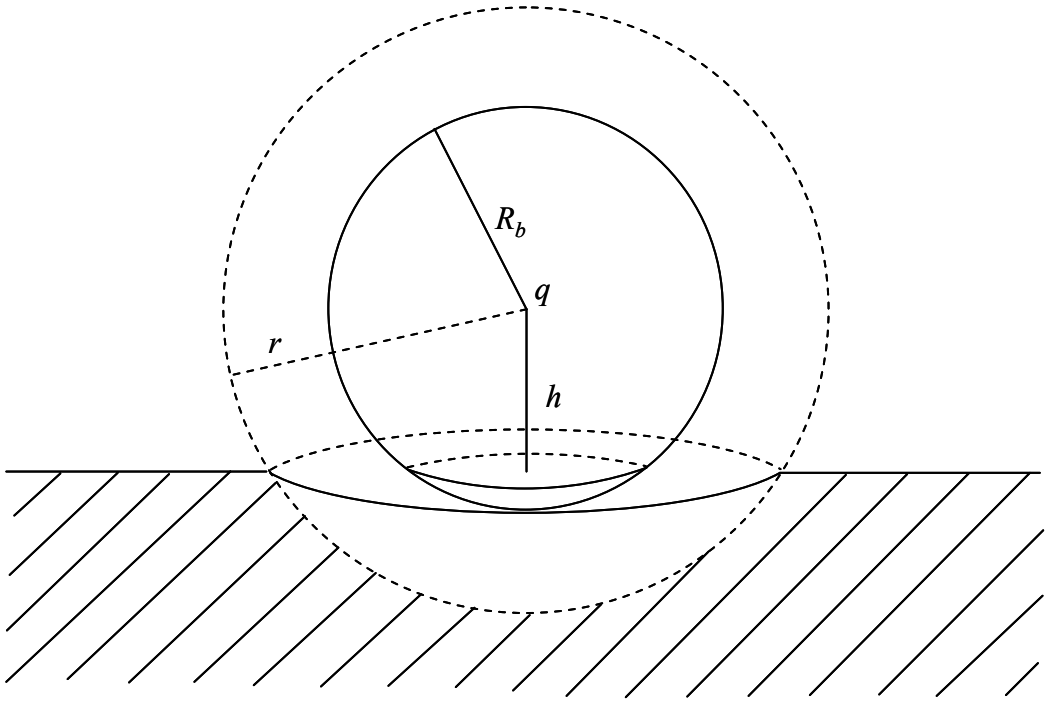


Figure E.1 Sphere for integration of free energy of solvation for an ion q on the distance h eabove the solid surface($h \le R_b$).

Electrostatic potential energy, U, of a continuous charge distribution in a volume V may be expressed in terms of electrostatic field generated by the charge throughout the volume V (Guru and Hiziroğlu, 2004). Although the charge distribution of the ion is not continuous, the discontinuity is outside of the volume of interest and the potential energy equation provided by Guru and Hiziroğlu may be applied

$$U = \frac{1}{2}\int_V \varepsilon_0 \varepsilon_r E^2 dV \tag{E.1}$$

where E is the magnitude of the electrostatic field $\mathbf{E}$ in an element of the volume dV, calculated as $E = |\mathbf{E}|$, while ε_0 and ε_r have earlier been defined as the electric permittivity of vacuum and relative dielectric constant, respectively.

For a point charge q placed in the center of a sphere with radius r, the electrostatic field on the surface of the sphere is calculated as

$$E = \frac{1}{4\pi\varepsilon_0\varepsilon_r}\frac{q}{r^2} \tag{E.2}$$

The volume of interest here is the volume outside of the sphere R_b (Figure E.1), but also above the solid surface. The integration is, therefore, performed by adding infinitesimally thin spherical shells from R_b to ∞, and subtracting the part of each shell that lies inside the solid surface.

The volume dV_S of a spherical shell of radius r and infinitely small thickness dr is calculated as the product of the area of shell surface and the thickness

$$dV_S = 4\pi r^2 dr \tag{E.3}$$

The volume of the spherical dome inside the solid surface is, similarly, calculated as

$$dV_D = 2\pi r(r-h)dr \tag{E.4}$$

where $2\pi r(r-h)$ is the area of the sphere (spherical dome) that is inside the solid surface, for the sphere whose center is on a distance h from the surface. Thus, the volume dV for integral in equation (E.1) can be obtained by subtracting the volume dV_D from the volume dV_S

$$dV = 4\pi r^2 dr - 2\pi r(r-h)dr = 2\pi r(r+h)dr \tag{E.5}$$

and for $h \le R_b$ the integral in equation (E.1) can now be solved

$$U = \frac{1}{2}\int_{R_b}^{\infty} \varepsilon_0 \varepsilon_r \left(\frac{1}{4\pi\varepsilon_0\varepsilon_r} \frac{q}{r^2} \right)^2 2\pi r(r+h)dr \tag{E.6}$$

$$U = \frac{1}{2}\int_{R_b}^{\infty} \varepsilon_0 \varepsilon_r \frac{1}{16\pi^2\varepsilon_0^2\varepsilon_r^2} \frac{q^2}{r^4} 2\pi r(r+h)dr = \frac{q^2}{16\pi\varepsilon_0\varepsilon_r}\int_{R_b}^{\infty} \frac{r+h}{r^3} dr \tag{E.7}$$

where integral on the left hand side of the equation can be decomposed into a sum of two table integrals of the form $\int \frac{dx}{x^2}$ and $\int \frac{adx}{x^3}$. Solving these, U becomes

$$U = \frac{q^2}{16\pi\varepsilon_0\varepsilon_r}\left[\left(-\frac{1}{r} \right)\Big|_{R_b}^{\infty} + h\left(-\frac{1}{2r^2} \right)\Big|_{R_b}^{\infty} \right] = \frac{q^2}{16\pi\varepsilon_0\varepsilon_r}\left(\frac{1}{R_b} + \frac{h}{2R_b^2} \right) \tag{E.8}$$

and finally

$$U = \frac{q^2}{16\pi\varepsilon_0\varepsilon_r} \frac{2R_b + h}{2R_b^2} \tag{E.9}$$

This is, however, only the potential energy in a dielectric environment with dielectric constant ε_r, such as water ($\varepsilon_r \approx 80$). In order to obtain the free energy of solvation (or, in this case, the contribution of implicitly represented water to it), one must find the difference of potential energies in a gas phase (in which $\varepsilon_r = 1$) and inside the dielectric environment. Thus, the free energy of solvation can be calculated as

$$\Delta G_{\text{es(c)}} = U^{sol} - U^{vac} = \frac{q^2}{16\pi\varepsilon_0\varepsilon_r} \frac{2R_b + h}{2R_b^2} - \frac{q^2}{16\pi\varepsilon_0} \frac{2R_b + h}{2R_b^2} \tag{E.10}$$

$$\Delta G_{\text{es(c)}} = \frac{q^2}{16\pi\varepsilon_0} \frac{2R_b + h}{2R_b^2}\left(\frac{1}{\varepsilon_r} - 1 \right) \tag{E.11}$$

If the whole sphere of Langevin dipoles is above the surface (i.e. $h > R_b$, Figure E.2), the integration is performed in two steps, where the first step integrates whole spherical shells from the radius R_b to h, whilst the second step (from h to ∞) extends the integration for the spherical shells partially immersed into the solid surface and is equivalent to the procedure shown in equations (E.1) to (E.11).

Integration of the inner space (R_b to h) uses the same basic principles (equation (E.1)), but, since none of the spherical shells from this space intersect the surface, the differential volume dV is equal to the volume of the whole spherical shell, dV_S, and the upper limit of the integration is h, rather than ∞. The contribution of this part of

the space is $\Delta G_{\text{es(c)}}^{\text{inner}} = \frac{q^2}{8\pi\varepsilon_0}\left(\frac{1}{R_b}-\frac{1}{h}\right)\left(\frac{1}{\varepsilon_r}-1\right)$. Summing up with the contribution from the outer space of the implicit solvent (h to ∞), which is, from equation (E.11), equal to $\Delta G_{\text{es(c)}}^{\text{outer}} = \frac{q^2}{16\pi\varepsilon_0}\frac{3}{2h}\left(\frac{1}{\varepsilon_r}-1\right)$, the overall contribution of implicit solvent to solvation free energy of a charged molecule on a distance h from the solid surface is

$$\Delta G_{\text{es(c)}} = \Delta G_{\text{es(c)}}^{\text{inner}} + \Delta G_{\text{es(c)}}^{\text{outer}} = \frac{q^2}{16\pi\varepsilon_0\varepsilon_r}\frac{4h-R_b}{2R_b h}\left(\frac{1}{\varepsilon_r}-1\right) \qquad \text{(E.12)}$$

Equations (E.11) and (E.12) are equivalent to the two cases ($h \le R_b$ and $h > R_b$) of equation (7.6).

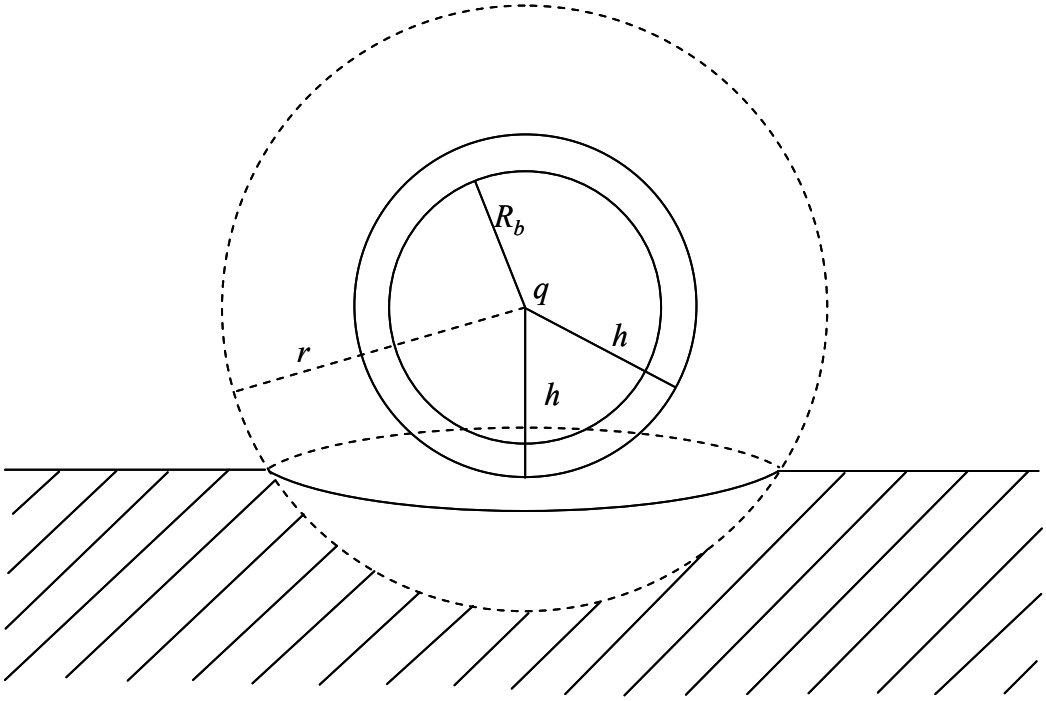


Figure E.2 Sphere for integration of free energy of solvation for an ion q on the distance h above the solid surface ($h > R_b$).

E.2 Solvation of a Dipole at the Water-Solid Interface

Derivation of the implicit solvent contribution to the free energy of solvation of a dipole with moment $\boldsymbol{\mu}$ and net charge equal to 0 is more complex than for the ion as, in addition to the dipole position, it also depends on its orientation. Calculation of $\Delta G_{\text{es(c)}}$ uses the approach described by Bell (Bell, 1931). Figure E.3 shows relevant spherical domains used in the derivation of free energy of solvation of a dipole. The

only difference, compared to the Bell's approach is decomposition of the inner domain to two domains, A and B, due to the presence of solid surface.

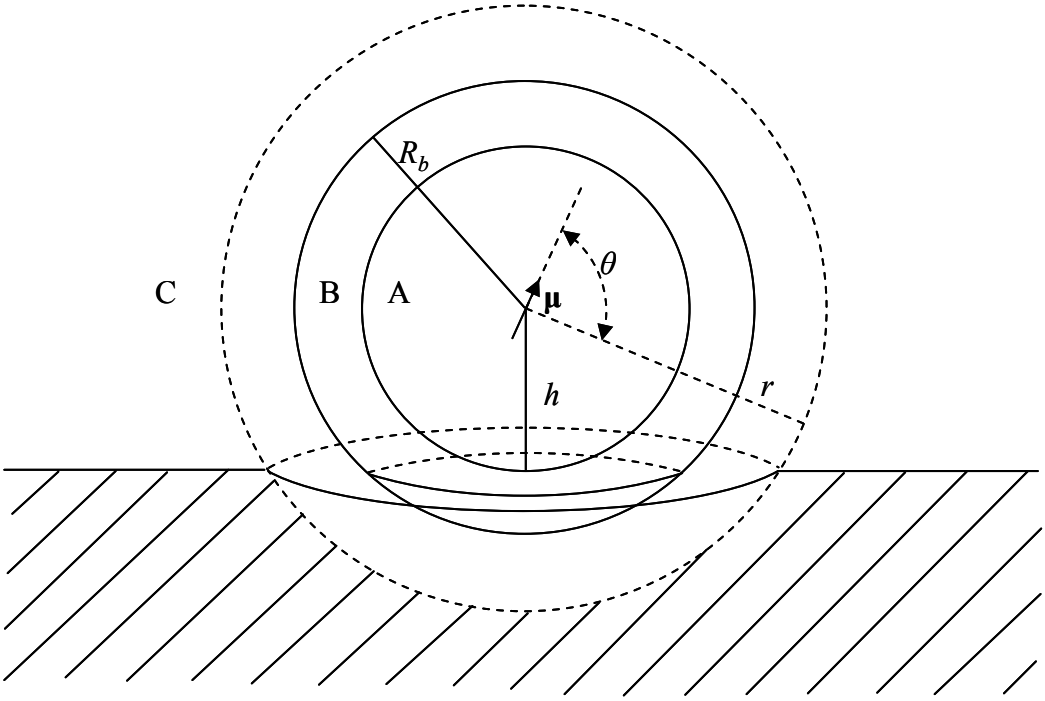


Figure E.3 Sphere for integration of free energy of solvation for a dipole $\boldsymbol{\mu}$ on the distance h above the solid surface ($h \le R_b$).

For the case shown in Figure E.3 ($h \le R_b$), the part of the solvent represented by Langevin dipoles is divided into two spherical domains, A and B. A is the sphere of radius h, i.e. the sphere that is centered in the point dipole $\boldsymbol{\mu}$ and touches the solid surface. The rest of the Langevin dipole domain is represented by a spherical shell B that extends from sphere A to the radius R_b. The domain outside of radius R_b is the implicitly represented solvent (domain C) and it is characterised with dielectric constant, ε_r, greater than 1 (for water solutions $\varepsilon_r \approx 80$). It should be noted that domains A and B are both characterised with $\varepsilon_r = 1$ since the solvent inside them is represented explicitly using Langevin dipoles.

Since the first spherical domain, A, is not interrupted by the presence of the solid surface, the integration of the contribution of that domain is identical to that proposed by Bell (1931)

$$\Delta G_{\mathrm{es(c)}}^{\mathrm{A}} = \frac{1}{2}\int_0^{\pi}\int_0^{h} \varepsilon_0 \frac{\mu^2}{16\pi^2\varepsilon_0^2}\left[\frac{2\lambda}{r^3h^3}\left(3\cos^2\theta - 1\right) + \frac{\lambda^2}{h^6}\right] 2\pi r^2 \sin\theta \, dr d\theta \qquad \text{(E.13)}$$

which, after solving, produces (Bell, 1931)

$$\Delta G_{\mathrm{es(c)}}^{\mathrm{A}} = \frac{\mu^2 \lambda^2}{24\pi\varepsilon_0 h^3} \tag{E.14}$$

where μ is the magnitude of the dipole $\boldsymbol{\mu}$, and λ is introduced by Bell in order to simplify the numerical representation (Bell, 1931)

$$\lambda = \frac{2\varepsilon_r - 2}{2\varepsilon_r + 1} \tag{E.15}$$

Integration of the domain B uses similar approach with two modifications. The first modification is in the limits for the integration. Whilst the domain A is integrated from 0 to h (radius of the spherical domain A), the integration of the domain B is performed from h to R_b, which, as for the solvation of ion, represents the radius of a sphere whose volume is filled by Langevin dipoles.

The second modification regards the intersection of the domain B by the solid surface (Figure E.3). Only the part of the domain B that is above the solid surface is considered for the calculation of the free energy of solvation, while the part that is immersed into the surface is neglected. Calculation of the contribution of the part above the solid surface has been performed by multiplying the volume of each infinitesimal spherical shell in the integration by its fraction above the solid surface. It should be noted that this procedure is only an approximation, but since our experience suggests that the overall implicit solvent contribution to the solvation free energy is very small compared to the contribution of the Langevin dipoles (up to several percents), this approximation is expected to be good enough for this purpose. For the sphere of radius r whose center is at a distance h from a solid surface ($h < r$), simple geometric manipulation shows that the fraction of the area of the spherical surface that is outside of the solid surface is $f = (r+h)/(2r)$.

Adding the fraction f and changing the integration limits in equation (E.13) produces

$$\Delta G_{\mathrm{es(c)}}^{\mathrm{B}} = \frac{1}{2}\int_0^{\pi}\int_h^{R_b} \varepsilon_0 \frac{\mu^2}{16\pi^2\varepsilon_0^2}\left[\frac{2\lambda}{r^3 R_b^3}\left(3\cos^2\theta - 1\right) + \frac{\lambda^2}{R_b^6}\right] 2\pi r^2 \frac{r+h}{2r}\sin\theta\, dr d\theta \tag{E.16}$$

which, after some mathematical manipulation, results in

$$\Delta G_{\mathrm{es(c)}}^{\mathrm{B}} = \frac{\mu^2\lambda^2\left(2R_b^3 + 3hR_b^2 - 5h^3\right)}{96\pi\varepsilon_0 R_b^6} \tag{E.17}$$

Contribution of the domain C is obtained using Bell's approach for the contribution of dielectric environment (equation (4) in (Bell, 1931)), with multiplying each spherical shell by the fraction *f* of the shell above the solid surface, as for the domain B

$$\Delta G_{\mathrm{es(c)}}^{\mathrm{C}} = \frac{1}{2}\int_0^{\pi}\int_{R_b}^{\infty} \varepsilon_0 \frac{\mu^2}{16\pi^2\varepsilon_0^2}\left[\frac{\varepsilon_r\left(1-\lambda\right)^2\left(3\cos^2\theta+1\right)}{r^6} - \frac{3\cos^2\theta+1}{r^6}\right] 2\pi r^2 \frac{r+h}{2r}\sin\theta\, dr\, d\theta \tag{E.18}$$

which, after simplification, produces

$$\Delta G_{\mathrm{es(c)}}^{\mathrm{C}} = \frac{\mu^2\left[\varepsilon_r\left(1-\lambda\right)^2-1\right]\left(4R_b+3h\right)}{96\pi\varepsilon_0 R_b^4} \tag{E.19}$$

Summing the contributions of all three domains (A, B and C), the total contribution of implicit solvent to the free energy of solvation of a dipole **μ** on a distance *h* from the solid surface is

$$\begin{aligned}\Delta G_{\mathrm{es(c)}} &= \Delta G_{\mathrm{es(c)}}^{\mathrm{A}} + \Delta G_{\mathrm{es(c)}}^{\mathrm{B}} + \Delta G_{\mathrm{es(c)}}^{\mathrm{C}} \\ &= \frac{\mu^2}{96\pi\varepsilon_0 R_b^6 h^3}\frac{\varepsilon_r-1}{\left(2\varepsilon_r+1\right)^2}\left[\left(\varepsilon_r-1\right)\left(16R_b^6-20h^6\right) - 4\left(2\varepsilon_r+1\right)h^3R_b^3 - 9h^4R_b^2\right]\end{aligned} \tag{E.20}$$

for $h \le R_b$.

Calculation of $\Delta G_{\mathrm{es(c)}}$ for a point dipole **μ** further away from the surface ($h > R_b$, Figure E.4) uses an analogous procedure, but divides the bulk solvent into domains B (spherical shell with inner radius of R_b and outer radius equal to *h*) and C (set of infinitesimally thin spherical shells partially immersed into the solid surface and extending from radius *h* to ∞), whilst the domain A now represents the whole sphere in which the solvent is represented by Langevin dipoles and for which $\varepsilon_r = 1$.

The contributions of the domain A can be calculated as

$$\Delta G_{\mathrm{es(c)}}^{\mathrm{A}} = \frac{1}{2}\int_0^{\pi}\int_0^{R_b} \varepsilon_0 \frac{\mu^2}{16\pi^2\varepsilon_0^2}\left[\frac{2\lambda}{r^3R_b^3}\left(3\cos^2\theta-1\right)+\frac{\lambda^2}{R_b^6}\right] 2\pi r^2\sin\theta\, dr\, d\theta \tag{E.21}$$

or, after simplification

$$\Delta G_{\mathrm{es(c)}}^{\mathrm{A}} = \frac{\mu^2\lambda^2}{24\pi\varepsilon_0 R_b^3} \tag{E.22}$$

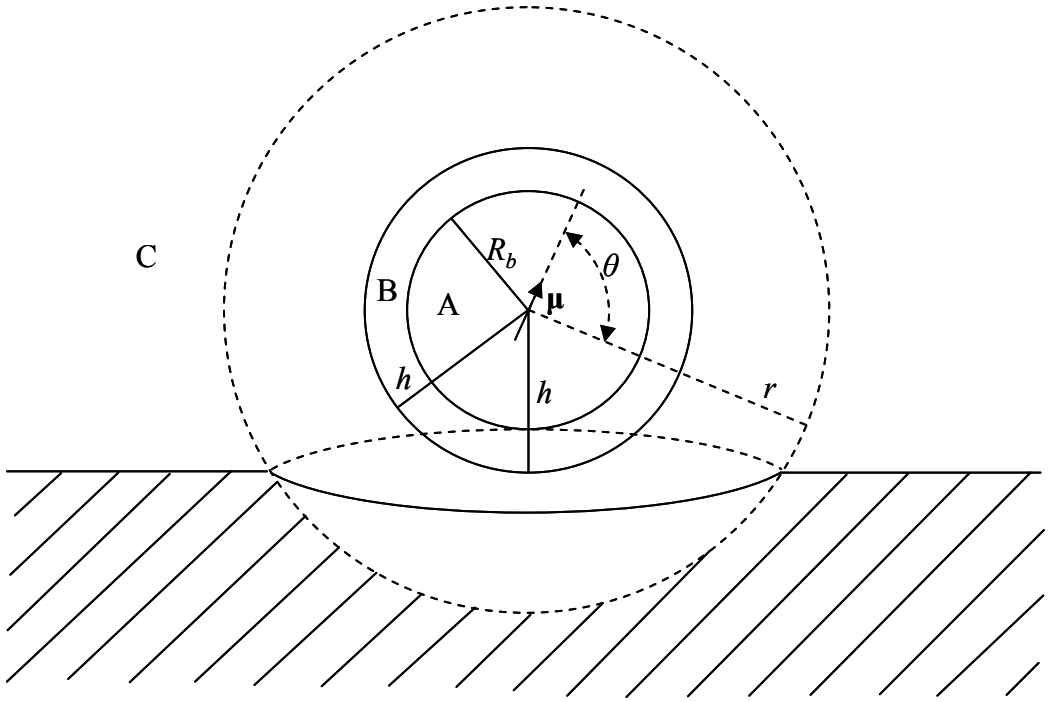


Figure E.4 Sphere for integration of free energy of solvation for a dipole $\boldsymbol{\mu}$ on the distance h above the solid surface ($h < R_b$).

The contribution of the domain B is

$$\Delta G^{\mathrm{B}}_{\mathrm{es(c)}} = \frac{1}{2}\int_0^{\pi}\int_{R_b}^{h} \varepsilon_0 \frac{\mu^2}{16\pi^2\varepsilon_0^2}\left[\frac{\varepsilon_r\left(1-\lambda\right)^2\left(3\cos^2\theta+1\right)}{r^6} - \frac{3\cos^2\theta+1}{r^6}\right] 2\pi r^2 \sin\theta dr d\theta \tag{E.23}$$

or simplified

$$\Delta G^{\mathrm{B}}_{\mathrm{es(c)}} = \frac{\mu^2\left[\varepsilon_r\left(1-\lambda\right)^2-1\right]\left(h^3-R_b^3\right)}{12\pi\varepsilon_0 R_b^3 h^3} \tag{E.24}$$

Finally, the domain C is included with the following contribution

$$\Delta G^{\mathrm{C}}_{\mathrm{es(c)}} = \frac{1}{2}\int_0^{\pi}\int_{h}^{\infty} \varepsilon_0 \frac{\mu^2}{16\pi^2\varepsilon_0^2}\left[\frac{\varepsilon_r\left(1-\lambda\right)^2\left(3\cos^2\theta+1\right)}{r^6} - \frac{3\cos^2\theta+1}{r^6}\right] 2\pi r^2 \frac{r+h}{2r} \sin\theta dr d\theta \tag{E.25}$$

which, when simplified, gives

$$\Delta G^{\mathrm{C}}_{\mathrm{es(c)}} = \frac{7\mu^2\left[\varepsilon_r\left(1-\lambda\right)^2-1\right]}{96\pi\varepsilon_0 h^3} \tag{E.26}$$

Summing up equations (E.22), (E.24) and (E.26), the contribution of implicit solvent to the solvation free energy of a point dipole on a distance h from solid surface for $h > R_b$ becomes

$$\begin{aligned} \Delta G_{\mathrm{es(c)}} &= \Delta G^{\mathrm{A}}_{\mathrm{es(c)}} + \Delta G^{\mathrm{B}}_{\mathrm{es(c)}} + \Delta G^{\mathrm{C}}_{\mathrm{es(c)}} \\ &= \frac{\mu^2}{96\pi\varepsilon_0 R_b^3 h^3} \frac{\varepsilon_r - 1}{\left(2\varepsilon_r + 1\right)^2} \left[\left(4\varepsilon_r - 1\right) R_b^3 - 8\left(2\varepsilon_r + 1\right) h^3 \right] \end{aligned} \tag{E.27}$$

Combining equations (E.20) and (E.27) covers all dipole positions above the solid surface, thus corresponding to equation (7.7).

References (Appendix)

Bell R.P. (1931). "The electrostatic energy of dipole molecules in different media". *Trans. Farad. Soc.*, **27**, 797-802.

Born M. (1920). "Volumen und Hydratationswärme der Ionen". *Z. Phys. A: Hadrons Nucl.*, **1**, 45-48.

Cantor C.R., Schimmel P.R. (1980). Part I: The Conformation of Biological Macromolecules. W. H. Freeman and Company, San Francisco.

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

Dauber-Osguthorpe P., Roberts V.A., Osguthorpe D.J., Wolff J., Genest M., Hagler A.T. (1988). "Structure and energetics of ligand binding to proteins: *Escherichia coli* dihydrofolate reductase-trimethoprim, a drug-receptor system". *Proteins: Struct., Funct., Genet.*, **4**, 31-47.

Florián J., Warshel A. (1997). "Langevin Dipoles Model for ab Initio Calculations of Chemical Processes in Solution: Parametrization and Application to Hydration Free Energies of Neutral and Ionic Solutes and Conformational Analysis in Aqueous Solution". *J. Phys. Chem. B*, **101**, 5583-5595.

Guru B.S., Hiziroğlu H.R. (2004). Electromagnetic Field Theory Fundamentals. 2nd Edition, Cambridge University Press, Cambridge.

Hille B., Catterall W.A. (2006). "Electrical Excitability and Ion Channels", in *Basic Neurochemistry: Molecular, Cellular and Medical Aspects*, 7th Edition, G.J. Siegel, R.W. Albers, S.T. Brady, D.L. Price, eds. Elsevier, Amsterdam,

Jorgensen W.L., Maxwell D.S., Tirado-Rives J. (1996). "Development and Testing of the OPLS All-Atom Force Field on Conformational Energetics and Properties of Organic Liquids". *J. Am. Chem. Soc.*, **118**, 11225-11236.

Kitano M., Kuchitsu K. (1973). "Molecular Structure of Acetamide as Studied by Gas Electron Diffraction". *Bull. Chem. Soc. Jpn.*, **46**, 3048-3051.

Kitano M., Fukuyama T., Kuchitsu K. (1973). "Molecular Structure of N-Methylacetamide as Studied by Gas Electron Diffraction". *Bull. Chem. Soc. Jpn.*, **46**, 384-387.

Mizushima S.-i., Simanouti T., Nagakura S., Kuratani K., Tsuboi M., Baba H., Fujioka O. (1950). "The Molecular Structure of N-Methylacetamide". *J. Am. Chem. Soc.*, **72**, 3490-3494.

Momany F.A., McGuire R.F., Burgess A.W., Scheraga H.A. (1975). "Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interactions, and intrinsic torsional potentials for the naturally occurring amino acids". *J. Phys. Chem.*, **79**, 2361-2381.

Nemethy G., Gibson K.D., Palmer K.A., Yoon C.N., Paterlini G., Zagari A., Rumsey S., Scheraga H.A. (1992). "Energy parameters in polypeptides. 10. Improved geometrical parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides". *J. Phys. Chem.*, **96**, 6472-6484.

Pauling L. (1940). The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry. 2nd Edition, Cornell University Press, Ithaka, NY.

Ramachandran G.N., Ramakrishnan C., Sasisekharan V. (1963). "Stereochemistry Of Polypeptide Chain Configurations". *J. Mol. Biol.*, **7**, 95-99.

Rappé A.K., Casewit C.J. (1997). Molecular Mechanics Across Chemistry. University Science Books, Sausalito, Calif.

Steele W.A. (1974). The Interaction of Gases with Solid Surfaces. Pergamon Press, Oxford.