A Global Optimization Approach for Searching Low Energy Conformations of Proteins

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch–Naturwissenschaftlichen Fakultät
der Universität Basel

von

Shantanu Roy
aus Indien

Basel, 2010

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von:

Basel, den 15.09.2009

Prof. Dr. Stefan Goedecker

Prof. Dr. Martin J Field

Prof. Dr. Torsten Schwede

Prof. Dr. E. Parlow

To my parents

ABSTRACT

De novo protein structure prediction and understanding the protein folding mechanism is an outstanding challenge of Biological Physics. Relying on the thermodynamic hypothesis of protein folding it is expected that the native state of a protein can be found out if the global minimum of the free energy surface is found. To understand the energy landscape or the free energy surface is challenging. The structure and dynamics of proteins are the manifestations of the underlying potential energy surface. Here the potential energy function stands on a framework of all-atom representation and uses purely physics-based interactions. For the solvated proteins the effective free energy is defined as an implicit solvation model which includes the solvation free energy, along with a standard all-atom biomolecular forcefield. A major challenge is to search for the global minimum on this effective free energy surface. In this work the Minima Hopping Algorithm (MHOP) to find global minima on potential energy surfaces has been used for protein structure prediction or in general finding the lowest energy conformations of proteins. Here proteins have been studied both *in vacuo* and in the aqueous medium. For short peptides starting from a completely extended conformation we could find conformational minima which are very close to the experimentally observed structures.

Contents

1 Introduction **1**

1.1 Protein Structure : Amino acids sequence 1

1.2 Protein folding and Structure prediction 8

1.3 Simulational Approach 11

1.3.1 Conformational Sampling 12

1.3.2 Forcefields 14

1.3.3 Solvation Models 15

1.4 Scope of the thesis 16

2 Energy Landscape **19**

2.1 Global Optimization and BEP 19

2.2 Conformational correspondence : Forcefield vs Density Functional methods 29

3 Methodological developments **45**

3.1 Implementation of Minima Hopping Algorithm 45

3.1.1 DIMERSOFT 46

3.1.2 Mode Decomposition 47

3.1.3 Enhanced Feedback 48

3.2 Preconditioned Steepest Descent 49

3.3 DIISIP : A new preconditioned-DIIS Geometry optimization scheme . . . 50

4 Biomolecules in gas phase: Conformational global optimization 61

4.1 Alanine dipeptide . . . 66

4.2 Met-enkaphalin . . . 66

4.3 $AcAla_8Lys + H^+$ . . . 71

4.4 $AcPheAla_5Lys + H^+$ and $AcPheAla_{10}Lys + H^+$ . . . 73

4.5 Helix-Turn-Helix motifs . . . 75

4.6 Pyridine-Pyrimidine oligomer . . . 76

5 Protein Structure Prediction (PSP) within All Atom Framework 87

5.1 Model and Method . . . 89

5.1.1 Forcefield . . . 89

5.1.2 Simulation Method . . . 91

5.2 Chignolin (1UAO) . . . 91

5.2.1 Results of MHOP Simulation . . . 91

5.3 Tryptophan Zipper (1LE1) . . . 104

5.3.1 Results of MHOP Simulation . . . 104

5.3.2 MHOP and MD . . . 107

5.4 Conclusions . . . 111

6 Summary 113

A Programs and Analysis tools used 117

Chapter 1

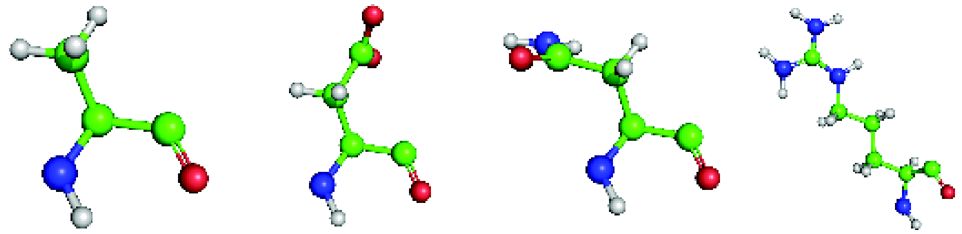
Introduction

One of the most important class of biomolecules present in every living organisms is protein. It evolved through the selective pressure to execute specific biological functions. In cells they are synthesized on ribosomes - and they are involved in functions ranging from catalysis of diverse chemical reactions to maintain the chemical potential across the cell membranes. The functional properties of proteins are dependent on the structures. The three-dimensional structure of these linear chain molecules are compact and “folded”. To understand its biological function, its structure has to be determined accurately. This chapter starts with a note on the very basics of protein structure. The next sections review the theoretical scientific approach towards understanding the protein structures. Finally, the scope of the thesis is presented.

1.1 Protein Structure : Amino acids sequence

The basic monomeric unit of a protein is an amino acid. There are twenty naturally occurring amino acids. All of the twenty amino acids have a central carbon atom (C_α), to which are attached a hydrogen atom, an amino group(NH2), and a carboxyl group(COOH). A side chain which is attached as to the C_α atoms is unique for each amino acid. There are twenty different naturally occurring amino acids shown in fig. 1.1. There are very rare occurrences of some other amino acids in proteins too, e.g selenocysteine or pyrrolysine.

Amino acids are linked together by the formation of peptide bonds to form a polymeric chain. When A peptide the carboxyl group of the an amino acid reacts with the amino group of the other, with the release of a water molecule a peptide bond is formed, as shown in fig. 1.2. The N-terminus is one end of the chain where the amino group is kept intact and a C-terminus is the other end of the chain where carboxy group stays. The protein backbone is the repetitive sequence of $[-N - C_\alpha - C-]$.

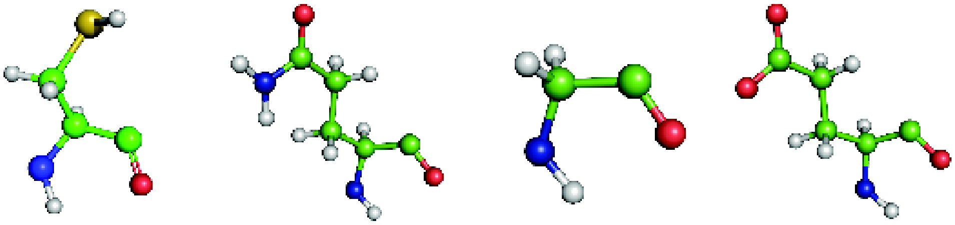


Ala (A)
Alanine

Asp (D)
Aspartate

Asn (N)
Asparagine

Arg (R)
Arginine

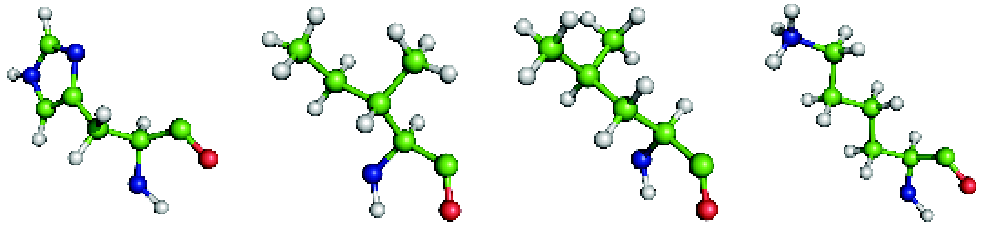


Cys (C)
Cystine

Gln (Q)
Glutamate

Gly (G)
Glycine

Glu (E)
Glutamine

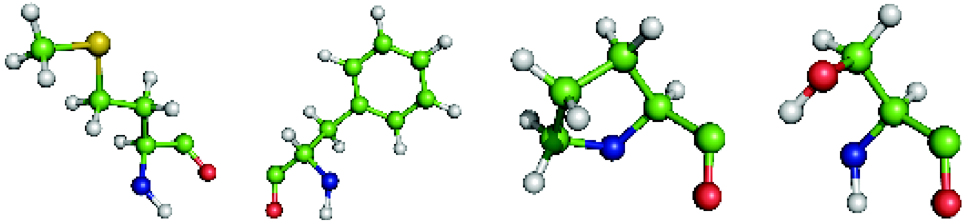


His (H)
Histidine

Ile (I)
Isoleucine

Leu (L)
Leucine

Lys (K)
Lysine

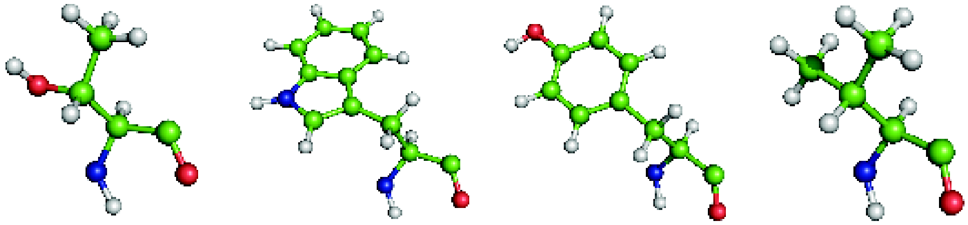


Met (M)
Methionine

Phe (F)
Phenylalanine

Pro (P)
Proline

Ser (S)
Serine



Thr (T)
Threonine

Trp (W)
Tryptophan

Tyr (Y)
Tyrosine

Val (V)
Valine

Figure 1.1: Twenty naturally occurring amino acids: Their structure, name, three-letter code and one-letter code. The structures are color coded with carbon(green), nitrogen(blue), oxygen(red) , hydrogen(white) and sulphur(orange)[37, 38]

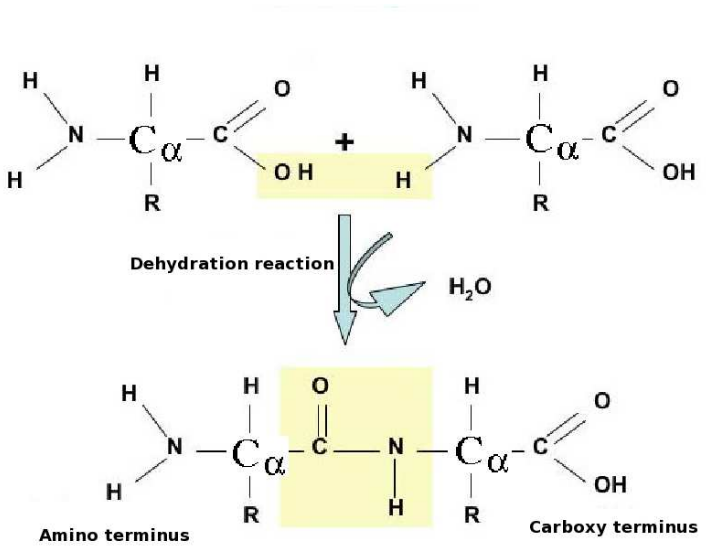


Figure 1.2: Formation of a peptide bond. NH_3^+ and COO^- are connected to the C_α atom.

Depending upon the chemical structure of the side chain, the amino acids are divided into three different classes(Branden and Tooze, 1999). The first class comprises those with strictly hydrophobic side chains Ala(A), Val(V), Leu(L), Ile(I), Phe(F), Pro(P), and Met(M). The second class includes four charged residues Asp(D), Glu(E), Lys(K) and Arg(R) and the third class comprises those with polar side chains Ser(S), Thr(T), Cys(C), Asn(N), Gln(Q), His(H), Tyr(Y) and Trp(W). The amino acid glycine(G) has only a hydrogen atom as the side chain and thus is the simplest of all the twenty amino acids. The amino acid proline(P) is also different from the rest as it is the only amino acid where both ends of the sidechain are covalently bound to the main chain.

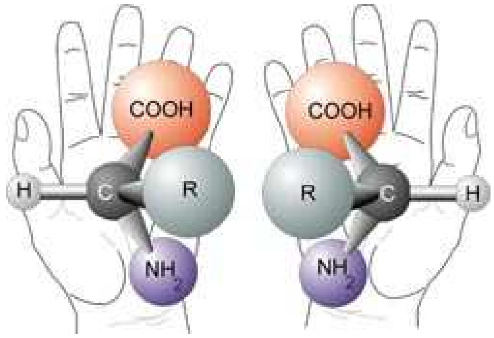


Figure 1.3: L-chiral and D-chiral forms of an amino acid.

All amino acids (except glycine) are chiral molecules which can exist either in the L or the D-form (see fig.1.3). Biological systems depend on specific detailed recognition of molecules involving differentiation between chiral forms. Amino acids are found in only one of the chiral forms, the L-Form, during protein synthesis. There is, however, no obvious reason why the L-form was chosen during the evolution and not the D-form(Weatherford and Salemme, 1979; Mason, 1984).

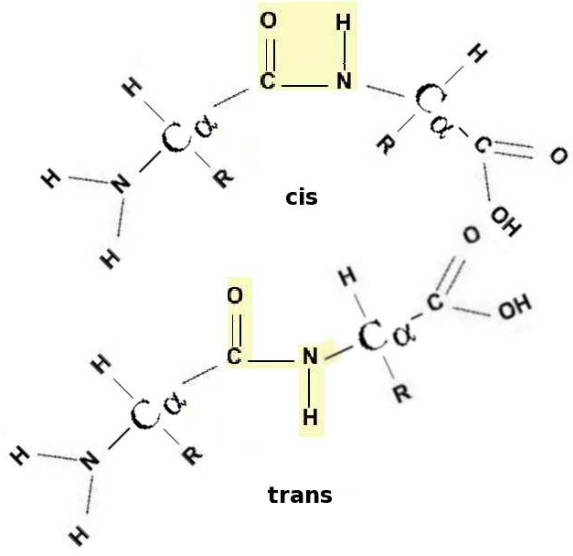


Figure 1.4: *cis* and *trans* forms of a peptide bond.

X-ray diffraction studies of crystals of small peptides by Linus Pauling and R. B. Corey indicated that the peptide bond is rigid, and planer. Pauling explained that this is largely a consequence of the resonance interaction of the amide, or in other words the ability of the amide nitrogen to delocalize its lone pair of electrons onto the carbonyl oxygen. The partial double bond turns the amide group into planar, posing in either the cis or trans isomers(shown in fig. 1.4). In the unfolded state of proteins, the peptide groups are free to isomerize and adopt both isomers; however, in the folded state, only a single isomer is adopted at each position (with rare exceptions). There is a clear preference for the trans form in most peptide bonds (roughly 1000:1 ratio in trans vs cis populations). However, X-Pro peptide groups tend to have a roughly 3:1 ratio. For all the amino acids except proline, the energy difference between cis and trans states is very large(Ramachandran and Mitra, 1976). But for proline, the energy difference between cis and trans states is small (Lesk, 2001) presumably because the symmetry between the C_α and C_δ atoms.

ϕ, ψ, ω angles Around the backbone of protein the three dihedral angles ϕ, ψ and ω are defined. The dihedral angle around the bond $N - C_\alpha$ is known as ϕ and the dihedral angle around the bond $C_\alpha - C$ is known as ψ. The dihedral angle ω is around the peptide bond $C - N$ and because of the planar peptide bond plane it is restricted to values 0^o or 180^o. As most residues in proteins have trans peptide bonds, the main chain conformation of each residue is determined by ϕ and ψ.

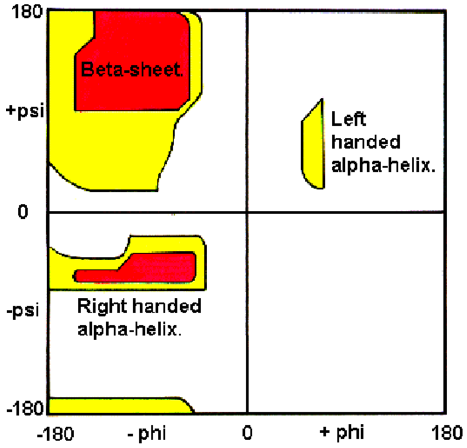


Figure 1.5: A schematic Ramachandran plot is shown - the major secondary structures are pointed out.

Only a fraction of the combinations of ϕ and ψ produce sterically allowed conformations. Such a ϕ-ψ plots showing the allowed regions is known as the Ramachandran plot (see fig. 1.5).

Protein structure

Primary Structure A protein is a sequence of different amino acids. The protein structures are classified into four categories depending upon the amount of information known. Primary structure describes the sequence of amino acids starting from amino(N) terminus to carbonyl(C) terminus. The primary sequence is written in either the one-letter code or three-letter code, e.g., NLYIQWLKDGGPSSGRPPPS (one letter code) or Asn-Leu-Tyr-Ile-Gln-Trp-Leu-Lys-Asp-Gly-Gly-Pro-Ser-Ser-Gly-Arg-Pro-Pro-Pro-Ser (three letter code) for the tryptophan-cage protein (PDB code:1L2Y).

Secondary Structure A secondary structure is about describing a protein in terms of frequently-observed regular structural elements. The most important secondary structures are as follows :

- Helix This is a spiral structure where tightly coiled backbone forms the inner part of the helix and the side chains project outwards. According to the hydrogen bonding pattern the helices are classified e.g. α-helix($i+4 \rightarrow i$ hydrogen bonding, i.e. there is a hydrogen bond between every i^{th}-residue and $i+4^{th}$ residue, angles $\phi \sim -57^o$, $\psi \sim -47^o$), 3_{10}-helix ($i+3 \rightarrow i$ hydrogen bonding, angles $\phi \sim -49^o$, $\psi \sim -26^o$), π-helix ($i+5 \rightarrow i$ hydrogen bonding, angles $\phi \sim -57^o$, $\psi \sim -70^o$) etc.

- Beta Sheet β-sheet is almost fully extended with multiple strands coming close to each other and the adjacent strand bound by inter-strand hydrogen bonds. According to the parallel or anti-parallel alignment of the adjacent strands there are Parallel β-strand ($\phi \sim -119^o$, $\psi \sim +113^o$ and Antiparallel β-strand ($\phi \sim -139^o$, $\psi \sim +135^o$.

There are eight types of secondary structure that Dictionary of Secondary Structure of Proteins (DSSP)[26] defines:

- G = 3-turn helix (3_{10}-helix). Min length 3 residues.
- H = 4-turn helix (α-helix). Min length 4 residues.
- I = 5-turn helix (π- helix). Min length 5 residues.
- T = hydrogen bonded turn (3, 4 or 5 turn)
- E = extended strand in parallel and/or anti-parallel β-sheet conformation. Min length 2 residues.
- B = residue in isolated β-bridge (single pair β-sheet hydrogen bond formation)
- S = bend (the only non-hydrogen-bond based assignment)

If the structure element falls within none of the above categories then that loop or irregular structure element is often called “random coil” or “coil”.

Tertiary Structure The tertiary structure is formed by the assembly of secondary structural elements along with loops into a three dimensional arrangement. The tertiary structure mainly has a hydrophobic core with charged residues on the surface of protein. The charged residues on the surface gives the protein its biological activity and is thus responsible for its biological function. Tertiary structures of proteins (independent folding chains) can still assemble themselves under physiological conditions in order to perform specific functions.

Quaternary Structure For a multi-domain protein the quaternary structure defines the overall arrangement of the tertiary structures of each of the domains.

Driving forces

Formation of the native state is a global property of a protein[37, 38]. In most cases, the entire protein (or at least a large part) is necessary for stability. This is because of the long-range nature of the stabilizing interactions which eventually bring the distant parts of an unfolded protein into spatial proximity by the folding process. Proteins are only marginally stable, and achieve stability only within narrow ranges of conditions of solvent and temperature. Outside of these regions proteins lose their definite compact structure, and even their helices and sheets, and take up states with disorder in the backbone conformation and specific interactions among residues. The chemical interactions responsible for the formation and stability of the native state and the biochemical properties and functions of a protein are as follows.

- Hydrogen bonding: Certain groups in proteins can form hydrogen bonds with water or other protein groups. The main chain has one H-bond donor (N-H) and H-bond acceptor (C=O) for each amino acid. In addition, some polar side-chains can form hydrogen bonds. The main chain, containing peptide groups, must pass through the interior, and some polar side chains are also buried. They thereby lose their interactions with water. To recover the energy, buried polar atoms form protein-protein hydrogen bonds. The standard secondary structures, helices and sheets, are achieved by the formation of hydrogen bonds by the main chain atoms.

- Hydrophobic effect: For proteins to take their native states in the aqueous environments, hydrophobic residues bury themselves in the interior and charged residues come on the surface. The accessible surface area of the protein, calculated from a set of atomic coordinates, measures the thermodynamic interaction between the protein and water.

- van der Waals forces and packing of protein: The packing of atoms in protein interiors contributes in two ways to the stability of structure. One is the exclusion of hydrophobic atoms from contact with water. The other is the dispersive attraction between the protein atoms.The cohesion of ordinary substances shows the existence of attractive forces between atoms and molecules. As the matter does not collapse, there must be limits to how far it can be compressed. This observation leads to the presence of repulsive forces at short range. The most general type of interatomic force, the van der Waals force, reflects this principle: The nearer the atoms, the stronger the attractive force, until the atoms are in contact, at which the forces become repulsive and strong. To maximize the total cohesive force, the atoms have to be brought close to each other. It is the requirement for a dense packing that imposes a requirement for structure in the interior of a protein. It produces a fit of the elements of secondary structure packed together in protein interiors.

- Covalent and coordinate chemical bonds: Some proteins contain covalent chemical bonds between side chains. These covalent bonds such as disulphide bridges between cystine residues are quite common.

1.2 Protein folding and Structure prediction

The functional property of a protein depends upon its three dimensional structure. Under physiological conditions, a particular sequence of amino acids in a polypeptide chain folds into a compact three-dimensional structure. This three dimensional structure, due to the specific properties, makes a protein perform a specific biological function. These single chains, which are folded into a respective three dimensional structure, can still assemble together to form more complex functional units. To understand the biological function of a protein, one needs to measure or predict its three dimensional structure from its amino-acid sequence. This prediction problem is still unsolved and remains one of the most basic challenges in biophysical chemistry. The fundamental reason why the prediction problem remains unsolved lies in the large size of the conformational space that is accessible to a single protein (Branden and Tooze, 1999; Berg et al., 2001).

Anfinsen's Dogma In his pioneering work, C. B. Anfinsen, showed that the necessary information for the polypeptide chain to fold into its native structure is contained in its sequence of amino acids. Protein refolding especially demonstrated that the native conformation of many proteins is reproducibly formed even when the proteins are in isolation. This observation can be explained, if the native state is lower in free energy than all other conformations. This observation led to the thermodynamic hypothesis (Anfinsen, 1973) that at the environmental conditions (temperature, solvent concentration and composition, etc.) at which folding occurs, the native structure is a unique, stable and kinetically accessible minimum of the free energy. The native state is a sufficiently low free energy minimum which is stable over a long timescale if not the true global minimum in the free energy - and thus folding process corresponds to an overall reduction of the free energy. The stability of each possible conformation of a polypeptide chain depends on the free energy change between native and unfolded states given by equation: $\Delta G = \Delta H - T\Delta S$ where ΔG, ΔH and ΔS are the differences between free energy, enthalpy, and entropy respectively, of the native and unfolded conformation. The enthalpic difference is the difference associated with atomic interactions (electrostatic interactions, van der Waals potentials, hydrogen bonding) whereas the entropy term describes hydrophobic interactions, thereby including the dominant interactions in protein folding, namely, the hydrophobic effect, hydrogen bonding and configurational entropy. The free energy of stabilization of proteins under ordinary conditions is typically only a few Kcal mol1 and slight changes in the surrounding conditions can force a protein to adopt a completely different conformation. In an unfolded protein, the polypeptide chain can adopt different rotameric positions around ϕ and ψ torsional angels, and side chain can adopt different rotamers around their dihedral angles. When folded, the ϕ and ψ dihedral angles of the polypeptide chain are nearly restricted to a narrow range of values, as are majority of χ angles. This loss of freedom translates into a loss of configurational entropy. This loss of configurational entropy must be overcome by favorable interactions, such as hydrogen bonding, increase in solvent entropy, etc, in order to fold a polypeptide chain into a stable conformation.

Levinthal's paradox While the experiments by C.B. Anfinsen and co-workers demonstrated that many proteins can adopt their native conformation spontaneously, it immediately raised a fundamental problem known as Levinthals paradox(Levinthal, 1968). Anfinsens experiments suggested that the native state of a protein is thermodynamically the most stable state under biological conditions. But a polypeptide chain has enormous number of possible conformations (at least 2^{100} for an 100 amino acid protein considering are only two possible conformations per amino acid). If one estimates that each state is reached in 1ps from a related conformation, such a chain would take $\sim 2^{100}$ ps (considering one ps per conformation) or $\sim 10^{10}$ years (even more than the estimated age of universe) to sample all possible conformations and to find the lowest energy state. Levinthal thus concluded that a specific folding pathway must exist and that protein folding is under kinetic control rather than thermodynamic control.

Landscape Theory This above mentioned issue can be resolved by considering a balance between kinetics and thermodynamics in an energy landscape perspective. According to the energy landscape paradigm, the free-energy landscape has a small gradient in all conformations towards the native state. Even in the absence of a unique folding pathway the protein dynamics is guided towards the native state. Projected to low dimension, the free energy surface thus has a funnel like slope. The landscape perspective explains the process of reaching a kinetically accessible minimum in free energy (satisfying Anfinsens experiments) and doing so on a realistic timescale (satisfying Levinthals concerns). Funneling landscape notion allows the existence of kinetically convergent multiple folding routes on funnel-like energy landscapes and thus supports the new view of folding which finds the unique folding pathway to be not the correct solution to the kinetic problem Levinthal posed. The funnel theory includes ruggedness on the funnel surface(see fig.1.6). The main idea is that while the folding landscape resembles a funnel globally but is to some extent rugged locally, i.e. with traps in which the protein can be trapped along the folding route. The funnel guides the protein through many different sequences of traps toward the low free energy folded (native) structure. Here there is no pathway but a multiplicity of folding routes. For small proteins, discrete pathways emerge only late in the folding process when much of the protein has almost reached the native ensemble. The simple parts of the folding process, where most of the real molecular organization is going on, occur in the early events of folding and can be described using a few parameters statistically characterizing the protein folding funnel[175].

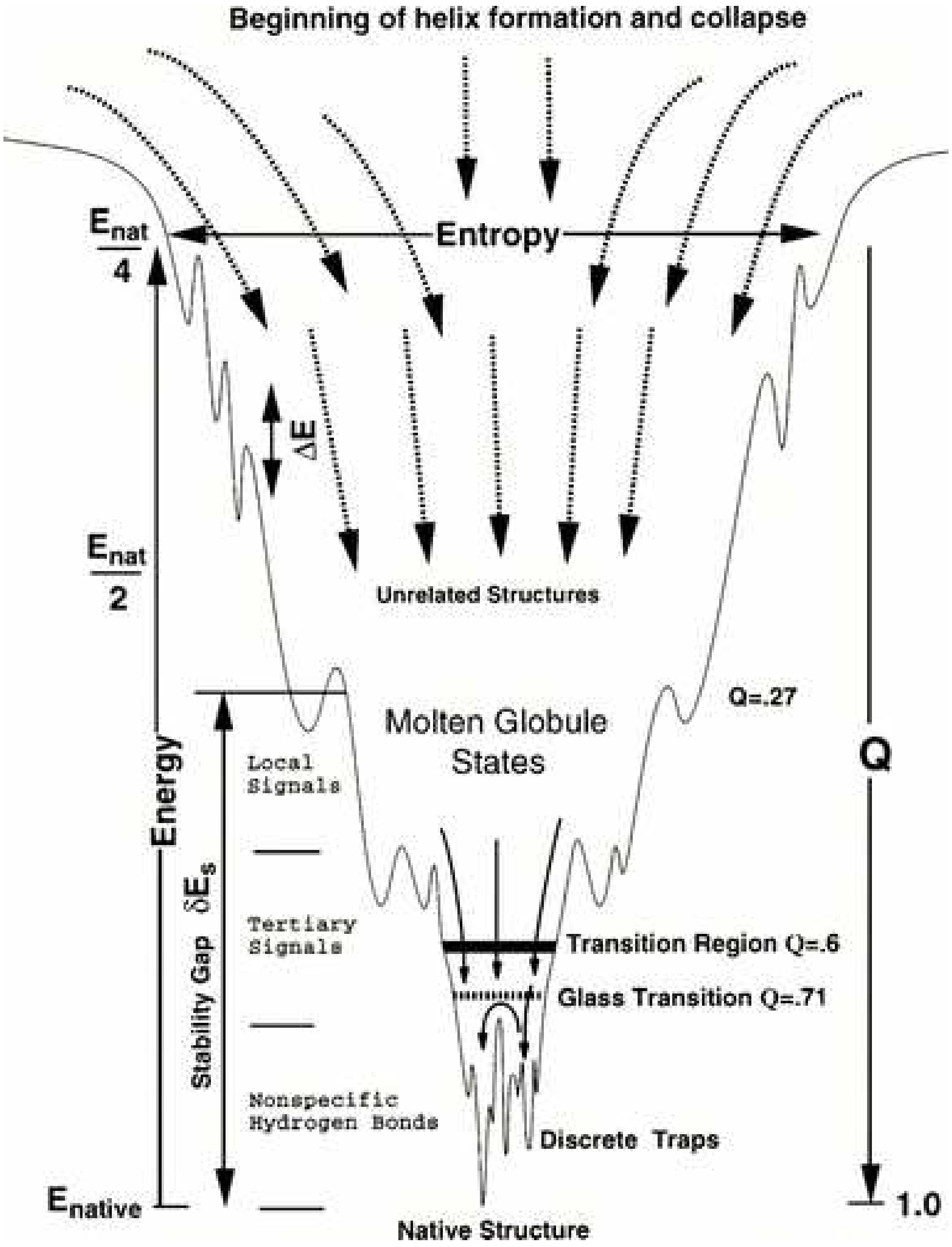


Figure 1.6: Schematic representation of the protein folding landscape (taken from Onuchic et al. (1997)[175])

Here the notion of a “unique pathway” folding model(Levinthal model) and extreme opposite to that, the “jigshaw puzzle” model of folding are apparently having very conflicting perspectives. The folding ’intermediates’ have been reported to have very specific regions of native-like structure. That is an indication that although the number of possible starting points for folding scale exponentially with the system size, the diversity in the transition region may not be large. At this point, with the inclusion of the concept of an ensemble of paths and the fact that for large systems a ’structure’ is meant by an ensemble of substates or superposition of stationary point in the landscape one can recover the two-state picture and the context of a reaction pathway. To describe whether the conformational changes happen through a “distinct pathway” or not, the term “distinct” perhaps becomes ambiguous. An appropriate grouping[3] or clustering of the conformations might contain a solution.

Spontaneity in folding There are folding accessory proteins that assist in folding *in vivo* so protein folding is not always a self assembly. Some of the them are: protein disulfide isomerases, peptidyl prolyl cis-trans isomerases and molecular chaperones. Molecular chaperones assist in protein folding by preventing or reversing aggregation (proteins hydrophobic parts sticking together in a misfolded way). Molecular chaperones specialize in multidomain and multisubunit proteins (and RNA/DNA but were not going there). They function by binding hydrophobic patches, releasing and rebinding as needed. They have inherent ATPase function; the energy released from ATP is used to perform their job. The two main classes of molecular chaperones are heat shock proteins and chaperonins. Some proteins are able to assemble on their own *in vivo*. This is evident since chaperones are not plentiful enough to fold every protein in the cell. Other proteins are always dependent on a chaperone for their efficient folding. Under cell stress conditions (particularly heat) proteins get misfolded or stuck together. In these conditions, molecular chaperones are crucial to protein folding for many proteins. Peptidyl prolyl cis-trans isomerase (PPIase) assists in isomerizing peptide bonds to the cis conformation before proline residues. This is crucial in making sharp turns in proteins. X-Pro peptide bonds are 90% trans and PPIases job is to catalyze the cis-trans flip. This flipping helps the protein folding to gain speed up. Protein disulfide isomerases (PDI) catalyze disulfide interchange reactions, allowing proteins to get to their native pairing of disulfide links faster. PDI has one S atom on a surface flanked by a groove of hydrophobic residues.

1.3 Simulational Approach

Any computational approach to study a chemical system requires a mathematical model to calculate the energy of the system as a function of its conformation. Central to the success of the study is the quality of the mathematical model used. For smaller chemical systems studied in the gas phase, quantum mechanical(QM) approaches are appropriate and feasible. Walter Kohn and John A. Pople jointly won the Nobel prize in chemistry in 1998 for the development of the density-functional theory and computational methods in quantum chemistry. However, these methods are typically limited to system of approximately hundred atoms or less, although approaches to treat large systems are under development. Systems of biophysical or biochemical interest typically involve macromolecules that contain thousands of atoms plus their surrounding environment. In addition to the large size of the system, the inherent dynamical nature of biomolecules require long simulation times, i.e. many energy calculations. Many processes of biophysical relevance occur on microsecond to millisecond time scales, while the individual time step of the methods commonly used today are of the order of femtosecond. Thus the energy function might be subjected to over 10^8 energy evaluations in a single simulation.

1.3.1 Conformational Sampling

There are two main approaches in performing molecular simulations: the stochastic (Monte Carlo) and the deterministic (Molecular Dynamics). Recent comparisons reveal that for polypeptide folding Monte Carlo takes 2-2.5 times smaller computational effort [18] than a comparable molecular dynamics study.

Stochastic This is based on exploring the energy landscape by random changes in the geometry of the molecule. In this way, a large area of the configurational space is searched. In Monte Carlo simulations, the system has no memory between two steps, i.e., the probability that the system might revert to its previous state is as probable as choosing any other state. As a result of stochastic simulation, the large number of configurations are accumulated and the energy function is calculated for each of them. This data can then be used to calculate thermodynamic properties of the system. Monte Carlo is not a deterministic method and does not offer time evolution of the system in a form suitable for viewing, but is well suited for investigating systems in certain ensembles. Monte Carlo simulations often gives rapid convergence of the calculated thermodynamic properties for small molecules[5]. The stochastic and hybrid methods include Monte Carlo with minimization (MCM) whereby combinatorial optimization with Monte Carlo is combined with energy minimization to find local minima; related approaches are basin hopping [3], and BB, a branch-and-bound method [13], electrostatically driven Monte Carlo (EDMC), self-consistent basin-to-deformed basin mapping (SCBDBM) that locates large regions of conformational space containing low-energy minima by coupling them to some of the greatly reduced number of minima on a highly deformed surface.

Deterministic The deterministic approach,e.g. molecular dynamics, actually simulates the time evolution of the molecular system and provides us with a trajectory of the system. Newtons or Lagranges equations are solved to obtain the coordinates and momenta along the simulation trajectory. Alternative approaches are based on solving Langevins equations when the solvent is treated implicitly with added friction and noise terms corresponding to the solvent effect. The information generated from simulations can in principle be used to fully characterize the thermodynamic state of the system. In practice, most simulations are interrupted long before there is enough information to derive absolute values of thermodynamic functions, however the differences between thermodynamic functions corresponding to different states of the system are usually computed quite reliably. In molecular dynamics, the evolution of the molecular system is studied as a series of snapshots taken at very close time intervals (usually of the order of femtoseconds). For large molecular systems the computational complexity is enormous and supercomputers or special attached processors have to be used to perform simulations spanning long enough periods of time to be meaningful. Typical simulations of small proteins including surrounding solvent cover the range of tens to hundreds of nanoseconds, i.e., they incorporate millions of elementary time steps.

Evolutionary Conformational space annealing (CSA) [14] is a frequently used genetic type algorithm; it combines essential aspects of the build-up procedure and a genetic algorithm and searches the whole conformational space in early stages, and then narrows the search to smaller regions with low energy. CSA has been applied to find the lowest energy structures of proteins and to protein sequence alignment [9, 14].

Hierarchical A hierarchical approach is based on carrying out an extensive coarse-grained search followed by a more detailed search on a potential energy surface of higher accuracy. An example is the scheme of CSA method using a united-residue (UNRES) force field first, following which the set of families of low-energy UNRES conformations obtained with CSA to be converted to an all-atom representation, and the search is continued with the EDMC method [3, 15]. In the context of global optimization of clusters the method dual-minima hopping(DMHM)[47] has been applied by a combined usage of density functional method and a forcefield.

Some of the successful algorithms of both categories are based on molecular dynamics (MD) and Monte Carlo (MC) techniques and include replica exchange [176], parallel tempering [177], simulated tempering [178], stochastic tunneling [179], REMUCA [180], MUCAREM [180], metadynamics [181] and hyperdynamics [182]. Applications to biomolecules of global optimization a pproaches by other groups include a study of a tryptophan-zipper (1LE1) using basin-sampling techniques with the PFF02 forcefield [127], investi gation of the folding pathway of β-hairpins using the activation-relaxation algorithm [183], and several examples of basin-hopping simulatio ns [184, 185, 186, 187].

1.3.2 Forcefields

Atomistic energy functions fulfill the demands required by computational studies of biochemical and biophysical systems. Empirical force fields use atomistic models, in which atoms are the smallest particles in the system rather than the electrons and nuclei used in quantum mechanical descriptions. The mathematical equations in these empirical energy functions include relatively simple terms to describe the physical interactions that dictate the structure and dynamical properties of biological molecules. These simplifications allow for the computational speed required to perform a large number of energy evaluations on biomolecules in their environment. Empirical energy functions were first used for small organic molecules, where it was referred as molecular mechanics, but are now regularly applied to biological systems. Some of the standard force fields available for biomolecular simulations are :

- AMBER : Assisted Model Building with Energy Refinement [28]
- CHARMM : CHemistry at Harvard Macromolecular Mechanics [29]
- GROMOS : GROningen MOlecular Simulation [30]
- OPLS : Optimized Potentials for Liquid Simulations [102]
- PFF : Protein Force Field [201]

For handling big biomolecular system for long timescale simulations required for the folding studies there have been many coarse-grained forcefield proposed in the recent past. A few of them are :

- ECEPP (Scheraga group)
- UNRES (Scheraga group)
- Martini forcefield (S.J. Marrink et al.)
- OPEP (Derreumaux)
- Simfold (Fujitsuka et al.)

These have been used for protein folding studies or in general biomolecular simulations.

Correlation among different Forcefields Although peptide folding has its own behavior in the nature, its actual presentation on the computer simulation is strongly dependent of the strategies we apply, especially the force field or energy model we use. In the past, many famous all-atom force field models have been proposed, like CHARMM [29], AMBER [28], GROMOS[30] and OPLS [102]. The force fields have been reported to be successful in various application. But actually these fields have their own preferred models. Okamoto and co-workers test these four typical force fields on the folding of short α-helix and β-hairpin with generalized ensemble method [31]. After comparing the secondary structure content in different replicas, they conclude that AMBER94 have excellent performance in α-helix simulation and GROMOS96 behaves well in β-hairpin simulation. AMBER96, CHARMM22 and OPLS-AA/L present proper tendencies to both of them. So, it is clear that for different types of peptides, choose a special force field is quite critical. Sometime it determines the final success of the simulation. Recently, AMBER has made a great improvement in the potential energy form and related parameters to fit the experimental data and high-level QM computations, like AMBER99 [32] and AMBER03 [33]. To compare the performance of these AMBER force fields and their revised versions: AMBER99m1, AMBER99m2, AMBER99off, Lwin and Luo[34] carried out many simulations for C-terminal β-hairpin from protein G [35, 36]. They discuss the structure distribution, folding thermodynamics and folding pathway in different conditions. The final results show that AMBER99ci and AMBER03 produce a good agreement with experiment data, like nuclear Overhauser effect (NOE) and native contacts. Furthermore, the free energy landscapes for these two force fields are a little different. AMBER99ci has a partial folded state in the landscape and AMBER03 does not. But this difference does not change their general agreement in various thermodynamics properties.

1.3.3 Solvation Models

The solvents are treated in the simulations either through the explicit solvent models, which treat the solvent in atomic detail, or through implicit solvent models, which generally replace the explicit solvent with a dielectric continuum. Explicit solvent models offer some of the highest levels of detail, they carry a burden of an additional thousands of degrees of freedom, this slows the computation. They generally require extensive sampling to converge properties of interest. SPC model, TIP3P and TIP4P etc are the examples of explicit models of water molecules.

Implicit solvent models trade detail and some accuracy for the “pre-equilibration” of solvent degrees of freedom and elimination of sampling for these degrees of freedom. Because of such pre-equilibration, implicit solvent methods generally require less computational effort and have become popular for a variety of biomedical research problems. Such implicit approaches include Poisson-Boltzmann and Generalized Born treatments of biomolecular solvation. The Poisson-Boltzmann equation (PB) describes the electrostatic environment of a solute in a solvent containing ions. The Generalized Born (GB) model is an approximation to the Poisson-Boltzmann equation. It is based on modeling the protein as a sphere whose internal dielectric constant differs from the external solvent. GBSA is simply a Generalized Born model augmented with the hydrophobic solvent accessible surface area SA term. The use of this model in the context of molecular mechanics is known as MM/GBSA. Although this formulation has been shown to successfully identify the native states of short peptides with well-defined tertiary structure[25], the conformational ensembles produced by GBSA models in other studies differ significantly from those produced by explicit solvent and do not identify the protein's native state[24]. It seems to produce artifacts like, overstabilized salt bridges possibly due to insufficient electrostatic screening or higher-than-native alpha helix population. Many Variants of the GB model have also been developed [27]. As ad-hoc quick strategies to estimate solvation free energy there are methods like ASA-based model involving the calculation of a per-atom solvent accessible surface area [23]. Another strategy is implemented for the CHARMM19 force-field and is called EEF1 [21]. EEF1 is based on a Gaussian-shaped solvent exclusion. EEF1 additionally utilizes a distance-dependent dielectric and the ionic side-chains of proteins are simply neutralized. The hydrophobic effect was added in EEF1 model and it is called Charmm19/SASA[22].

The implicit solvation schemes do not take into account the viscosity that water molecules impart by random collision with the solutes through the van der Waals repulsion. Although this makes the sampling of configurations and phase space much faster, it can lead to misleading results when kinetics are of interest. Viscosity can be added by using Langevin dynamics instead of Hamiltonian dynamics and choosing an appropriate damping constant for the particular solvent. Regarding the hydrogen-bonds with water, the average energetic contribution of protein-water hydrogen bonds may be reproduced with an implicit solvent. However, the directionality of these hydrogen bonds will be missing.

1.4 Scope of the thesis

In this thesis we explore an alternate approach for protein structure prediction and folding that is based on the Anfinsens hypothesis that most proteins are in thermodynamic equilibrium with their environment in their native state. For proteins of this class the native conformation corresponds to the global optimum of the free energy of the protein. We know from many problems in physics and chemistry that the global optimum of a complex energy landscape can be obtained with high efficiency using stochastic optimization methods. These methods map the folding process found in nature onto a fictitious dynamical process that explores the free-energy surface of the protein. By construction these fictitious dynamical processes not only find the conformation of lowest energy, but typically characterize the entire low-energy ensemble of competing metastable states. Since the total free energy change for protein folding under physiological conditions is small, often only a few kcal/mol, a characterization of the low-energy ensemble of thermodynamically accessible protein conformations may be sufficient not only to predict the structure of the protein, but also to characterize the folding process. The technique that has been used here is global optimization of the effective free energy function. The global optimization technique is *Minima Hopping Algorithm*. The effective free energy function has been described in detail in section 5.1.1. The aim of the work that the thesis addresses can be summarized as follows :

1 Efficient implementation of global optimization scheme for finding the global minimum and many of the low energy conformations proteins

2 To Use an all-atom forcefield based Effective Free Energy model for predicting protein structure and later on for structural refinement

3 To develop new algorithms (i.e.geometry optimization) to make the scheme more efficient

Chapter 2 deals with a couple of investigations done on the potential energy landscape which would in turn help in enhancing the performance of the minima hopping algorithm. The effect of Bell-Evans-Polanyi principle on molecular dynamics and global optimization has been tested using different potential functions which actually were modelled on different physical systems, e.g. Argon cluster, silicon cluster etc. The second part of this chapter would focus on the conformational energy landscape of protein described by a biomolecular forcefield and density functional methods.

Chapter 3 contains a detailed account of the methodological aspects. Different optimization techniques are described here.

Chapter 4 describes the application of minima hopping global optimization method for biomolecules in gas phase.

Chapter 5 describes the application of minima hopping algorithm for protein structure prediction and folding. The studies were conducted using an implicit solvation model of water. We predicted the putative global minima of two β-hairpins and discussed the minima hopping search profiles.

Chapter 2

Energy Landscape

To understand the energy landscape one way is to analyze different stationary points, their stability, the study of the barrier heights etc. In the first part of this chapter, the Bell-Evans-Polanyi principle that is valid for a chemical reaction that proceeds along the reaction coordinate over the transition state is extended to molecular dynamics trajectories that in general do not cross the dividing surface between the initial and the final local minima at the exact transition state. Our molecular dynamics Bell-Evans-Polanyi principle states that low energy molecular dynamics trajectories are more likely to cross into the basin of attraction of a low energy local minimum than high energy trajectories. In the context of global optimization schemes based on molecular dynamics our molecular dynamics Bell-Evans-Polanyi principle implies that using trajectories that have an energy that is only somewhat higher than the energy necessary to overcome the barriers lead fastest to the global minimum of funnel like energy landscapes. In the second part of this chapter, a comparative study of two kinds of energy landscapes, one of the OPLS forcefield and the the other one of density functional interactions would be presented.

2.1 Global Optimization and BEP

The Bell-Evans-Polanyi (BEP) principle is a conceptual tool in chemistry that is introduced in standard textbooks on physical chemistry [39],[40]. It gives a relation between the free energy ΔG released in a chemical reaction and the activation free energy ϵ_a for the reaction. It is generally assumed to be well obeyed for chemically similar reactions. It was qualitatively first put forward by Brønsted [41] who observed that strongly exothermic reactions have a low activation energy. A more quantitative relation was then derived by Polanyi *et al*[42],[39] who approximated the potential energy surface by straight lines. This approximation leads to a linear relation between the activation energy ϵ_a and the free energy of the reaction ΔG:

$$\epsilon_a = k_1 + k_2 \Delta G \quad , \tag{2.1}$$

where k_1 and $k_2 > 0$ are constants that depend on the slopes of the lines. A more accurate approach by Marcus[43],[40] approximates the potential energy surface by two parabolas centered at the two local minima of the energy, which leads to an additional quadratic term in Eq. 2.1.

In a chemical reaction, the reaction coordinate connects the educt A with the product B. In this article we will study the BEP principle not for this hypothetical path along the reaction coordinate but for molecular dynamics (MD) trajectories that cross the dividing hypersurface between the two basins of attraction of two local minima on the potential energy surface. The notions of educt and product are replaced by the notions of initial and final local minima in this context. We will show that the BEP principle is also valid in the context of MD. Since our study requires the calculation and statistical evaluation of a very large number of local minima and saddle points, we will initially base our study on a Lennard Jones cluster containing 55 atoms [44] for which stationary points can be calculated rapidly.

We will first investigate how well the traditional BEP principle is satisfied for these Lennard Jones clusters. To do so we have searched for more than 130000 first order saddle points G_i^s on the potential energy surface connecting energetically low local minima. Subsequently we have moved the system by a small amount away from the saddle point along the two directions where the curvature is negative, i.e we moved the system in the direction of the eigenvector associated with the negative eigenvalue of the Hessian matrix and in the negative direction of this eigenvector. These two points served as the starting points for a local geometry optimization that led us in the two closest local minima with energies E^a and E^b. In this way we have generated a set of pairs of local minima together with the saddle points that connect them. Fig.2.1 and Fig.2.2 show a scatter plot of $\Delta G = G_i^b - G_i^a$ versus the activation energy $\epsilon_a = G_i^s - G_i^a$ and the red line in the same figure shows a histogram with averages of the $G_i^s - G_i^a$. Each pair of local minima contributed two data points to these plots since one can surmount the barrier by going from the minimum A to minimum B as well as by going from minimum B to minimum A.

The scatter plots in Fig.2.1 and Fig.2.2 show that there is no strict linear correlation between the barrier height ϵ_a and the energy difference ΔG between the two minima. For small barrier heights one can find both high energy and low energy minima behind the barrier. However, the BEP principle holds as a negation. If one goes over high barriers it is extremely unlikely that one will end up in a low energy minimum. The better correlation for large activation energies is simply due to the fact that ΔG can not become larger than ϵ_a. On the other hand, the red line in Fig.2.1 and Fig.2.2 shows that there is a good linear relation if one averages over ΔG. Good linear Bell-Evans-Polanyi relations have been found in calculations of dissociative chemisorption of various molecules[45].

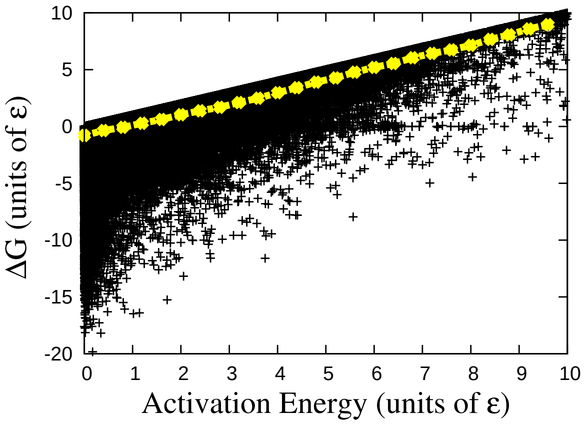


Figure 2.1: The relation between the activation energy $G^s - G^a$ and the reaction energy $G_i^b - G_i^a$ for more than 130000 saddle points in a Lennard Jones cluster of 55 atoms. All the energies plotted here are free energies at T = 0, i.e. just energies. The red line is the same data but averaged within 25 bins along the x axis.

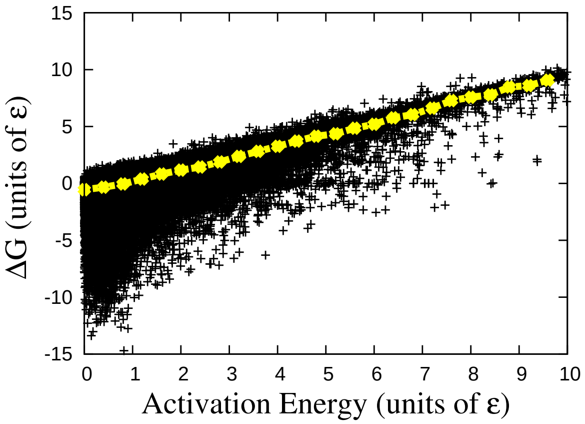


Figure 2.2: Same as Fig. 2.1 however at a temperature (T = 30K) which is below the melting point (50 K) of this weakly bound system[3]. The entropy contribution was calculated in the harmonic approximation from the vibrational frequencies [40]. The figure shows that the free energy has essentially the same behaviour as the energy. The fact that some points are above the diagonal shows that some shallow minima of the potential energy surface are not any more minima of the free energy surface. In principle these points should be eliminated, but we left them in the Figure since they indicate the size of the entropic corrections.

Kinetic rate theory gives the rate constant for a reaction as

$$k = \frac{k_B T}{h} exp(-\epsilon_a/(k_B T)) = \frac{k_B T}{h} \frac{Q_s}{Q_a} exp(-(E^s - E^a)/(k_B T)) \quad , \tag{2.2}$$

where E_a and E_s are the energies of the initial minimum and of the saddle point and Q_s and Q_a are the partition functions corresponding to the saddle point and the initial minimum respectively. Combining this formula with the linear BEP relationship of Eq. 2.1 gives a formula where the speed of the reaction depends only on the energy of the final minimum E^b relative to the initial minimum E^a.

$$k = \frac{k_B T}{h} \frac{Q_s}{Q_a} exp(-(k_1 + k_2(E^b - E^a))/(k_B T)) \tag{2.3}$$

On the macroscopic level a chemical reaction proceeds along a molecular dynamics trajectory. Its energy is determined by the temperature T. The above formula reflects therefore our MDBEP principle. At low temperature one will rarely find MD trajectories that cross into high energy local minima E^b. This statement may sound similar to the well known fact that an ergodic system obeys the Boltzmann distribution and will be therefore preferentially found in low energy regions. Our statement is however not on thermodynamic equilibrium distributions but on the dynamics of the system. The derivation of the above formula (Eq. 2.3) has several weak points. As we have seen before the BEP principle for the energies (Fig. 2.1) holds only on average for similar processes. The rate equation (Eq. 2.2) is itself derived using several approximations. In particular it only holds for trajectories which cross the dividing surface close to the transition state and it is thus not valid for very high energy MD trajectories. Up to now we have also neglected the dependence of the partition function Q_s at the saddle point on the temperature. Q_s is a measure of the size of the dividing surface that is accessible at a certain temperature. The area of this surface increases as the energy of the MD trajectory relative to the saddle point increases. Hence the crossing area is larger for energetically lower saddle points and this effect increases thus the preference of MD trajectories for crossings into the basins of attraction of low energy minima. In addition to this dependence of Q_s on the kinetic energy of the MD trajectory, i.e. on the temperature we have also empirically found a dependence of Q_s on the height of the saddle point. The positive curvatures of the potential energy surface near low energy saddle points is typically larger and so their entropy associated to Q_s becomes smaller (Fig. 2.3). This decreases the preference of MD trajectories for crossings into low energy basins.

Because of all the uncertainties listed above, we will now present numerical experiments to verify the MDBEP principle. In all these experiments the kinetic energy of the MD trajectories was considerably larger than the minimum energy required to be able to overcome the transition states. Fig.2.4 shows the results of the first numerical experiment. For a large number of MD trajectories that start with random directions but fixed kinetic energy E_{kin} from a certain minimum with energy E_a we have recorded how many times this trajectory reaches the basin of attraction of neighboring minima with energy E_b. To

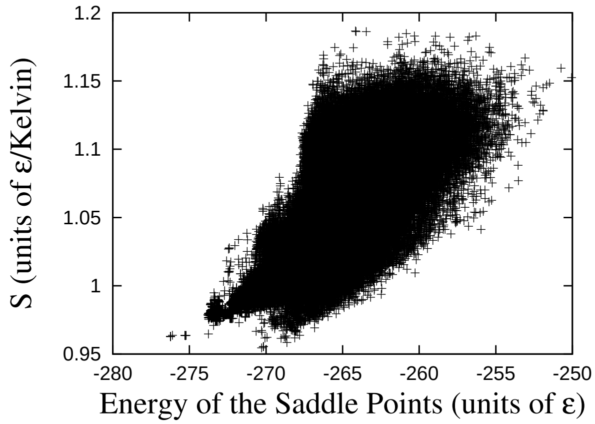


Figure 2.3: Entropy vs height of the saddle point (in energy unit, ϵ). As in Fig 2.2 the entropy was calculated in the harmonic approximation

check whether the MD trajectory has crossed into another basin of attraction steepest descent geometry optimizations were started after every 20 MD steps. Once the crossing occurred the MD run was stopped.

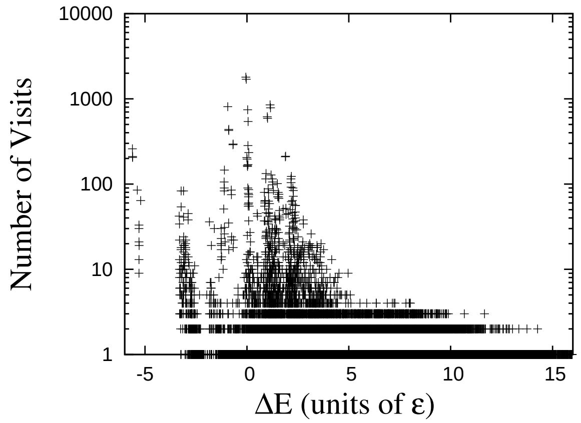


Figure 2.4: The number of visits as a function of $E_b - E_a$ for a MD trajectory with a kinetic energy of 4.0ϵ per atom.

In Fig.2.4 we then plot the number of visits as a function of $E_b - E_a$. We see that it is orders of magnitude more likely that the MD trajectory crosses into low energy basins than into high energy basins. By varying the kinetic energy of the trajectory we can tune the strength of the preference for low energy minima. Low energy trajectories have a much stronger preference for low energy minima than high energy trajectories as shown in Fig 2.5. We will denote this correlation as the MDBEP principle: low energy MD trajectories are more likely to lead into the basin of attraction of a low energy local minimum than high energy trajectories. The activation energy of the original BEP principle has thus been replaced by the energy of the trajectory. As can be seen from Fig.2.1 and Fig.2.4, both the traditional BEP principle and our MDBEP principle are only valid in an average sense. As we will see this validity in the average sense is sufficient in the context of global optimization.

Methods for global geometry optimization are an active area of research, as can be decided from the large number of publications in this field. A basic problem in this context is to construct moves that on the one hand rapidly lead downward in energy and on the other hand avoid trapping [50, 51, 52] in a local minimum that is not the global minimum. We will exemplify this issue in the context of the minima hopping method (MHM)[46, 47]. In the MHM the system moves from one local minimum to another by a combination of MD and local geometry optimizations. With the MD part one jumps from one minimum into the basin of attraction of another minimum. The subsequent local geometry optimization part brings us then into the local minimum of this basin of attraction. From the MDBEP principle we expect that low energy MD trajectories are the most efficient for global optimization. Fig.2.6 and Fig.2.7 show that there is indeed a very strong correlation between the energy of the MD trajectory and the number of minima that are visited before the global minimum is found. The data for Fig.2.6, Fig.2.7, Fig.2.8, Fig.2.9 and the figure Fig. 2.10 were obtained by performing MHM runs that are stopped once the global minimum is found for different but fixed kinetic energies E_{kin} (i.e. $\beta_1 = \beta_2 = \beta_3 = 1$ using the notation of ref.[46]) in a reasonably chosen energy interval. Subsequently we plot the values of E_{kin} versus the number of local minima that were visited before the global minimum was found. The potential energy of the local minimum from which the MD trajectory starts is set to zero. In this way the kinetic energy is the total energy of the MD trajectory and by energetic reasons it can not cross barriers higher than E_{kin} relative the starting minimum. Only new and accepted local minima are counted. In order to achieve better statistics we perform for each fixed E_{kin} 100 MHM runs (for Fig.2.6 the average is taken over 1000 runs), and we take for the plots the averaged number of visited local minima.

The Lennard Jones 55 cluster whose behaviour is shown in Fig.2.6 is a system for which it is very easy to find the global minimum since it has a one funnel structure. Other Lennard Jones clusters such as the 38 atom cluster whose behaviour is shown in Fig.2.7 have two or more funnels[3]. In this case low kinetic energy MD trajectories will rapidly lead into a funnel which is not necessarily the funnel containing the global minimum. Once the system is trapped in a wrong funnel a sufficiently large kinetic energy is evidently required to escape from it. Fig.2.7 however shows that also in this case the efficiency of the global optimization is mainly determined by how rapidly the bottom of a funnel is reached and high energy trajectories are thus less efficient than low energy trajectories even though they can more easily escape from any wrong funnel.

Even for one funnel structures there is of course a lower limit to the kinetic energy. Once it is too low no barriers can any more be overcome and the system gets trapped. One has thus to reconcile two opposite requirements on the kinetic energy of the MD trajectories.

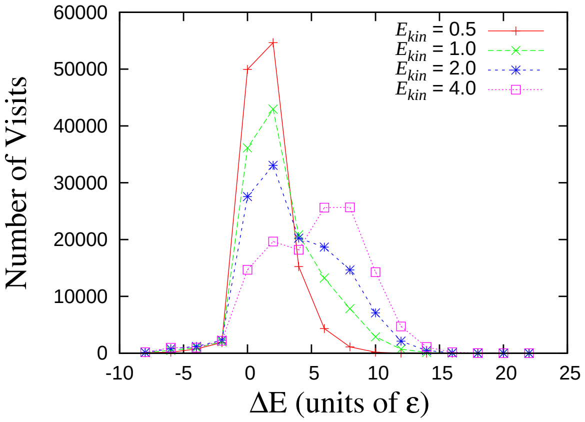


Figure 2.5: The number of visits as a function of $E_b - E_a$ summed over energy bins of length 2 for 4 MD trajectories with different kinetic energies. The curve for an energy of 4.0 ϵ represents the same data as the scatter plot in Fig 2.4.

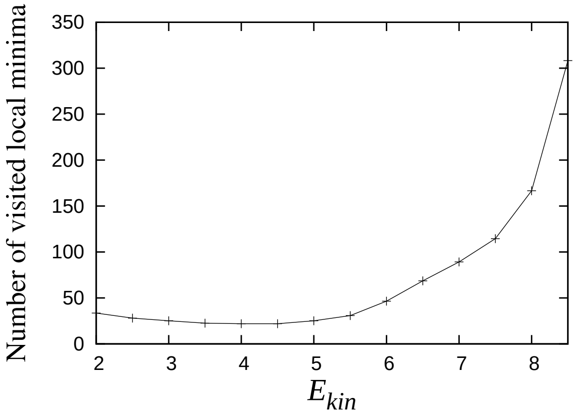


Figure 2.6: The MDBEP principle for the Lennard-Jones cluster of 55 atoms.

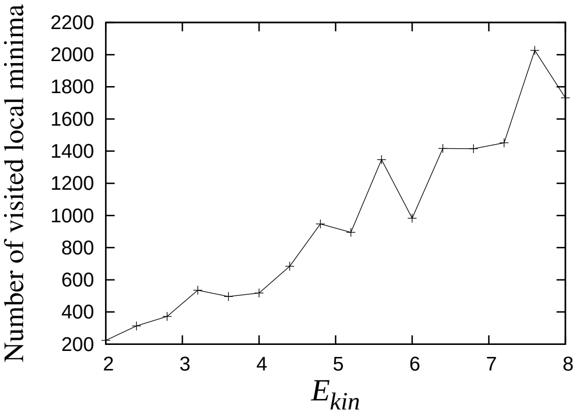


Figure 2.7: The MDBEP principle for the Lennard-Jones cluster of 38 atoms.

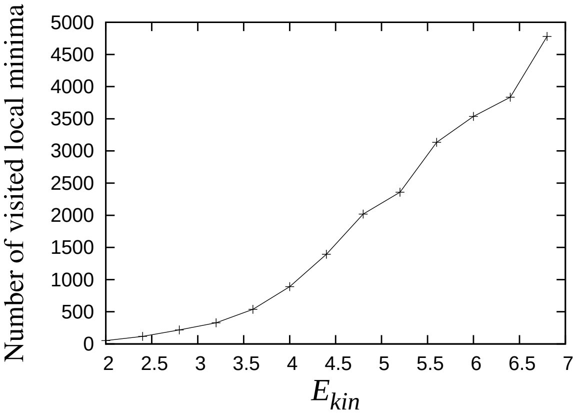


Figure 2.8: The MDBEP principle for the Morse cluster cluster of 38 atoms with $\rho = 6.0$

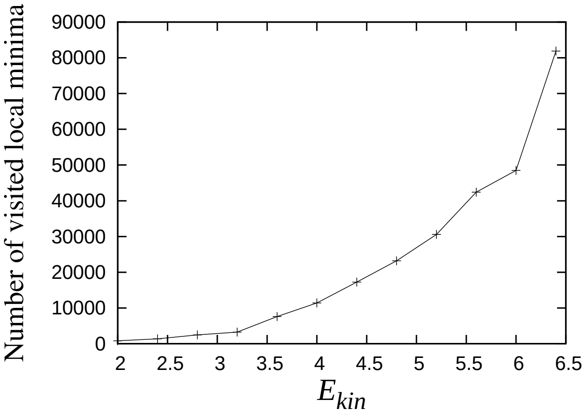


Figure 2.9: The MDBEP principle for the Morse cluster of 38 atoms with $\rho = 10.0$

This is done in a very efficient way in the minima hopping method. If the system goes down in one funnel it explores new local minima and the kinetic energy of the trajectories used to hop from one minimum to another one is reduced. Once the system gets trapped the kinetic energy is increased through a feed back mechanism and the system can escape from any funnel. Minima hopping keeps a history list of all the minima that were previously visited and the feed back is activated if old minima are revisited. Since escapes from a funnel occur seldom one can achieve in the minima hopping method very low average energies for the MD trajectories without being trapped.

Fig. 2.8 and Fig. 2.9 present our results for Morse clusters of 38 atoms with $\rho = 6.0$ and $\rho = 10.0$. Large values of ρ lead to a interaction that varies over shorter length scales. As a consequence the potential energy surface becomes more rugged and has significantly more local minima. As a consequence considerably more minima are visited before the global minimum is found. The global optimization is however also in this case more efficient for low energy trajectories which implies that the MDBEP principle is well observed for very rugged potential energy surfaces.

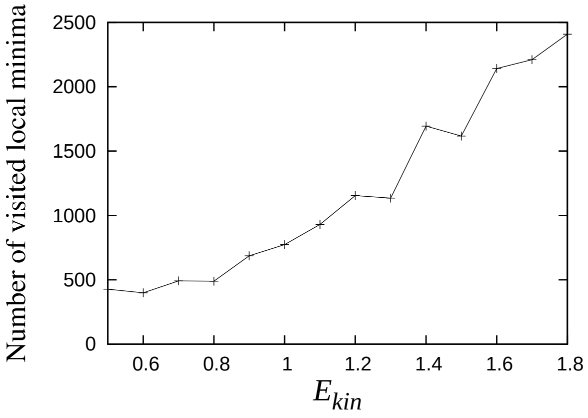


Figure 2.10: The MDBEP principle for the Lenosky tight binding cluster of 20 atoms.

Fig.2.10 presents our results for the Si_{20} cluster [48] within the Lenosky tight binding scheme [49]. In contrast to the Lennard Jones and Morse potentials the silicon tight binding scheme has much more complicated interactions that depend not only on the distance between atoms but also on the quantities like the bond angles. Tight binding schemes are the simplest way to treat solid state systems at a quantum mechanical level. The Lenosky tight binding scheme gave a very good agreement with the DFT energies[47] and can be considered as a reliable approximation to a precise density functional treatment of silicon clusters. The fact that low energy trajectories lead again faster into the global minimum indicates that the MDBEP principle is also valid for realistic interactions and in particular for quantum mechanical interactions.

The fact that for small values of E_{kin} the global minimum is found after having visited only a small number of local minima does not imply that the computational time in the MHM is continuously decreasing with smaller values of E_{kin}. If E_{kin} is getting too small the system has to make a huge number of attempts before succeeding to escape from the basin of attraction of the current minimum and this will actually lead to an increase in the computer time (Fig.2.11). For this reason it is also in practice virtually impossible to explore the behaviour of trajectories with lower energy than those shown in Figs.2.6, 2.7, 2.8, 2.9 and 2.10. Fig. 2.6 shows however that the minimum of the CPU time is reached when the number of minima visited becomes small. The BEP principle is thus not only of conceptual interest but can in practice also help to save CPU time.

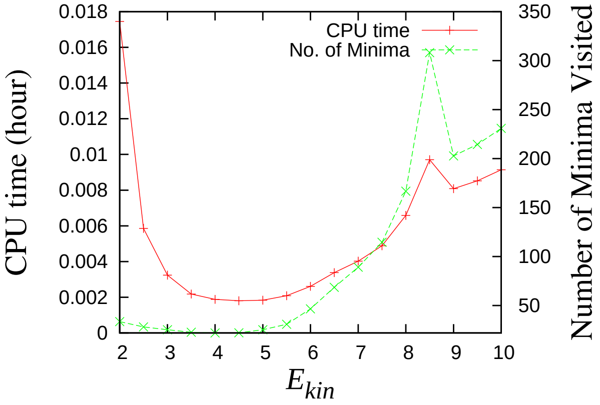


Figure 2.11: The average CPU time (left Y-axis) along with the average number of distinct local minima visited (right Y-axis) before reaching the global minimum for the Lennard-Jones cluster of 55 atoms are plotted against the Kinetic energy of the MD trajectory per atom (E_{kin}).

In practice, the short computation time can be obtained by giving the MD trajectories initial velocities that have large components in the subspace of low curvature of the Hessian matrix. Due to the fact that low energy saddle points often lie at the end of low-curvature modes[53, 54, 55] one can in this way even with low energy trajectories very rapidly escape from the present minimum. A similar gain in efficiency was found in the context of global optimization using random moves if those moves were biased in the direction of the low curvature modes [56]. In summary, we have shown that the BEP principle can be extended to MD trajectories with high energies which cross from one basin of attraction into another one far from the transition state. We call this extended principle MDBEP principle. It says that MD trajectories with lower energy are more likely to lead into basins of attraction of low energy configurations than very high energy trajectories. In the context of global optimization this principle can be used to improve the efficiency of existing MD based methods by tuning the energy of the MD trajectories.

2.2 Conformational correspondence : Forcefield vs Density Functional methods

The aim here is to study the energetic and structural correspondence on a few sets of conformations between a biomolecular forcefield and Density Functional theory. Biomolecular landscapes are often described as "rugged". That would mean the presence of many funnels, traps etc. To describe it even more microscopically, one would expect many sets of local minima separated from each other with considerably high barriers - to overcome which the system has to cross the lifetime of each of those metastable conformations. A solvated protein is obviously expected to have a larger amount of those conformations than while *in vacuo*. In gas phase one can get rid of the large number of degrees of freedom of the solvating system and thus the landscape is simpler - the description of the landscape by a forcefield should thus be more accurate. So in an ideal case one should expect a one-to-one correspondence between each of the local minimum in a present in the landscape corresponding to a biomolecular all-atom forcefield and an energy landscape corresponding to the first-principle description of a protein. As in the forcefields the description of the bonded interactions are in the form of a harmonic potentials and the parameters are mostly derived from spectroscopic experiments their accuracy is much more than the non-bonded interactions. The non-bonded interactions are prone to be erroneous. The short range part of the non-bonded interactions are very important in deciding the exact packing and compactness of a protein.

The first-principle method used here is Density functional theory(DFT). It has become the standard computational chemistry method. While using DFT one encounters the problem of choosing a an appropriate exchange-correlation functional. Based on the performance in many energetic assessments related to small molecules, B3LYP[60, 61] is the most widely used exchange correlation functional. Although it reproduces the geometries of smaller and larger molecules very well but it can fail in describing the energies of van der Waals molecules, hydrogen-bonded systems, reaction barrier heights and larger molecules [62, 63, 64, 65, 66]. PBE [59] Although the LDA[58] approximation generally strongly overestimates weak interaction energies[67, 68, 69] it has been reported to be working better than many other relatively more sophisticated functional [66]. Apart from LDA and B3LYP functional we have also used a pure GGA functional - PBE [59] and a dispersion-corrected LDA.

This study has involved two protein systems. We are defining "system A" to be a peptide of $AcAla_{14}LysGly_{3}Ala_{14}Lys + 2H^{+}$ and "system B" to be a peptide of sequence $AcPheAla_{10}Lys + H^{+}$.

Methodology

Here, we used the OPLS all-atom force field (OPLS-AA) [102] for the potential energy of the biomolecular system as implemented in the DYNAMO modeling library [103, 104]. We conducted short molecular dynamics based moves followed by local optimizations to search the neighbourhood of a preselected minimum. This starting point minimum could have arbitrary. From several thousand neighbouring local minima we calculated the root-mean-squared-deviation (RMSD) of all the structures with respect to the starting point minimum. Then we chose the closest few of the local minima for the ab-initio geometry optimization. For once we took the 10 lowest energy minimum from a global optimization run using minima hopping algorithm (described in chapter 4) on the system B ($AcPheAla_{10}Lys{+}H^{+}$) For the ab-initio geometry optimization we gave used - (i) LDA functional implemented in BigDFT[57] package using daubechies basis sets and (ii) the level B3LYP/6-31G* implemented in Gaussian package[96]. The geometry optimizations were done upto a few hundred steps for all the conformations studied.

Results

We have conducted three sets of calculations.

(RUN 1) Ab-inito geometry optimization of structurally very close 25 conformations of system A and reoptimizing them using the forcefield.

(RUN 2) Ab-inito geometry optimization of structurally very close 24 conformations of system B and reoptimizing them using the forcefield.

(RUN 3) Ab-inito geometry optimization of energetically lowest 10 conformations of system B.

The "RUN 1" and "RUN 2" are for selecting the minima lysing in the neighbourhood of each other. For "RUN 3" we tried to chose the conformation from the energetic point of view.

RUN 1

The top 25 conformational minima of system A have been picked up and they were optimized in DFT-LDA scheme up to $\sim$ 200 steps of geometry optimization. The DFT-optimized conformation have been re-optimized in the OPLS forcefield. The table 2.1 shows the energy of the starting OPLS minimum, the initial DFT-LDA energy, final DFT-LDA energy and the OPLS energy of the re-optimized conformation. The re-optimization was done for identifying the distinct conformations after the DFT based geometry optimizations. From the 25 conformations we could finally have 19 conformations to be stable in DFT-LDA scheme. We pick up a pair of such minima which fall back in one of them. The conformation 22 and 25 are shown in fig. 2.12. They were having rms deviation $0.067\mathring{A}$ and OPLS energy difference $254.24 KJmol^{-1}$ and the conformation no. 22 was found to be unstable in DFT. All the steps of DFT geometry optimization have been saved and their corresponding OPLS energy and different components were calculated. To check which OPLS local minimum the DFT optimization geometry lead to, we optimized all the steps of DFT geometry optimization. All the energies are shown in fig.2.13. Its clear that for the conformation 22, along the DFT geometry optimization path from the start upto a few steps, leads back to the staring OPLS minimum then afterwards it starts leading to a new minimum. So for such an unstable conformation two local minima can coalesce while conducting a point-to-point landscape transformation from OPLS to DFT.

For the conformation 22 we conducted another two sets of geomtry optimization in dft using a PBE functional scheme and a dispersion-corrected DFT scheme. The conformation was not stable in any of them. The OPLS energy and different components for each step of DFT geometry optimization is shown in fig 2.14.

RUN 2

The top 24 conformational minima of system B have been picked up and they were optimized in B3LYP/6-31G* DFT scheme up to ~ 100 steps of geometry optimization. The DFT-optimized conformation have been re-optimized in the OPLS forcefield. The table 2.2 shows the energy of the starting OPLS minimum, the initial DFT-B3LYP energy, final DFT-B3LYP energy and the OPLS energy of the re-optimized conformation. The re-optimization was done for identifying the distinct conformations after the DFT based geometry optimizations. From the 24 conformations we could finally have 17 conformations to be stable in DFT-B3LYP scheme. We pick up a pair of such minima which fall back in one of them. The conformation 10 and 14 are shown in fig. 2.16. They were having rms deviation $0.1993\mathring{A}$ and OPLS energy difference $254.33 KJmol^{-1}$ and the conformation no. 10 was found to be unstable in DFT. All the steps of DFT geometry optimization have been saved and their corresponding OPLS energy and different components were calculated. To check which OPLS local minimum the DFT optimization geometry lead to, we optimized all the steps of DFT geometry optimization. All the energies are shown in fig.2.17. Its clear that for the conformation 10, along the DFT geometry optimization path from the start upto a few steps, leads back to the staring OPLS minimum then afterwards it starts leading to a new minimum. So for such an unstable conformation two local minima can coalesce while conducting a point-to-point landscape transformation from OPLS to DFT.

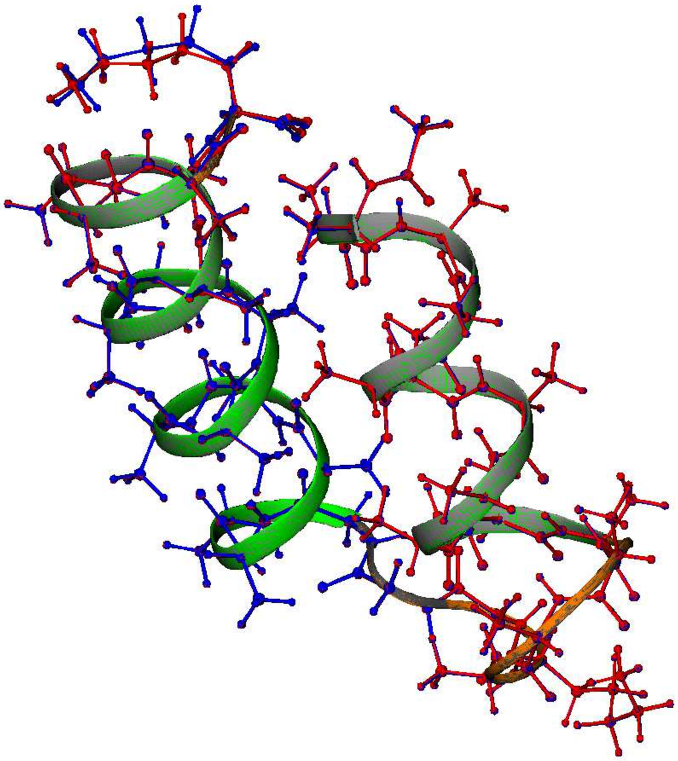


Figure 2.12: Two closely lying minima of OPLS forcefield(conformation number 22 (blue) and 25(red) from the table2.1) having rms deviation $0.067\AA$ and energy difference $254.24KJmol^{-1}$, which will eventually coalesce in the ab-initio geometry optimization(DFT-LDA).

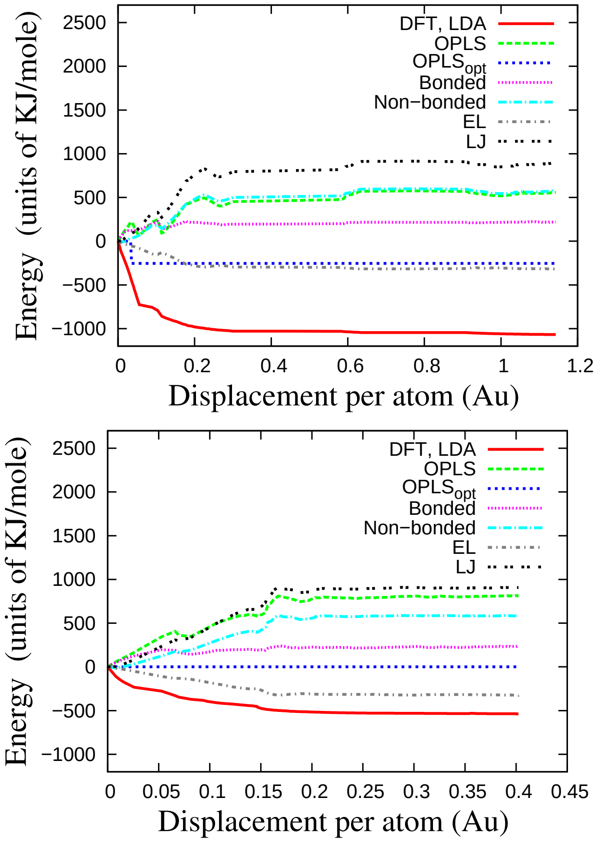


Figure 2.13: Conformation 22 and Conformation 25 of the table 2.1 : the different components of the forcefield energy along the path of geometry optimization in DFT.

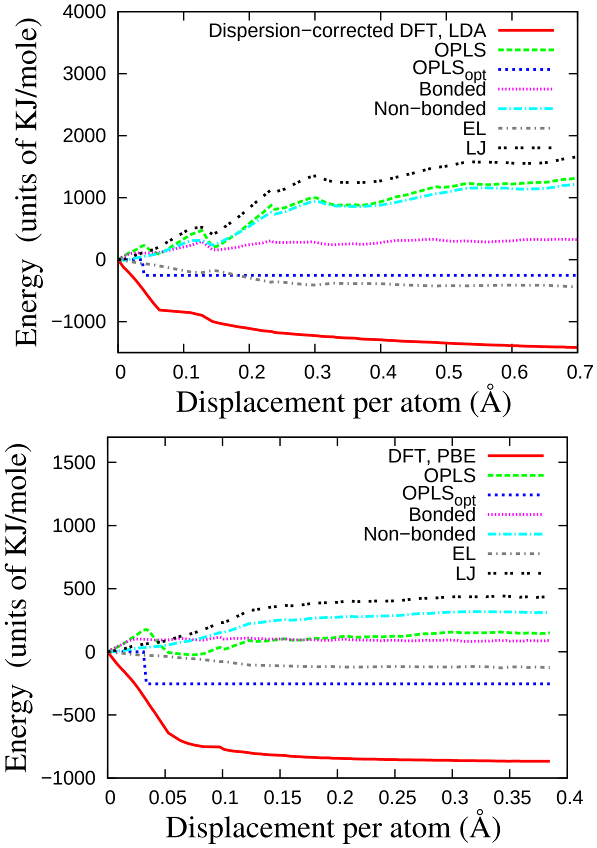


Figure 2.14: The different components of the forcefield energy along the path of geometry optimization in DFT with PBE functional and Dispersion-corrected DFT with LDA functional for the Conformation no. 22 from the table 2.1.

Conformation number	OPLS Energy Starting Minimum	DFT(LDA) Energy before optimization	DFT(LDA) Energy after optimization	OPLS Energy after reoptimization of the DFT minimum
1	-6.413	-17.171	-1105.811	-284.512
2	-263.378	-557.444	-1075.632	-263.378
3	-18.701	-6.023	-1076.116	-256.579
4	-209.558	-494.684	-1009.179	-209.558
5	-188.403	-467.125	-976.799	-188.403
6	0.000	0.000	-1081.420	-284.512
7	-6.813	-15.898	-1081.259	-251.413
8	-22.162	-29.983	-1101.629	-284.512
9	-255.699	-560.022	-1082.073	-255.699
10	-77.980	-163.817	-1074.519	-249.110
11	-267.541	-565.333	-1088.707	-267.541
12	-249.110	-555.110	-1079.144	-249.110
13	-258.878	-555.261	-1074.907	-258.878
14	-246.417	-539.674	-1052.935	-246.417
15	-234.265	-544.534	-1067.457	-234.265
16	-257.412	-570.584	-1091.109	-257.412
17	-283.230	-587.106	-1111.180	-283.230
18	-282.796	-582.074	-1103.299	-282.796
19	-0.170	-5.944	-1104.365	-284.512
20	-223.615	-521.935	-1055.992	-223.615
21	-251.413	-556.737	-1086.905	-251.413
22	-18.248	-40.730	-1104.374	-272.482
23	-229.648	-534.582	-1062.300	-229.648
24	-182.725	-480.408	-995.600	-182.726
25	-272.482	-574.967	-1103.380	-272.482

Table 2.1: Summary of the optimization-reoptimization test on the 25 conformational minima of OPLS forcefield (ref. fig.2.15)

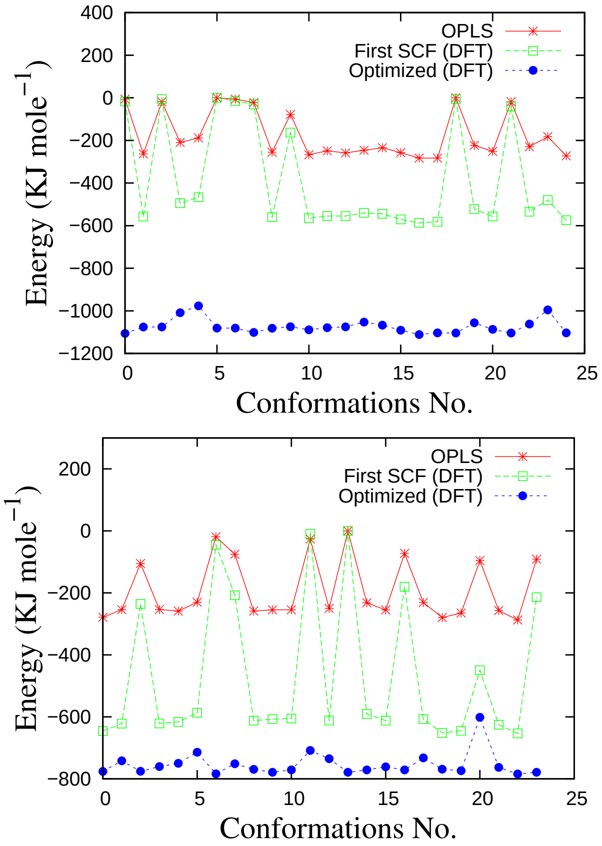


Figure 2.15: OPLS energy of the starting minima, energy in a single-point calculation and energy after geometry optimization in DFT of both RUN 1 on system A and RUN 2 on system B shown all together.

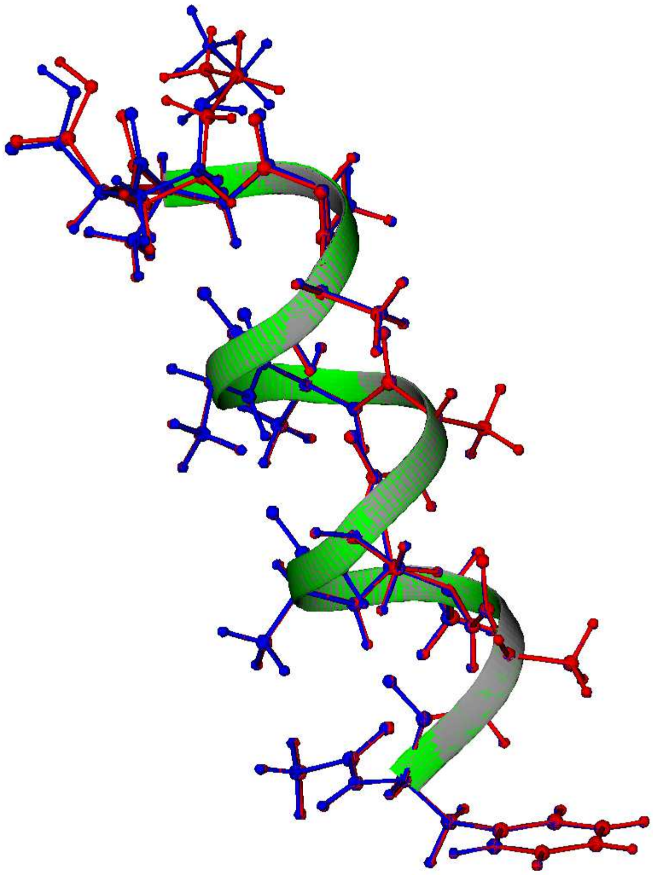


Figure 2.16: Two closely lying minima (conformation 10 (red) and 14 (blue) from the table2.2) of OPLS forcefield (rms deviation $0.1993\mathring{A}$, energy difference $254.33KJmol^{-1}$), which will eventually coalesce in the high accuracy ab-initio geometry optimization (DFT-B3LYP).

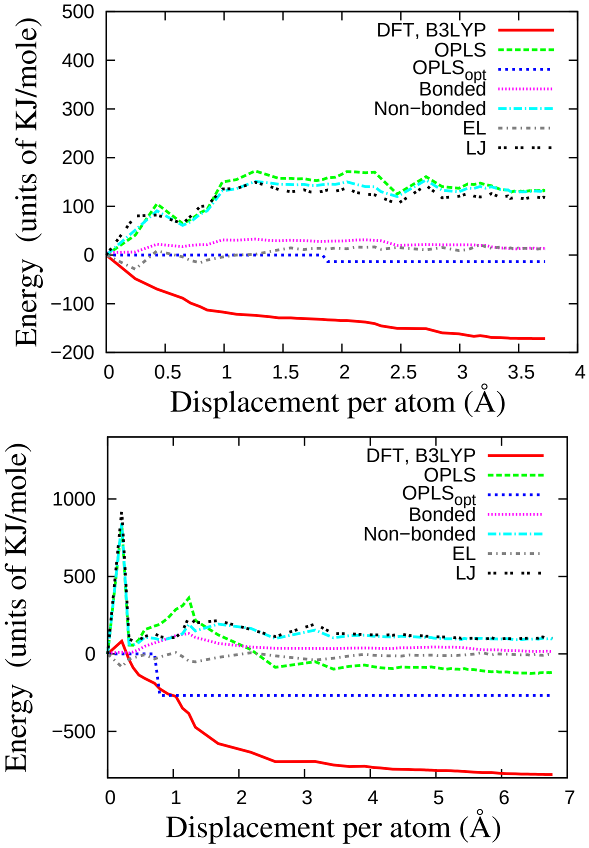


Figure 2.17: Conformation 10 and Conformation 14 shown in fig. 2.16 : the different components of the forcefield energy along the path of geometry optimization in DFT.

Conformation number	OPLS Energy Starting Minimum	DFT(B3LYP) Energy before optimization	DFT(B3LYP) Energy after optimization	OPLS Energy after reoptimization of the DFT minimum
1	-278.382	-645.347	-776.179	-278.382
2	-254.041	-621.469	-741.758	-254.041
3	-105.170	-235.567	-775.768	-278.382
4	-253.784	-621.371	-760.178	-258.636
5	-258.636	-616.379	-750.006	-258.636
6	-230.144	-586.687	-714.390	-230.144
7	-18.802	-45.156	-783.943	-287.331
8	-75.459	-207.778	-751.369	-253.784
9	-258.793	-611.955	-769.161	-266.901
10	-254.339	-607.203	-778.792	-267.835
11	-254.331	-605.538	-771.112	-254.331
12	-25.982	-9.274	-708.224	-278.416
13	-250.115	-611.834	-734.758	-252.564
14	0.000	0.000	-778.780	-267.835
15	-231.911	-590.410	-771.113	-254.331
16	-254.926	-612.015	-761.035	-254.926
17	-72.671	-180.076	-771.155	-254.331
18	-230.893	-607.202	-732.339	-230.893
19	-279.365	-651.830	-768.721	-279.365
20	-264.973	-644.543	-774.010	-264.973
21	-95.780	-449.410	-601.234	-95.780
22	-256.897	-625.240	-762.895	-256.897
23	-287.331	-653.404	-784.042	-287.331
24	-91.007	-213.942	-778.679	-267.835

Table 2.2: Summary of the optimization-reoptimization test on the 24 conformational minima of OPLS forcefield on system B. (ref. fig.2.15)

In the fig. 2.15 we showed that the OPLS energy of the starting minima, energy in a single-point calculation and energy after geometry optimization in DFT of both RUN 1 on system A and RUN 2 on system B shown all together. The agreement on the structural ranking between a forcefield and DFT is apparently very poor. The lines of the relative energy of the conformations (in the fig. 2.15 became smoother upon DFT geometry optimization.

RUN 3

Here from the lowest 10 conformations found by a conformational search, 8 were found to be stable in DFT-B3LYP scheme. We are showing one set the conformations which were found to fall back in the same minimum in DFT-B3LYP scheme. The conformation 2-nd lowest and 6-th lowest are shown in fig.2.18.

All the steps of DFT geometry optimization have been saved and their corresponding OPLS energy and different components were calculated. To check which OPLS local minimum the DFT optimization geometry lead to, we optimized all the steps of DFT geometry optimization. All the energies are shown in fig.2.19. Its clear that for the conformation 6, along the DFT geometry optimization path from the start upto a few steps, leads back to the staring OPLS minimum then afterwards it starts leading to a new minimum.

Non-bonded interaction

If we look at the fig. 2.13, fig. 2.17 and fig. 2.19 we can see that the forcefield minimum has always been pulled away in the DFT minimization by providing some amount of non-bonded energy. The repulsive lennard-jonnes of the forcefield is showing a very high energy for all the DFT minima. The overall change in the total OPLS energy upon a DFT geometry optimization is thus positive. The attractive coulombic interaction tries to compensate but only up to a certain extent. There is a clear disagreement in the non-bonded interactions between a forcefield scheme and an ab-initio scheme. The bonded interactions dont change very much in the structural change.

Packing and compactness The effect of a hard van der Waal repulsion and the compensating electrostatic interactions with a crude approximation on the point charges parametrized on small molecules should be seen in the overall packing of a folded conformation. In the fig. 2.20 we see that upon a DFT-LDA geometry optimization all the conformations attain a smaller radius of gyration and so, get an overall compactness and the atoms goes closer to each other within the repulsive/forbidden zone of the classical form of Lennard-jones repulsive part. But a different picture comes when we wee the

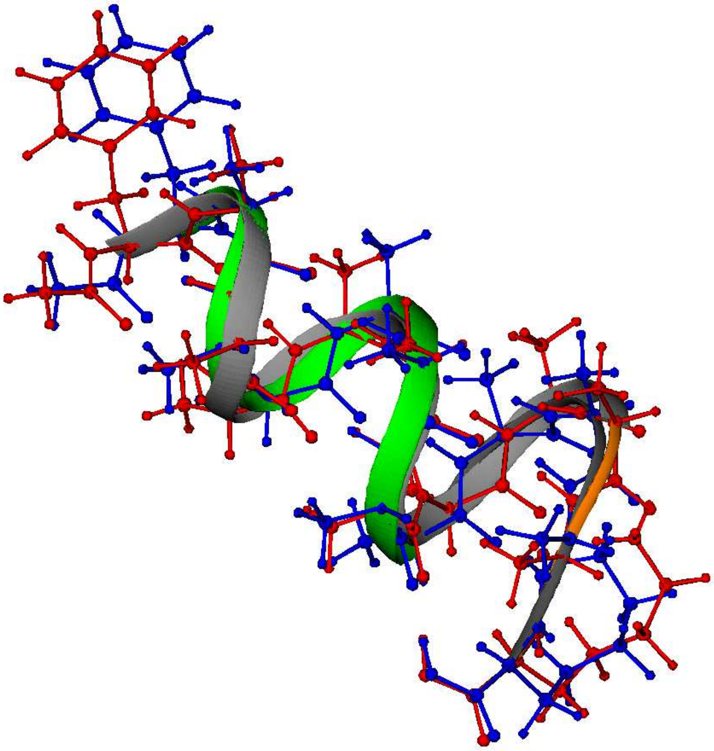


Figure 2.18: Two energetically closely lying minima of OPLS forcefield (conformation 2 (blue) and conformation 6 (red) ; rms deviation $1.08\AA$, energy difference $6.85KJmol^{-1}$), which will eventually coalesce in the high accuracy ab-initio geometry optimization (DFT-B3LYP).

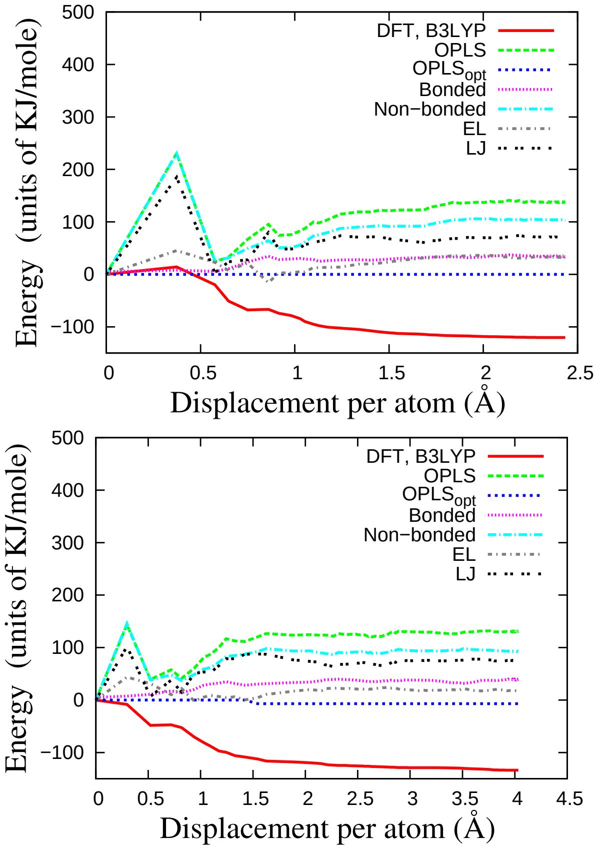


Figure 2.19: Conformation 2 and Conformation 6 shown in fig. 2.18 : the different components of the forcefield energy along the path of geometry optimization in DFT.

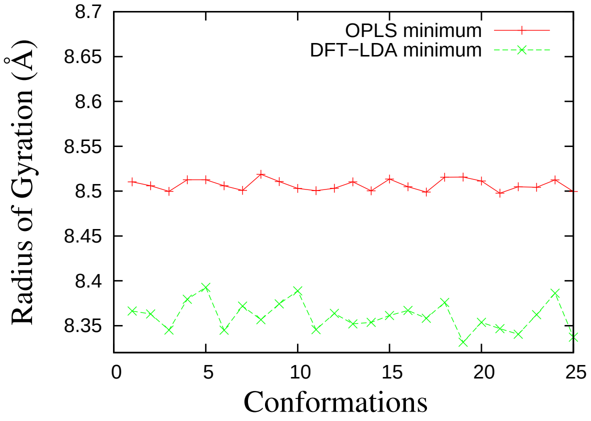


Figure 2.20: Radius of gyration of each of the 25 conformations listed in table 2.1 and their corresponding minima in DFT-LDA.

optimized structures of DFT-B3LYP. The DFT-B3LYP optimized structures have larger radii of gyration. It has been pointed out at beginning that B3LYP has always failed to portray a realistic van der Waal interaction [62, 63, 64, 65, 66] so although it is a higher level density functional and so reliable in producing the conformational density of states its energetic ranking may not be correct in such a case.

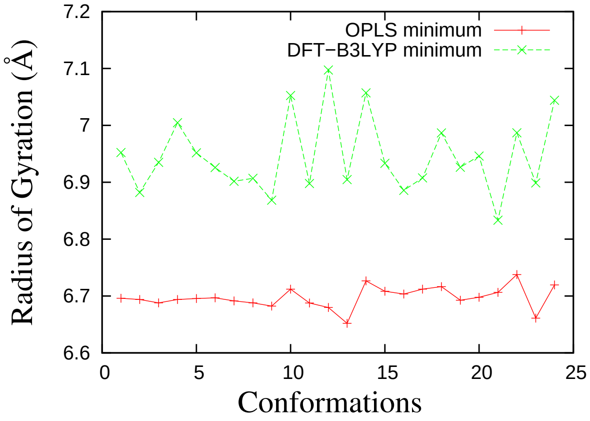


Figure 2.21: Radius of gyration of each of the 24 conformations listed in table 2.2 and their corresponding minima in DFT-B3LYP.

CONCLUSIONS

We compared energy landscapes of proteins in several levels of accuracy, e.g that of forcefields, LDA, B3LYP and PBE density functionals. We found that the forcefield can give rise to many fake minima which can be both structurally close or energetically close to the real minima. The non-bonded forces of a biomolecular forcefield should be corrected to produce better agreement with the higher level theory. The Lennard-Jones part of the forcefield is found to be harder than a density functional description. With a unrealistically hard Lennard-jones there is a low chance to find the real compactness and packing of the real proteins.

Chapter 3

Methodological developments

This chapter deals with some of the algorithms which were implemented or developed in course of the global optimization studies.

3.1 Implementation of Minima Hopping Algorithm

We use the Minima Hopping Algorithm(MHOP) [124] for structure prediction. Minima hopping [124] is an efficient algorithm for finding the global minimum on the potential energy surface of a polyatomic system. The complete description of the algorithm may be found in reference [124]. This has been used previously by for global optimization studies of Lennard-Jones [124] and silicon [125] clusters. It possesses an efficient feedback mechanism that makes use of the search history.

- It uses MD instead of random moves (as in basin-hopping [196]) to jump over reasonably low barriers using the Bell-Evans-Polanyi principle [129]. The MHOP process generates a collection of small MD trajectories from which the search pathway is traceable even though each trajectory can have different initial temperatures and velocity distributions. Although the MHOP search pathway and the folding pathway obtained from a continuous MD simulation may be different, the major funnel transitions should be common to both
- the MHOP process incorporates a learning mechanism from the search history and keeps on changing the temperature of the MD and the acceptance/rejection criterion systematically to avoid revisiting configurations. The length of the MD part of the algorithm can be adjusted to adapt better to the nature of the landscape, e.g. for funnel-like landscapes longer MD may be more useful than for more rugged landscapes

- the algorithm is easily parallelized to accelerate sampling and it is straightforward to ensure that each processor has access to the search history of all processors.

Algorithm - flowchart

Here a flowchart (see fig. 3.1) has been drawn to illustrate the algorithm briefly. The stopping criteria for the loop is not generic, so was not shown.

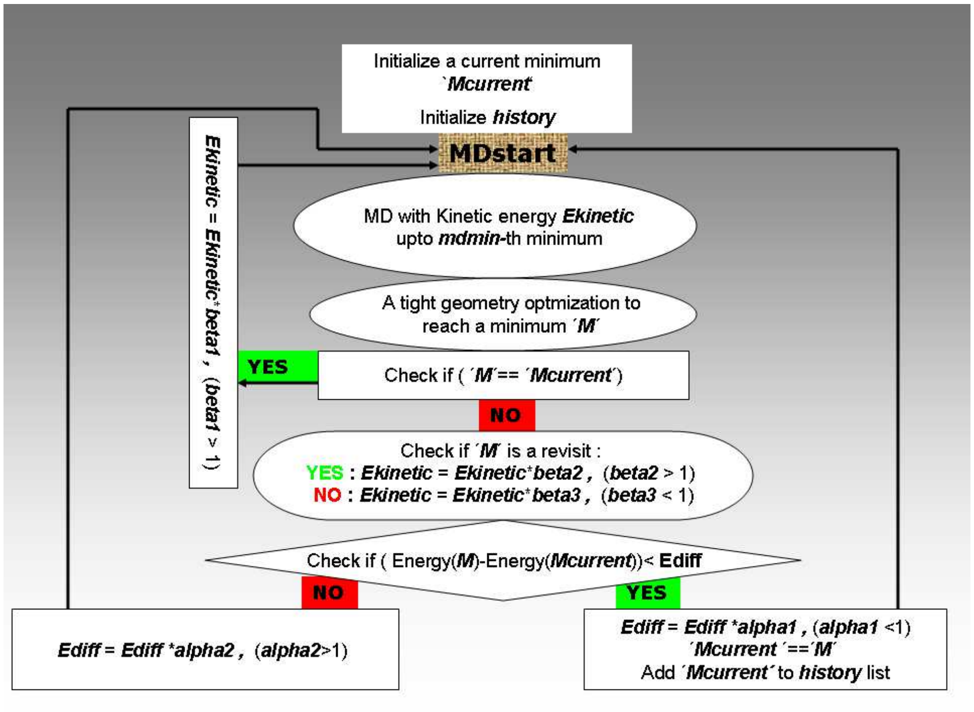

Initialize a current minimum
`Mcurrent`
Initialize history
MDstart
Ekinetic = Ekinetic*beta1 , (beta1 > 1)
MD with Kinetic energy Ekinetic
upto mdmin-th minimum
A tight geometry optmization to
reach a minimum ´M´
YES
Check if (´M´== ´Mcurrent´)
NO
Check if ´M´ is a revisit :
YES : Ekinetic = Ekinetic*beta2 , (beta2 > 1)
NO : Ekinetic = Ekinetic*beta3 , (beta3 < 1)
Check if (Energy(M)-Energy(Mcurrent))< Ediff
NO
YES
Ediff = Ediff *alpha2 , (alpha2>1)
Ediff = Ediff *alpha1 , (alpha1 <1)
´Mcurrent´==´M´
Add ´Mcurrent´ to history list

Figure 3.1: Flowchart of minima hopping algorithm.

3.1.1 DIMERSOFT

MD escape trials in the MH algorithm need an initial velocity distribution which is then rescaled to fit the desired kinetic energy. The velocities are randomly directed for each atom with Gaussian distributed magnitudes. Regardless of the actual distribution chosen it has proved very useful to use *softening*, to choose velocities along low-curvature directions. In this way one can typically find MD trajectories with a relatively small energy that cross rapidly into another basin of attraction. In the original MH method low kinetic energy trajectories could only be obtained by using large values for `mdmin` which results in long trajectories. A direction of low curvature is found using a modified iterative dimer method which only uses gradients, no second derivatives need to be calculated [128]. Starting at a local minimum $\mathbf{x}$ with an escape direction $\hat{\mathbf{N}}$ the method calculates a second point $\mathbf{y} = \mathbf{x} + d\hat{\mathbf{N}}$ at a distance d along the escape direction. The forces are evaluated at $\mathbf{y}$ and the point is moved along a force component $\mathbf{F}^{\perp}$ perpendicular to $\hat{\mathbf{N}}$:

$$\mathbf{F}^{\perp} = \mathbf{F} - (\mathbf{F} \cdot \hat{\mathbf{N}})\hat{\mathbf{N}}$$

$$\mathbf{y}' = \mathbf{y} + \alpha \mathbf{F}^{\perp}$$

$$\hat{\mathbf{N}}' = \frac{\mathbf{y}' - \mathbf{x}}{|\mathbf{y}' - \mathbf{x}|}.$$

After a few steps the iteration is stopped before a locally optimal lowest curvature mode is found. Initial velocities for the MD escape are then chosen along the final escape direction $\hat{\mathbf{N}}$.

If the softening procedure is executed until it converges the performance drops again. It is important not to overdo softening. Always escaping into the *same* soft mode direction of a given minimum reduces the possibilities of different escape directions and therefore weakens the method. A good indicator was the mean kinetic energy during a run. For a few softening iterations the value decreases whereas it starts to increase again at a certain number of softening iterations. We set the iteration count to the value where the mean kinetic energy was minimal. The overall impact of this scheme has been tested in detail in section 4.

3.1.2 Mode Decomposition

This is a scheme for searching along slow vibrational modes. To concentrate searching in slow vibrational modes we did not use simple random initial velocities in the MD simulations and instead took out the component of the initial random velocity which is in the space of all bond stretching and bond bending vibrations. Let $\vec{V_0}$ be the initial $3N$-dimensional random velocity vector, where N is the number of atoms. If there are N_{bonds} bonds and N_{angles} angles in the protein then $N_{\text{ba}} = N_{\text{bonds}} + N_{\text{angles}}$. Let the vectors corresponding to each of the bonds or angles be denoted $\vec{T_l}$, where $l = 1, 2, ..., N_{\text{ba}}$. Now we define the l-th bond vector $\vec{T_l}$ corresponding to the bond between the connected atoms,

i and j or the angle vector $\vec{T_l}$ corresponding to the angles $(i,\, k,\, j)$ as follows :

$$\vec{T_l} \;=\frac{1}{\sqrt{2(r_{ij})^2}}\begin{bmatrix} 0 & \\ \vdots & \text{upto 3(i-1)} \\ 0 & \\ +(x_i - x_j) & \\ +(y_i - y_j) & \\ +(z_i - z_j) & \\ 0 & \\ \vdots & \text{from (3i+1) to 3(j-1)} \\ 0 & \\ -(x_i - x_j) & \\ -(y_i - y_j) & \\ -(z_i - z_j) & \\ 0 & \\ \vdots & \text{from (3j+1) onwards} \\ 0 & \end{bmatrix}$$

where $r_{ij} = \sqrt{((x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2)}$. We write $\sum_{l=1}^{N_{\mathrm{ba}}} C_l \vec{T_m} \cdot \vec{T_l} = \vec{V_0} \cdot \vec{T_m}$. By solving this linear system of equations we get the coefficients C_l, and then the new velocity vector $\vec{V}_{\mathrm{modified}} = \vec{V_0} - \sum_{l=1}^{N_{\mathrm{ba}}} C_l \vec{T_l}$ which is rescaled to match the temperature. This new velocity is a random vector in the subspace free from any vibrational mode related to bond stretching or angle bending.

3.1.3 Enhanced Feedback

In original version of minima hopping algorithm `ekin` is increased by a factor β_2 if the current minimum has already been visited before — regardless of the number of previous visits. An enhanced feedback method uses a value of β_2 depending on the previous visits according to

$$\beta_2 = \beta_2^0 \times (1 + c \log N) \tag{3.1}$$

where β_2^0 is the original value of 1.05 and N the number of previous visits to this minimum. The parameter c has been set to 0.1 after tests on bigger Lennard-Jones clusters and gold systems. This feedback mechanism reacts slightly stronger if the minimum is visited many times. If the system has only one energy funnel this enhanced feedback can even be slightly disadvantageous since it increases the kinetic energy too much and thus weakens the BEP effect of MD. The increased feedback mechanism improves the efficiency however considerably for large systems where the system can be trapped in huge structural funnels. If a cluster has for instance both low energy icosahedral and decahedral structures it takes a very long time for the MH algorithm without enhanced feedback to switch from one structure to the other.

3.2 Preconditioned Steepest Descent

As a geometry optimizer steepest descent converges slower than the other advanced geometry optimizers like conjugate gradient - provided the starting point is near to the quadratic region. In a highly nonquadratic part of the energy landscape it works more efficiently than the others. In most of the efficient implementation of advanced geometry optimizers, SD is used for the first few iterations to bring the force down from the very high force of the starting conformation.

Here we have tested a preconditioner for steepest descent for applying on the protein systems in an all-atom forcefield representation. The all-atom forcefields contain the terms related to the bond stretching, angle bending etc. In the section 3.1.2, a mode decomposer was described which gets a vector free from its component on the subspace composed by all the vectors corresponding to the bond stretching and angle bending motions. We use that mode decomposer to precondition the force in the following way.

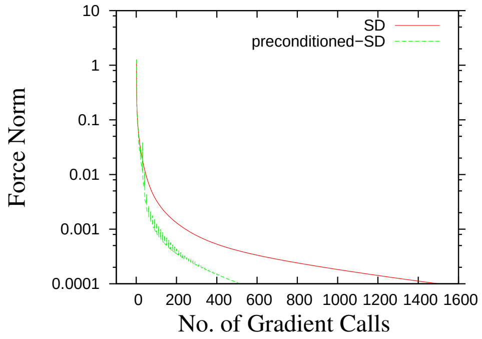


Figure 3.2: SD vs preconditioned SD - the force norm is plotted against the no. of gradient calls.

If $\vec{F}$ is the total force, then by a decomposition described in the section 3.1.2 we get a component of the force which is $\vec{F}_{\text{modified}} = \vec{F} - \sum_{l=1}^{N_{\text{ba}}} C_l \vec{T}_l$. Then the preconditioned force ($\vec{F}_{\text{precon}}$) would be $\vec{F}_{\text{precon}} = \alpha * \vec{F}_{\text{modified}} + (\vec{F} - \vec{F}_{\text{modified}})$, where α is the gain in the step size for the soft mode components. Here we can seperately decompose the components and the angles components. Then we can have the the preconditioned force ($\vec{F}_{\text{precon}}$) to be $\vec{F}_{\text{precon}} = \vec{F}_{\text{bond}} + \alpha_1 * \vec{F}_{\text{angle}} + \alpha_2 * (\vec{F} - \vec{F}_{\text{bond}} - \vec{F}_{\text{angle}})$. The values of α_1 and α_2 were determined using a trial and error method. The implementation of this preconditioned steepest descent was done by working on the combination frequency of the preconditioning step and the values of α_1 and α_2. In the fig.3.2 the performance of the preconditioned-SD is shown applying it on a peptide of sequence $AcAla_8Lys + H^+$(using OPLS forcefield) where we have called the preconditioner once in every 10 steps and the value of α_1 was ≈ 5.0 and α_2 was ≈ 40.

3.3 DIISIP : A new preconditioned-DIIS Geometry optimization scheme

A new geometry optimization scheme is presented. We constructed an iterative preconditioner coupled within the DIIS scheme calling it a Direct Inversion in the iterative subspace with iterative preconditioning algorithm (DIISIP). The preconditioner works on a model Hamiltonian consisting of a simple bond and angle representation for the fast vibrations in the system. The model can be built using *a priori* knowledge of the system curvatures or can be parametrized on-the-fly by an approximate gradient-matching. An efficient hybrid implementation of this algorithm with other geometry optimizers is tested. A number of clusters, biomolecules have been optimized using this scheme.

INTRODUCTION

Local optimization of geometry is a very important methodological aspect of computational chemistry. Essentially it is a search for a stationary point on a potential energy surface (PES). In practice the term geometry optimization is most often used in the context of searching for a local minimum than the stationary points with one or more than one negative eigenvalues. Many of the theoretical studies, specially those involve calculations of transition state, barrier height, heats of reaction, or vibrational spectra require an efficient method for geometry optimization. In the context of global optimization of PES, the local optimization is required most often. Specially for the algorithms like basin hopping, minima hopping etc, an accurate geometry optimization is almost indispensable.But geometry optimization becomes difficult and slow if it involves a huge number of variables due the presence of many thousands of atoms specially for Biomolecules, polymers and nanostructures etc. To circumvent the scaling issue in geometry optimization there are efforts in the community for designing geometry optimizers having a better scaling [70, 71]. Also if the calculation of gradient becomes very time consuming, e.g. for a higher level ab-initio calculation, even the regularly converging geometry optimization can face the time issue, so there any decrease in the computational cost is welcome.

A variety of algorithms for geometry optimization are widely used in computational chemistry [72, 73]. Geometry optimization methods can be categorized into three major groups. Zeroth order methods use only the functions value for optimizing the function - this is used specially if the gradient calculation is numerically prohibitive. First-order methods use just the analytic first derivatives to search for stationary points. Some of the commonly used first-order methods are steepest descent method(SD), conjugate gradient method(CG), direct inversion of the iterative subspace (DIIS) method etc. Second-order methods use both analytic first and second derivatives, assuming a locally quadratic model for the potential energy surface and a NewtonRaphson step ($\nabla\vec{x}= H^{-1}\vec{g}$, where $\vec{x}$ is the coordinate vector, $\vec{g}$ is the gradient, H is the hessian matrix) for the minima search. While second-order optimization schemes need fewer steps to reach convergence than first-order methods[88]. This kind of an approach can quickly become very expensive with increasing system size because the explicit computation of the Hessian scales as $O(N^3)$ to $O(N^4)$, where N is the system size. Quasi-Newton methods are intermediate between the first and second-order approaches. A initial evaluation of the Hessian is done using some inexpensive method. Subsequently, the Hessian is regularly updated using the first derivatives [74, 75, 76, 77, 78, 79, 80, 81]. The quasi-Newton approach is comparable in computational cost to first-order methods. The Quasi-Newton approach L-BFGS algorithm has been observed to be significantly efficient than methods like conjugate gradient[3]. So this kind of a Quasi-Newton approach in terms of speed better than first-order methods, although an efficient implementation of DIIS scheme has been expected to be very effective[95, 3].

In DIIS method, to reduce the number of iterations required to reach convergence, a least-squares minimization scheme is used[89, 90]. At a given iteration, the optimizer constructs a linear combination of approximate error vectors from previous iterations. The coefficients of the linear combination are then determined so as to best approximate, in a least squares sense, the null vector. The coefficients are then used to generate the function variable for the next iteration. It is efficient in both converging the wave function and optimizing the geometry[91, 92, 93, 94, 95, 96, 97].

The idea of preconditioning is brought to improve the condition number of the second-derivative matrix. All the first-order geometry optimizers improve upon incorporating a good preconditioner and using preconditioned gradients in place of the gradients. In the ideal case a preconditioning matrix $A = H^{-1}$, where H is the hessian matrix. Its known that the use of a NewtonRaphson step when the current position is far away from a quadratic region can lead to big step sizes in the wrong direction. It makes optimizer unstable. The stability of a NewtonRaphson geometry optimization is enhanced by controlling the step size using techniques such as rational function optimization (RFO) [82, 83] or the trust radius model [84, 85, 86, 87]. The same way the preconditioned force can at times lead to instability. In the spirit of a trust radius model an adaptive step-size or frequent restarts can help out.

Here we shall describe a preconditioner which would be used in the DIIS method. This preconditioner works as an iterative optimization of the gradient of a second Hamiltonian which is constructed out of a model describing the fast vibrations present in the system. Because here the preconditioning is a multi-step optimization itself, we call it Direct Inversion in the iterative subspace with iterative preconditioning algorithm (DIISIP). It is then applied on a few small clusters interacting via Lennard-Jonnes potential, lenses potential and a small alanine-based peptide. Finally, an efficient hybrid implementation of DIISIP, which is its combination with the SD algorithm has been tested on bigger peptides.

Methodology

It is a preconditioned DIIS method with a variable preconditioning, i.e. the preconditioning operater(P_m) is specific to the iteration number m. Applying this operator we get a preconditioned gradient $\vec{g_m}$ as follows :

$$\widetilde{\vec{g_m}} = P_m\vec{g_m} = P_m\vec{\nabla}\phi(\vec{c_m}), \tag{3.2}$$

where, for a given position vector c_m we have the gradient $\vec{g_m} = \nabla\vec{\phi}\vec{c_m}$. The preconditioned gradient vector $\widetilde{\vec{g_m}}$ is found by successive smoothing steps to reduce the components of the high frequency vibrations. These vibrations present in a molecular system are coming out of the bond stretching or the angle bending motion.

Preconditioner We construct a bonded network by searching all atom pairs (i, j) within a cut-off distance ($\sim$ (r_{eq}) , $r_{eq} \approx$ equilibrium distance or typically the physical bond length between i^{th} and j^{th} atoms). At first we write $\vec{r} = \vec{c_m}$. The corresponding Hamiltonian is as follows :

$$H_{\mathrm{b}} = \sum_{\mathrm{i<j}} K_{ij}(r - r_{\mathrm{eq}})^2, \tag{3.3}$$

$$\vec{g_r} = \vec{\nabla}(H_{\mathrm{b}}) \tag{3.4}$$

$$\tag{3.5}$$

This preconditioning is done through a set of iterative smoothing operations on $\vec{g_m}$. Typically $\sim$ 15 steps are needed for conjugate gradient optimization based smoothing (i.e. the preconditioning) of the force $\vec{g_m}$ of eq. 3.2.

DIISIP The preconditioned diis method is explained below. At the m-th step we have is an exact solution of a quadratic minimization problem and $\vec{c_i}, (i = 1, ..., m)$ a set of m approximate solution vectors, $\vec{g_i}, (i = 1, ..., m)$ are the gradient vectors and the error vectors are $\vec{e_i}, (i = 1, ..., m)$. The error vector $\vec{e_i}$ is defined as

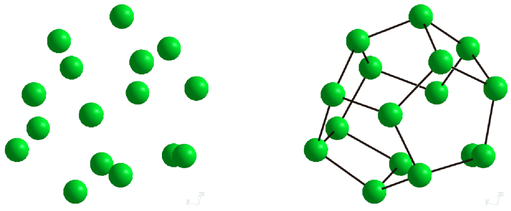
$$\vec{e_i} = P_m\vec{g_i}, (i = 1, ..., m)$$

We have the Lagrangian

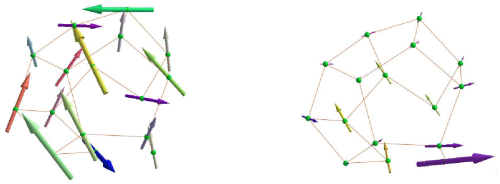
$$L =< \hat{e}_i|\hat{e}_j > -\lambda(\sum_i d_i - 1) \tag{3.6}$$

where for $i = 1, ..., md_i$ are the coefficients to be used to form a new solution vector $\vec{c_{m+1}}$ from the old vectors. and λ is the Lagrange multiplier which couples the constraint that

$$\sum_i^m d_i = 1$$



Si$_{16}$ cluster Bonded Si$_{16}$ cluster



Forces Preconditioned forces

Figure 3.3: The application of the preconditioner (The figures were prepared using the program V_SIM [106])

Taking the derivative of the Lagrangian L in eq. 3.6 with respect to $d_i, (i = 1, ..., m)$and λ we get a system of linear equations of the order $(m + 1)$ (eq. 3.7) .

$$\begin{bmatrix} e_1e_1 & e_1e_2 & \dots & e_1e_m & -1 \\ e_2e_1 & e_2e_2 & \dots & e_2e_m & -1 \\ \vdots & \vdots & \vdots & \vdots & -1 \\ e_me_1 & e_me_2 & \dots & e_me_m & -1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_m \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{3.7}$$

Solving the eq. 3.7, we get

$$\vec{\widetilde{c}}_m = \sum_{i=1}^{m} d_i \vec{c_i}$$

Under the assumption that we are in a quadratic region, we have

$$\begin{aligned}
\vec{\widetilde{g_m}} &= \vec{\nabla}\phi(\vec{\widetilde{c_m}}) \\
&= \vec{\nabla}\phi(\sum_{i=1}^{m} d_i\vec{c_i}) \\
&= \sum_{i=1}^{m} d_i\vec{\nabla}\phi(\vec{c_i}) \\
&= \sum_{i=1}^{m} d_i\vec{g_i} \\
\vec{c_{m+1}} &= \vec{\widetilde{c_m}} - \widetilde{P_m}\vec{\widetilde{g_m}}
\end{aligned}$$

So, $\vec{c_{m+1}}$ is the solution vector at the end of the iteration m and it goes for the next iteration.

RESULTS

Here in this section, the performance of the DIISIP optimizer is shown. We have used 16 silicon cluster (Lenosky forcefield[105]) for showing the preconditioned forces visually. The time-scaling of the preconditioner is tested on alanine peptide chain of length ranging from 10 to 900 residues. For the application of DIISIP we have chosen Lennard-Jonnes clusters of size 20, 50 and 150 , alanine peptide chain of length 4, 20, 40, 200 residues. Here, for the biomolecules the optimizations were carried out using the OPLS all-atom force field (OPLS-AA) [102] for the potential energy of the biomolecular system as implemented in the DYNAMO modeling library [103, 104]. For all methods, the geometry optimization is considered converged when the root-mean-square (RMS) force is less than $< 10^{-6} KJmole^{-1}$.

In fig. 3.3 we have shown a cluster system made of 16 silicon atoms. First we search atomic pairs within a certain distance (a little more than crystalline silicon-silicon bond length). The bonded network is built. A quadratic potential based model Hamiltonian is constructed - for this system the force constant is found by trial and error. (But for biomolecules which have the bond stretching component explicitly present in the potential function, we use them taking directly from the parameter file). A typical force vector and the preconditioned force is shown. The preconditioner can be seen to have actually smoothed the force considerably. In fig. 3.4 the time scaling of the preconditioner is plotted. It shows a square scaling. We don't need to have any calculation of gradient of the potential energy function during the preconditioning which is actually much more time consuming than this cheap preconditioning step.

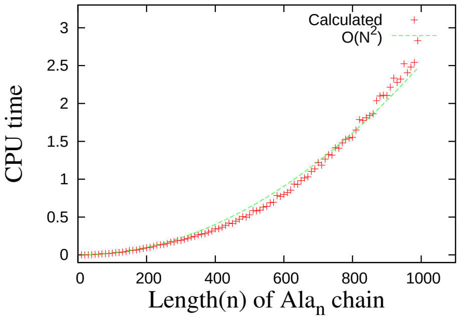


Figure 3.4: Scaling of the preconditioner with the system size. We plotted here the CPU time (in sec) needed for a single preconditioning step against the chain length(n) of polyalanine molecules(Ala_n)

In fig. 3.5 we compared the performance of conjugate gradient, DIIS and the DIISIP algorithms. DIISIP clearly outperforms others by converging $\sim$ 5 times faster. The same thing can be seen when we have applied DIISIP for a medium sized (150 atoms) Lennard-Jonnes cluster.

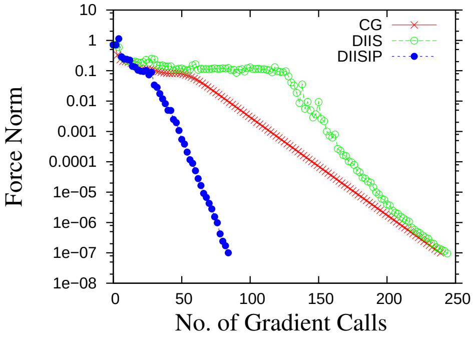


Figure 3.5: Force norm vs step number for geometry optimizations on a 16-atom Silicon cluster.

In fig. 3.6 we compared the regular DIIS and the DIISIP algorithms for a 20-residue alanine chain peptide. DIISIP converges fast here for this short peptide.

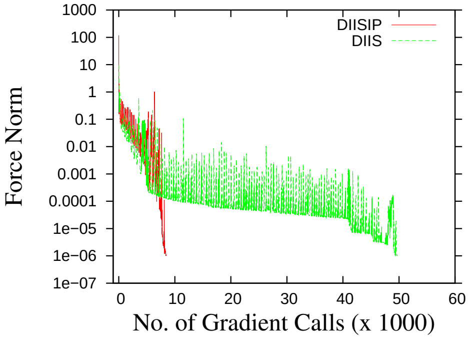


Figure 3.6: DIIS and DIISIP : Force norm vs step number for geometry optimizations on $(Ala)_{20}$ chain.

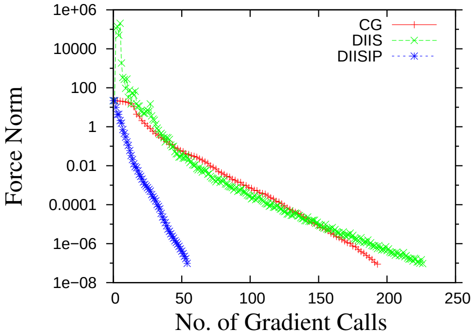


Figure 3.7: Force norm vs step number for geometry optimizations on Lennard-Jonnes 150 atom cluster.

The history length of DIIS is typically kept to be < 15 because of the stability issue. Here in fig.3.8 we tested the parameter history length for 4-residue alanine chain. It shows that a longer history length is allowable to get the best performance. But one has to take care of a probable instability which can occur by restarting or using a hybrid scheme. In the next subsection we discussed a couple of probable schemes which may help to extract the best out of this geometry optimizer.

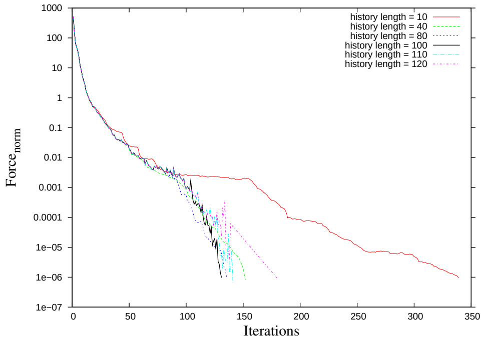


Figure 3.8: Effect of the DIIS history length on the Force norm (the system is $(Ala)_4$ chain.

Efficient implementation : Hybrid optimizer

We thus combine methods to gain a fast hybrid geometry optimization scheme:

Scheme 1
- (1) We do $10 - 100$ steps of steepest descent and go to step (2)
- (2) the optimization algorithm switches to DIISIP. Once there is a untrusted move ($fnrm/fnrm_{old} > 10.0$, $fnrm$ is the force form) we switch back to step (1) and the cycle continues.

Scheme 2
- (1) The geometry optimization begins with the steepest descent method for $20 - 100$ steps for the highly non-quadratic nature of starting position which is typically away from the quadratic region
- (2) Conjugate gradient up to the norm of the force coming around 10^{-2}
- (3) the optimization algorithm switches to DIISIP when the root-mean-square force of the latest point is smaller than 10^{-2}
- (4) whenever there is a untrusted move ($fnrm/fnrm_{old} > 10.0$, $fnrm$ is the force form) we switch back to steepest descent and continue that for $10 - 100$ steps and then go back to (3) and continue with DIISIP and the cycle continues.

The scheme 2 can give a very good stability. But for letting DIISIP face a real test of starting from a considerably high force norm with the help of just SD method, we started testing the scheme 1 (calling it SD-DIISIP) extensively before applying the scheme 2 (calling it CG-DIISIP).

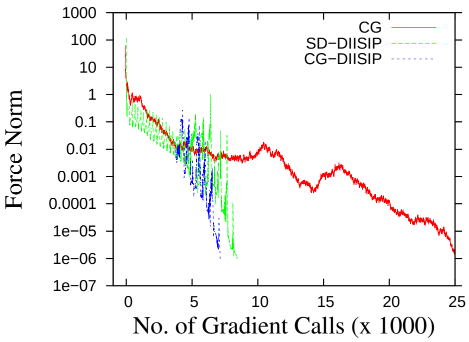


Figure 3.9: Force norm vs step number for geometry optimizations on $(Ala)_{20}$ chain.

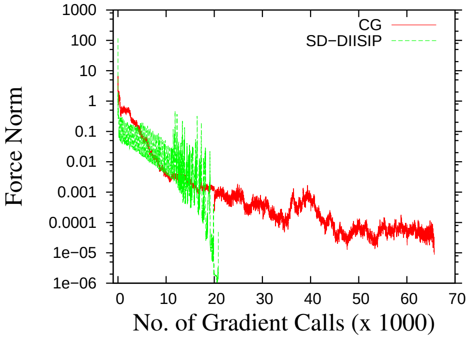


Figure 3.10: Force norm vs step number for geometry optimizations on $(Ala)_{40}$ chain.

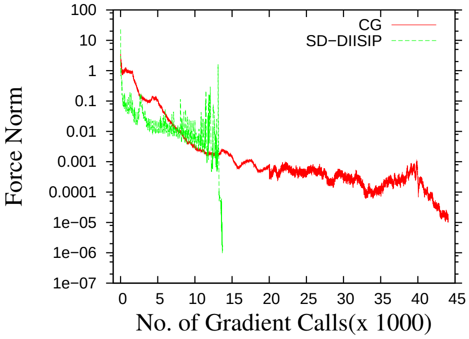
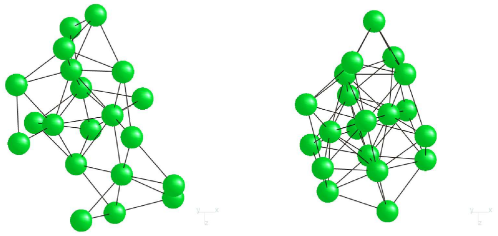


Figure 3.11: Force norm vs step number for geometry optimizations on $(Ala)_{200}$ chain.

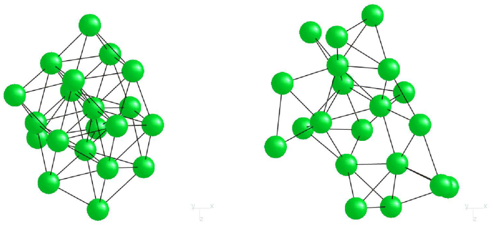
The figures 3.9,3.10 and 3.11 are the comparison of our best performing conjugate gradient implementation (coupled with SD) and the DIIS (coupled with SD). The SD-DIISIP work 3-5 time faster in all the cases. The fig.3.9 shows that CG-DIISIP implementation is even faster than the SD-DIISIP implementation.

Non-uniqueness of geometry optimization

As it is known that starting from any arbitrary point of the energy landscape for different geometry optimizers (e.g. conjugate gradient, diis, bfgs etc) the optimization path can lead the process into a final geometry optimized state which may not be identical as compared to each other. We present an example here. Starting with a 50 atom Lennard-Jones cluster we found that the energy of the optimized states for SD, CG and DIISIP method are all different from one another (see fig.3.12. The starting energy is -51.69, SD minimum -72.79, CG with -73.47 and DIISIP with -52.88 (in the units of ϵ of the LJ potential function). It shows that SD being exact in its definition is true to the connectivity present in the landscape. CG is exploratory and for that it may lead to a far-off minimum which may be the lowest in the neighborhood of connectivity. DIISIP is somewhat biased to the physical proximity while choosing out a mode reaching the "nearest" minimum. For a molecule where the covalent bonds are stiffer as compared



Initial SD minimum



CG minimum DIISIP minimum

Figure 3.12: Images of the minima using different geometry optimizers starting from the same initial structure (The figures were prepared using the program V_SIM [106])

to an inert gas cluster (LJ potential) the physical proximity is more correlated to the landscape connectivity - there, even if the geometry optimizers land up in different minima in some cases, the difference among those minima in terms of energy or the structural criteria is not expected to be very large.

CONCLUSION

Here minima hopping algorithm was described minima hopping algorithm. The *dimersoft* and mode decomposition schemes were described and in the following chapters it would be seen that those increased the efficiency minima hopping algorithm. We presented a preconditioner for the steepest descent algorithm. A geometry optimization method using DIISIP was developed. The DIISIP method uses a preconditioner determined iteratively using a Hamiltonian made of the short ranges interactions present in the system. We tested the stability of the algorithm with respect to the DIIS history. The hybrid implementations of the DIISIP algorithm coupled with steepest descent algorithm (SD-DIISIP) and with conjugate gradient (CG-DIISIP) implementation have been shown to outperform the performance of our best implementation of conjugate gradient algorithm.

Chapter 4

Biomolecules in gas phase: Conformational global optimization

Here we conducted a series of global optimization studies using Minima Hopping method(MHM) for a number of small peptides and a non-biological pyridine-pyrimidine oligomer in their gas phase with an all-atom OPLS forcefield. A softening scheme employed in the molecular dynamics based move part of the algorithm is tested here and usage of it in an effective way is worked out. The aim of the following study was to extend the usage of minima hopping algorithm towards a better understanding of some of those biomolecules which have been a sharp focus of the biochemists working on different aspects of the biomolecular interactions. The putative ground state of them have been analyzed. For a 5-residue peptide named Met-enkephaline we compared the dihedral angles of the energetically lowest structure with previous studies. The polyalalnes which form helices in vacuum when the c-terminal is appended with a positive charge has been used to understand the helix-coil transition on the conformational space. Going towards a bigger peptide system we studied the global optimization of a 34-residue peptide which were designed to form a helix-turn-helix motif. The minima-hopping search pathway suggests a parallel helix-turn-helix intermediate and an antiparallel helix-turn-helix global minimum. The self organization of alternating pyridine-pyrimidine oligomer which form into a helical superstructure has shown to be having an energy landscape where the minima hopping search pathway records distinctive gaps in conformational energy, but with a different charge set parametrized on a helical conformation instead of an extended conformation the landscape becomes smoother.

Introduction

The complexity within the spectrum of all interactions in a biomolecular system is vastly reduced if the environment is vacuum. The simpler electrostatics and comparatively small number of competing forces bring computational affordability and more accuracy in the study. Some pioneering experimental works [107, 108, 109, 114, 115, 110, 111, 112] on peptides in their gas phase have already been done. Those studies pointed towards understanding the helix-coil transition, hydrogen bond formation, the role of terminal charges, conformational heterogeneity at various temperatures etc.

Studies of protein folding using computational means based on first principles methods is a nontrivial problem. Here the study has to be aimed at the accurate portrayal of the subtle balance among the conformational energy, solvation free energy and the conformational entropy. A very accurately parametrized forcefield, clear and close-to-real model of the protein-solvent interaction including protein and water hydrogen bonding, fast sampling of the conformational space are among the shortcomings of this whole computational approach. Also, for the studies of biomolecules in solvent there can be issues specific for each of the standard experimental structure calculation techniques - e.g. crystallization for some proteins in x-ray crystallography or overlapping signals in NMR which make the whole problem diversely complicated. One can get rid of a major portion of the inaccuracies related to the modeling of protein-solvent interactions in the computational studies if proteins can be studied in gas phase. Now the introduction of high-resolution ion mobility measurements, which allow experiments to be performed on biological molecules in gas phase [114, 115] is a motivation behind working with this simplified problem - that is, to build a framework to more carefully study a subset of the interactions present rather than falling apart with the whole intractable problem of solving biomolecular structure and dynamics in aqueous or solvent environment. Also, from the perspective of biology, one can find a non-aqueous environment to be important in many case - e.g. membrane proteins which are involved in transport, recognition, and ligand-receptor binding have a low dielectric constant environment because a large portion of the surface is shielded from water, or water excluded hydrophobic cores of globular proteins with hydrophilic surfaces.

Also, for the numerical algorithms to be developed and tested very fast one needs some simple systems having simple interactions within, less-complicated but physically relevant and having some experimental studies already done and so working as a reference. So, the gas phase simulations of biomolecules, being computationally less expensive with many experimental gas phase studies going on around like mass-spectrometry or spectroscopy, serves as a good framework.

Here we have applied the minima hopping algorithm to find and study the lowest energy conformations of the following biomolecules.

(A) N-acetyl-alanine-N-methylamide or alanine-dipeptide

(B) Met-enkaphalin

(C) $AcAla_8Lys + H^+$

(D) $AcPheAla_5Lys + H^+$

(E) $AcPheAla_{10}Lys + H^+$

(F) $AcAla_{14}LysGly_3Ala_{14}Lys + 2H^+$

(G) Alternating pyridine-pyrimidine oligomeric strand

We applied this algorithm recently in a biomolecular structure prediction study [246]. Systems like alanine-dipeptide, met-enkaphalin which have been frequently used as model systems for many computational studies, here have been used as the benchmark systems. The lowest energy minima of those two systems have been shown and their internal coordinates have been tabulated. For these systems the lack of experimental studies hinders the possibility of validate the theoretical prediction, so we compared our predictions with the ones already reported before. For an unusual system like a pyridine-pyrimidine oligomer, we conducted the charge parametrization following an ab-initio geometry optimization (B3LYP/6-31G*) and then used those for the global optimization study. In the current study, we tried to match some of the structural properties with the experimental observables, if those are available in literature. For $AcAla_{14}LysGly_3Ala_{14}Lys + 2H^+$ peptide we calculated the collision cross section of the conformations - this quantity is related to the drift time distribution in an ion-mobility experiment. For $AcAla_8Lys + H^+$ the dipole moment has been calculated using the point charges. For $AcPheAla_5Lys + H^+$ and $AcPheAla_{10}Lys + H^+$ the vibrational spectra (stick-spectra) have been calculated within a harmonic approximation using an ab-initio method and have been referenced to the infrared-ultraviolet double resonance spectra in the ref. [111].

Model and Method

Here, we used the OPLS all-atom force field (OPLS-AA) [102] for the potential energy of the biomolecular system as implemented in the DYNAMO modeling library [103, 104]. The energy expression has been discussed in the section 5.1.1.

The simulation method has been described in detail in the section 3.1 and in chapter 5. We had to incorporate the following changes :

- *Checking for the chirality changes.* A very high temperature MD simulation can change the chirality of certain protein groups whereas folding preserves chirality. To prevent this, we devised a filter which checks the chirality of all the residues and rejects chirality-changing moves immediately. In this way the global minimum search is restricted to the biologically-relevant free energy surface of the atomistic system.

- *Searching along slow vibrational modes.* To concentrate searching in slow vibrational modes we did not use simple random initial velocities in the MD simulations and instead took used a Dimer-Softening to generate an initial velocity towards a direction of low curvature. It has been described in section 3.1.1.

The Dimer-Softening scheme

In the minima hopping algorithm the initial velocity in the molecular dynamics based escape moves is usually generated with random number with a gaussian distribution, which is then rescaled to fit the desired kinetic energy (*ekinetic*). In the original MH method low kinetic energy trajectories could only be obtained by using large values for `mdmin` which results in long trajectories. Also, we have seen that we can get rid of the fast vibrations coming from the bonds and the angles and have an initial velocity directed towards a random direction on a subspace free from bond stretching vectors and angles bending vectors (section 3.1.1, and ref. [246])- this *softening* scheme actually enhances the performance of the search algorithm. In this way one can typically find MD trajectories with a relatively small energy that cross rapidly into another basin of attraction. Here we tested another *softening*[126] scheme to choose velocities along low-curvature directions.

A direction of low curvature is calculated using a modified iterative dimer method [128]. This method doesnt need the calculation of the second derivatives - the energy gradients suffice for it. Starting from a local minimum a second point is taken along the initial escape direction. These two point work as a dimer of the system. The first point always kept fixed. Now, along the perpendicular direction of the dimer the second point is moved by an amount proportional to the component of the force on this point perpendicular to the dimer. By this way, changing the orientation of the dimer gives a direction with a different curvature. This is repeated for a few steps before ending up into a direction having a low curvature (curvature tolerance). Initial velocities for the MD escape are then chosen along the final escape direction $\hat{\mathbf{N}}$.

Test of DIMERSOFT scheme on $AcAla_8Lys + H^+$ system

For a peptide $AcAla_8Lys + H^+$ 4.3 we tested this dimer method derived *softening* method (dimersoft) - firstly, to see the impact of the initial *dimersoft* direction in the MD trajectory and secondly, to find an optmimum value of the curvature upto which the direction has to be ”softened”.

Starting an MD trajectory with initial velocity along a ”softened” direction gives a remarkably different dynamics as compared to a trajectory with a starting velocity along a random direction. The fluctuation in the conformational energy is very different in the two cases (see fig. 4.1). The backbone of the protein moves quite a lot in the first case.

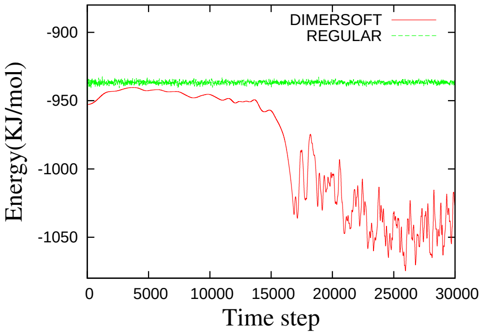


Figure 4.1: Energy profile of MD trajectory along a random direction and along a vector generated by Dimer-Softening scheme

We then have done 30 independent global optimization runs for this peptide for each of the different curvature values corresponding to the initial velocity directions of the MD escape. The fig. 4.2 shows that there is an optimum value or a range of values of curvature exists for such a search for a low-curvature direction.

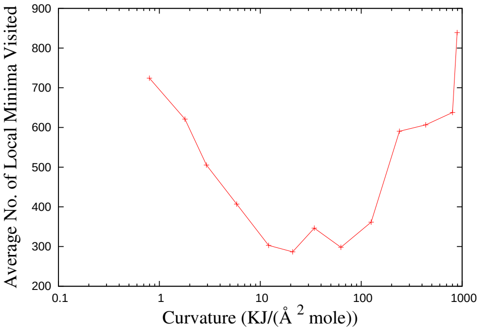


Figure 4.2: Average number of minima visited in minima hopping runs versus the curvature of the mode generated by Dimer-Softening scheme along which the molecular dynamics moves are conducted.

With a high value of curvature-tolerance the performance in global optimization is found to be bad - also, if the softening procedure is executed until it reaches a very low value of curvature the performance drops again. One can see in the figure 4.2 near the smaller values of the curvature that using an "ultra-soft" mode brings down the algorithmic performance as much as the modes with large values of the curvature. Always moving along the *same* soft mode direction of a given minimum make the possible escape routes fewer, which actually weakens the search algorithm. In minima hopping runs the mean kinetic energy is like an optimum filter for the high barriers and here the curvature is a sensitive filter. Here we observe that the curvature tolerance which gives the best performance of the algorithm is roughly the value where the mean kinetic energy starts increasing by leaving the flat region in the fig. 4.3.

Typically < 50 iterations are done with a step size $\alpha = 0.00014\mathring{A}$ (this is $\sim$ the step size used in the steepest descent) and a dimer length of $d = 0.01\mathring{A}$.

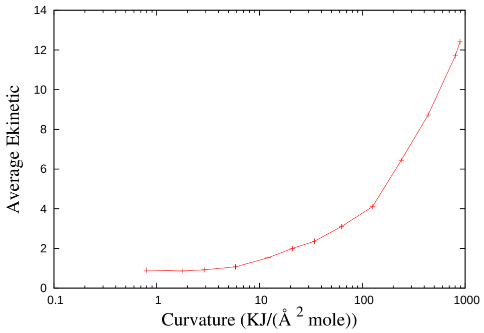


Figure 4.3: Average number of minima visited in minima hopping runs versus the curvature of the mode generated by Dimer-Softening scheme along which the molecular dynamics moves are conducted.

4.1 Alanine dipeptide

The alanine dipeptide molecule (N-acetyl-alanine-N-methylamide)[153] has been used as a model system for many computational studies of biopolymer structure and dynamics[154, 155, 156, 157, 158, 159, 160, 161, 162, 163, 164, 165]. For an initial benchmark simulation of minima hopping algorithm for peptides in gas phase we chose this molecule for its being able to adopt all conformational angles observed for a helix and β strand motifs in proteins[154, 155]. The structure and thermodynamics of alanine dipeptide have been characterized by using both theoretical and experimental methods[166, 167]. In a previous study [151] the conformational kinetics and the solvation effects of the molecule have been elucidated. The potential energy surface of alanine dipeptide exhibits approximately five minima, and the exact number depends on the details of the potential model. For this system the only pair of the ϕ(C-N-Cα-C) andψ(N-Cα-C-N) backbone dihedral angles present in it can serve as the coordinates onto which the potential energy surface can be projected.

Here, our goal was to conduct a thorough search to find the conformational minima and to compare the dihedral angles ϕ and ψ with the previous studies[151, 152] which using ab-initio methods found the stable conformations C_{7eq}, C_{7ax}, C_5, β, α_R, α_L etc. Table 4.1 shows the internal coordinates and the OPLS energies of the lowest 10 conformations in the minima hopping search. The last 3 columns are the information of the structures and energies reported in the references [151] and [152]. In our global optimization run we could not find any low energy stable β_1 or β_2 conformation. Here in fig.4.4 the global minimum of OPLS forcefield is shown.

4.2 Met-enkaphalin

Met-enkephalin is a model peptide which is used for computational studies quite often. It was identified from the enkephalin mixture from brains [137]. The sequence is Tyr-Gly-

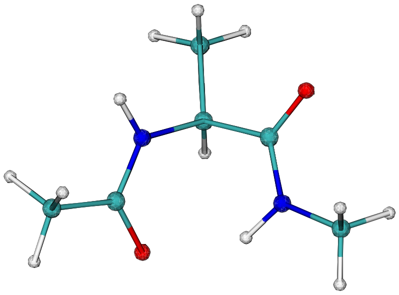


Figure 4.4: $C7_{eq}$, the global minimum structure of OPLS forcefield for alanine dipeptide. Internal coordinates are as follows: $\phi = -80.291$, $\psi = 67.8$, $\omega_1 = 179.33$ and $\omega_2 = -179.04$. (The figure is prepared using the program VMD [130])

Gly-Phe-Met. This pentapeptide is found in peripheral tissues and in brains (pituitary). It helps in many physiological processes. Previous experimental studies suggest that it can adopt many different structures in aqueous solutions[138]. The lowest energies of Met-enkephalin without explicit solvation effects were previously determined based on many potentials including the ECEPP/2 and ECEPP/3 potentials [139, 140, 141, 142, 143, 144, 145, 146, 147, 148, 149, 150]. From the minima hopping run we collected 30000

Table 4.1: Table of conformational energies and their internal coordinates. All the energies are in KCal/mole.

	Conformation	ϕ	ψ	OPLS Energy	ϕ	ψ	HF/6-31 +G**
1	$C7_{eq}$	-80.2	67.8	0	-85.8	79.0	0.00
2	$C5$	-143.6	161.4	1.29	-157.2	159.8	0.40
3	$C7_{ax}$	68.4	-55.0	2.55	76.0	-55.4	2.82
4	$C5$	-139.8	154.4	4.76	-	-	-
5	-	-100.7	113.3	5.5	-	-	-
6	α_R	-77.0	-22.3	6.0	-60.7	-40.7	4.35
7	$C5$	-144.4	161.5	6.2	-	-	
8	-	-157.2	-62.4	6.46	-	-	
9	-	-143.3	-56.6	8.3	-	-	-
10	α_L	46.4	51.3	8.95	67.0	30.2	4.76

MET5:OT2
TYR1:HT1
TYR1:HT3
MET5:OT1
GLY3:O
GLY3:H
MET5:OT1
GLY2:H
MET5:HT2
PHE4
GLY2:O
PHE4:H

OPLS Global Minimum of
Met-enkaphalin with
COO^- C-terminus

OPLS Global minimum of
Met-enkaphalin with
$COOH$ C-terminus

MET5:OT1
GLY2:H
MET5:HT2
PHE4:O
TYR1:HH
GLY3:O

Lowest ECEPP/3 energy minimum
of met-enkaphalin
with COO^- C-terminus

Lowest ECEPP/3 energy minimum
of met-enkaphalin with
$COOH$ C-terminus

Table 4.2: OPLS Global minimum of met-enkaphalin along with the lowest ECEPP/3 energy conformation

local minima and did geometry optimization using ECEPP/3 potential implemented in SMMP package[168, 169, 170]. We are aware that for a careful global optimization run with ECEPP/3 forcefield one may find conformations having lower ECEPP/3 energies - but optimizing all OPLS minima in ECEPP/3 forcefield would surely give us many of the lowest energy conformations in the ECEPP/3 energy landscape of the peptide. In this way we hope to cross-map a large portion of the energy landscape between the two forcefields, considering that it is a very small peptide. The lowest energy conformations corresponding to the OPLS and ECEPP/3 forcefields are shown in fig.4.2. The 10000 lowest OPLS energy conformation and the ECEPP/3 energies corresponding minima in ECEPP/3 forcefield is shown in fig.4.5. The lowest energy conformations in these two forcefields are distinctly different. The energetic correlation between these forcefields is very poor. In the table 4.3 we compared the internal coordinates of the the OPLS global minimum and the global minimum in ECEPP/3 forcefield reported in a previous study[150]. The internal coordinates of the lowest ECEPP/3 energy conformation from our study are also tabulated.

Table 4.3: Internal coordinates corresponding to the lowest energy minimum, both for OPLS forcefield and ECEPP/3

	OPLS		ECEPP/3		ECEPP/3 Previous study[150]	
	ϕ	ψ	ϕ	ψ	ϕ	ψ
TYR1	-	-35.0	-	-53.0	-	155.8
GLY2	-92.6	-49.7	162.6	155.8	-154.2	85.8
GLY3	-113.3	-124.8	77.7	-80.1	83.0	-75.0
PHE4	-68.0	-45.5	-145.2	22.4	-136.8	19.1
MET5	-65.9	-	-166.9	-	-163.4	-

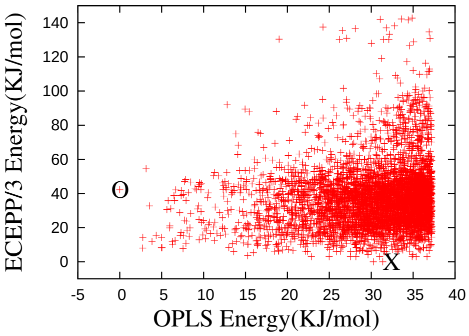


Figure 4.5: Energies of the OPLS minima and the corresponding ECEPP/3 minima. "O" is the OPLS global minimum and "X" is the conformation having lowest ECEPP/3 energy.

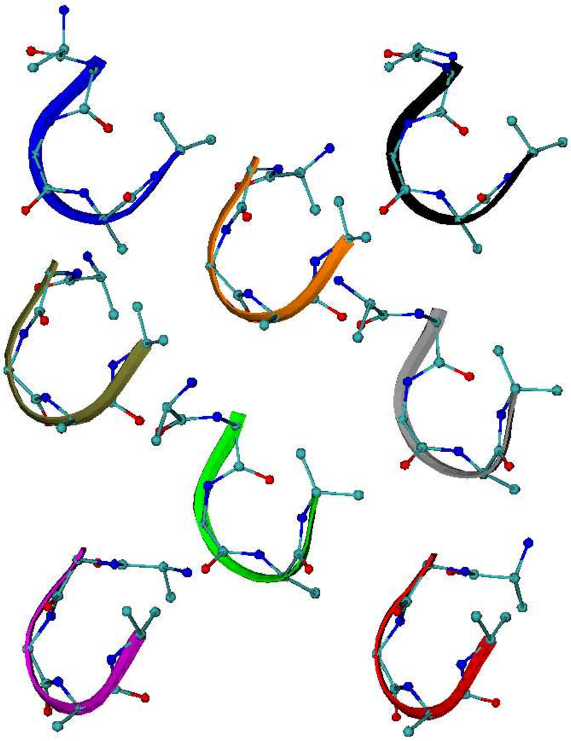


Figure 4.6: The global minima of all possible chiral combinations for the peptide sequence TYR-GLY-GLY-PHE-MET(metenkephalin). The color of the ribbon in each of these conformations represent a chiral combination which is listed in the table 4.4.

Mixed Chirality

For metenkephalin, out of the 5 residues 3 (TYR, PHE and MET) residues can have both the L-chiral and the R-chiral conformation, because GLY residue is achiral. So the landscape of the sequence TYR-GLY-GLY-PHE-MET can produce $2^3 = 8$ mixed chiral combinations. We took all of this combinations and conducted independent minima hopping runs and in the table 4.4 we have shown the energy of each of the lowest minimum corresponding to a particular chiral combination. In the fig. 4.6 all those lowest energy minima are shown. For this peptide it is observed that the global minimum of the potential energy surface is a mixed-chiral conformation. It is lower in energy by $\approx 6.58 KJmole^{-1}$ than the lowest among the pure chiral conformations.

4.3 $AcAla_8Lys + H^+$

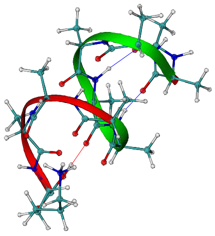
Alanine has one of the highest helix propensities - and without the solvent environment it forms a helix beyond a certain chain length. The protonated polyalanines which are $<$ 15 residues do not form helices in gas phase. Here we did a minima hopping run starting from a helical configuration of $(Ala)_8H^+$ sequence. The helical conformations are energetically higher than the globular conformations which actually was pointed out in the reference [107]. In this globular or collapsed conformation the N and C termini come close to each other. We show here in fig.4.3the minimum energy conformation from our run for this sequence. But adding a lysine at the C-terminus of even as small as 7 or 8 residue long polyalanine, a helical structure can be found in gas phase [107]. Lysine at the C terminus can help optimizing the hydrogen bond with the C-terminal backbone carbonyl groups and the interaction of its charge with helix dipole , thus to form a helical structure. The charge at N-terminal of a polyalanine without C-terminal lysine destabilize the helical conformation. For polyalanines, there have been a number of computational and experimental studies done - e.g. the helix-coil transition[116, 120, 118, 119]. There the question of the accuracy of the forcefield for such a study was raised - specially regarding the subtlety in the accuracy of the partial charge for a charged peptide.

Here we tried to find the global minimum of a $Ac(Ala)_8LysH^+$ chain as a part of the early-stage application of our algorithm(MHA). We started from a fully stretched configuration having an energy -951.24 KJ/mole after the initial geometry optimization. The global minimum which was found is helical with an energy -1335.88 KJ/mole. For a system like $Ac(Ala)_8LysH^+$ β-sheet structures are not stable in vacuum - here the β-sheet intermediates are of energy $\sim$100 KJ/mole higher than the global minimum.

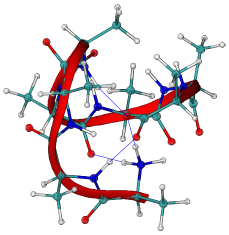
Table 4.4: Energy of the putative global minima of all possible chiral combinations for the peptide sequence TYR-GLY-GLY-PHE-MET(metenkephalin)

Chirality of TYR	chirality of PHE	Chirality of MET	Corresponding color in the fig.4.6	OPLS energy of the global minimum
L	L	L	blue	-618.458
R	L	L	black	-623.600
L	R	L	orange	-621.643
R	R	L	tan	-625.041
L	L	R	silver	-625.041
R	L	R	green	-621.643
L	R	R	magenta	-623.600
R	R	R	red	-618.458

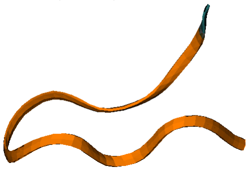
Structure



Global Minimum of
$Ac(Ala)_8LysH^+$



Global minimum of
$(Ala)_8H^+$



A beta-sheet shaped intermediate
of $Ac(Ala)_8LysH^+$ folding

Search pathway

This run was conducted on 16-processors and the run was continued up to searching 100000 energetically distinct minima. The putative global minimum was found within the first 10% of the simulation length. All the processors found the same energetic lowest conformation multiple times during the simulation. We pick up the search of one of the processors and followed the sequence of the local minima found by that processor - this way we traced out a pathway connecting the stretched peptide and the global minimum of that. In fig.4.7 and fig.4.8 the potential energy versus the index of the local minimum found during two of the Minima Hopping runs are shown. A few intermediate structures are also shown alongside the fig.4.7. The global minimum is shown in fig. 4.3 along with the hydrogen-bonds which stabilizes the helical structure of the peptide.

The lowest 5000 minima from the run were taken for comparing the energetic ranking with ECEPP/3 forcefield. The cross-correlation is shown in 4.9. The landscapes corresponding to the forcefields seems not be in agreement for this helical peptide, just like the case of met-enkaphalin. Moreover, the lowest energy conformations, shown in fig.4.10 suggests that in OPLS forcefield the global minimum is π-helical as compared to an $\alpha - helical$ conformation in ECEPP/3 forcefield. Of course, the protonation states were assigned on the NH_3 group in the LYS residue but the different partial charges may be responsible for this. In contradiction to this, the similarity in the fig. 4.9 (charged peptide) and

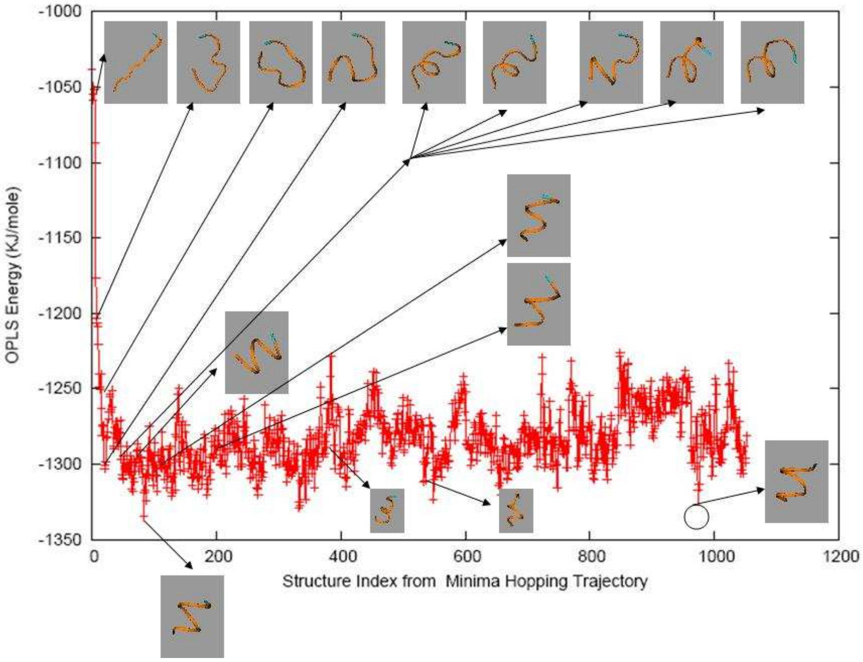


Figure 4.7: Minima hopping search profile - MD was started along a random direction. Energies of only the accepted conformations are shown.

fig.4.2(neutral peptide) suggests a bigger discrepancy in those forcefields with respect to each other. In ref. [136] the dominance of $\pi helical$ conformations have been described to be an artifact of the forcefields because the experimental observation of those helices are rare.

4.4 $AcPheAla_5Lys + H^+$ and $AcPheAla_{10}Lys + H^+$

This section is a work done following the reports in ref. [110, 111] where an application of infrared-ultraviolet (IR-UV) double-resonance spectroscopy was shown for the $AcPheAla_5Lys + H^+$ and $AcPheAla_{10}Lys + H^+$ peptides. The structural information was extracted from the pattern of spectral shifts of vibrational bands, mainly due to the internal hydrogen bonding. Because these experiments are done in a very low temperature, so the entropy effect is quite negligible here - and basically the potential energy surface should be very much similar to the free energy surface. At this low temperature the system can have many stable states with a certain distribution of the stability. So, not only the global minimum but many of the lowest energy conformations from the global optimization run can also have their places in the UV-photofragmentation spectrum. Along with the ab-initio analysis of the vibrational spectra of different low energy conforma-

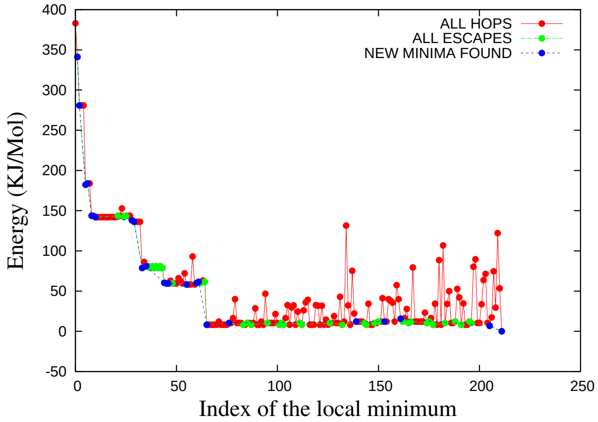


Figure 4.8: Minima hopping search profile - MD was started along a low-curvature mode, calculated using the dimer-softening scheme. This was the quickest of the all independent MHOP runs. In the search trajectory all hops, escapes and distinct new minima are shown.

tions one can hope to resolve the whole spectra of experimental data and can understand the modes of vibration, the relevant hydrogen-bonding ring, the overall hydrogen-bond pattern and related other structural elements.

The lowest 5000 minima from the minima hopping run were taken for comparing the energetic ranking with ECEPP/3 forcefield. The cross-correlation is shown infig. 4.13 for $AcPheAla_5Lys + H^+$ and in fig. 4.14 for $AcPheAla_{10}Lys + H^+$.

We conducted a few independent minima hopping runs to find the global minimum of each of the peptides. The minimum energy structure of $AcPheAla_5Lys + H^+$ peptide is shown in fig.4.11 and for $AcPheAla_{10}Lys+H^+$ in fig.4.12. For $AcPheAla_5Lys+H^+$ peptide, after picking up 100 lowest energy conformations we carried out single point energy calculation in the ab-initio scheme B3LYP/6-31G** using the Gaussian package[172]. In fig.4.15 we compared the OPLS energies with single point B3LYP/6-31G** energies for the lowest 100 conformations. Using the same ab-initio method, we optimized the geometries of the lowest 10 OPLS energy conformations. In the fig.4.16 we compared OPLS energies, B3LYP/6-31G** energies after the 1-st SCF cycle, density functional geometry optimized energies and the zero-point-corrected DFT energies. The energy rankings for the forcefield and DFT does not correspond well with each other for these structures. One of the minima with higher OPLS energy (12.2 KJ/mole) went relatively 7 mH lower in the B3LYP/6-31G** energy scale than the OPLS global minimum structure. For all these DFT optimized structures we followed up with a harmonic frequency analysis where the frequencies were scaled by a factor of 0.939 for comparison to the infrared spectra in [110, 111]. The stick-spectra is shown in fig.4.17. The N-H stretch region of the IR spectrum corresponding to the OPLS global minimum could not be matched with the IR-UV double resonance depletion spectrum corresponding any of the peaks in the UV-photofragmentation spectra. But the nearest match was the 8th lowest OPLS energy minimum (12.2 KJ/mole) which went to be the lowest in B3LYP/6-31G** energy after

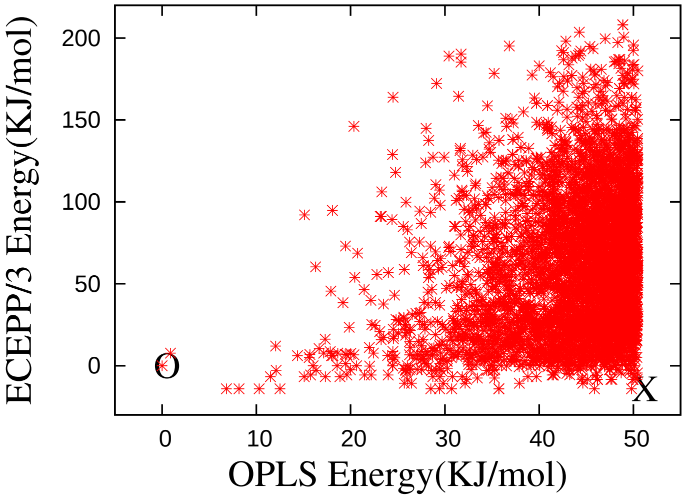


Figure 4.9: Energy correlation between OPLS and ECEPP

geometry optimization.

4.5 Helix-Turn-Helix motifs

Introduction

Moving towards bigger peptides which has more than one segment of secondary structure we chose a sequence designed to have a helix-turn-helix structure[117]. Motifs of this kind are known to cause DNA binding, calcium bonding etc. Here the sequence is AcA14KG3A14K+2H+. The alanine regions are supposed to be helices whereas the three glycine residues should break the helical structure at the connecting loop region - hence the two main conformations are expected the coiled coil helix and the extended helix.

The putative global minimum for this system is shown in fig. 4.18 - it has a coiled coil structure. The energy along the minima hopping trajectory of search is shown in fig. 4.20. In ref. [117] the cross-section of the lowest energy conformations are shown to be $460\mathring{A}^2$, whereas the lowest energy conformation in our simulation has a cross section of $472\mathring{A}^2$ and for the extended conformations the cross-section is $> 500\mathring{A}^2$. In fig. 4.19 the cross section of all the local minima from a successful run is shown. The cross-section was calculated using the hard-sphere-scattering-approximation. In the fig.4.21 we have drawn the energy map by projecting the conformational energies on a two dimensional surface comprising the collective variables RMS deviation with respect to a coiled-coil geometry (lowest energy structure) and the radius of gyration. We could observe parallel helices, having a cross-section of $480\mathring{A}^2$ as intermediates at an early stage of the search. This indicates that one possible mechanism of its folding can be in the following sequence

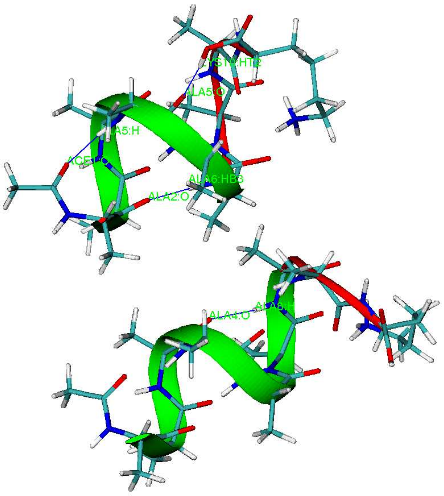


Figure 4.10: Minimum energy structure OPLS vs ECEPP

: (a) building of the secondary structure element(helix) at one end while no helix in the other half , (b) the loop going antiparallel to the helix, (c) a second helix forming along the a antiparallel direction to the loop and parallel to the first helix, (d) the second helix giving a 180^o rotation around the line connecting the centers of the helices (parallel to antiparallel transition) and (e) the optimization of the length of the loop and increasing the length of the second helix.

4.6 Pyridine-Pyrimidine oligomer

For this class of an oligomer or polymer chain the number of torsionally accessible conformations, which is dependent on the rigidity of the chain backbone is believed to be very small[122], i.e. the sterically non-prohibited minima are less abundant on their potential energy surface.

For the pyridine-pyrimidine system we conducted two sets of charge parametrization - first, for a stretched conformation(EXTENDED) and second, for a helical conformation(HELICAL). We optimized the geometries using an ab-initio method (B3LYP/6-31G*) using the Gaussian package[172]. The charges were fitted by matching the electrostatic potentials calculated by the same ab-initio method on 22255 points (on VDW surfaces in 4 layers with point density 6) by using Antechamber program of the AMBER package [171]. With these charges along with the other forcefield parameters [102] the topology was built. This way we had two sets of topology (EXTENDED and HELICAL)

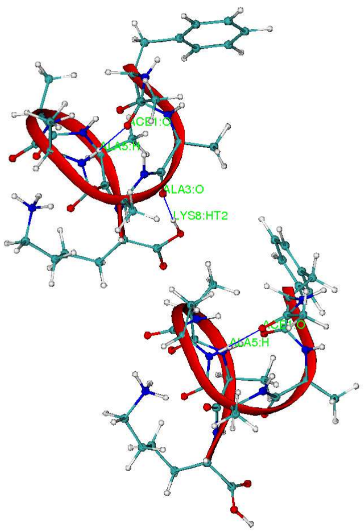


Figure 4.11: Putative global minimum of FA5K - Minimum energy structure OPLS vs ECEPP (The figure is prepared using the program VMD [130])

which were used for independent runs.

For both the parameter sets the global minimum energy corresponds to helical conformations. These two conformations (see in fig.4.23) are very similar and they are at 1.53$\mathring{A}$ all-atom RMS-deviation from each other. Though the minima hopping runs for these two parameter sets of the forcefield gives very different search profiles fig.4.24. There are big jumps in the energy profile along the search trajectory for the "EXTENDED" set of parameters and also for this case, it needs a longer search for finding the global minimum as compared to the "HELICAL" set. It points out the possibility that the energy landscape depend significantly on the partial charges for this kind of a system where the electron delocalization is prevalent. The figure 4.22 shows the difference in the two charge sets.

Conclusions

We studied here biomolecular structures in gas phase for a variety of systems. The algorithm remains successful in finding the putative global minimum and many low energy conformations efficiently which actually helped in understanding some of the experiments on peptides in gas phase like ion-mobility experiments or different spectroscopic experiments. With the help of ab-initio calculation we could identify conformations responsible for the infrared-ultraviolet double resonance spectra. From the comparative study of the predicted structures in different forcefields, it is apparent that its not easy to choose a

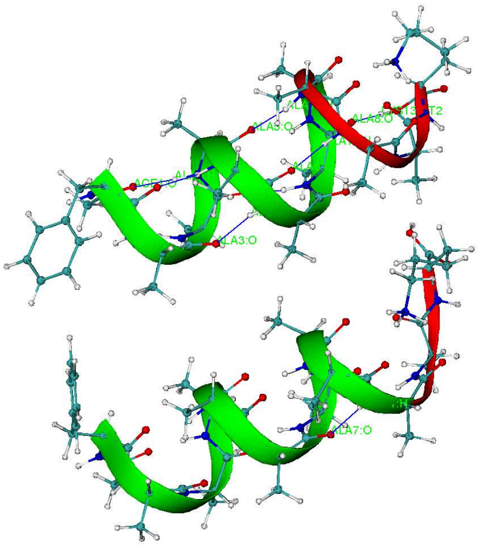


Figure 4.12: Putative global minimum of FA10K - Minimum energy structure OPLS vs ECEPP (The figure is prepared using the program VMD [130])

"correct" forcefield - this is indeed a bottleneck in the approach of computational structure prediction and dynamics. It is also seen in this study that the charge parametrization based on for specific conformations can be quite different from each other and this can make a considerable impact on the energy landscapes. Provided the forcefield parameters are correct, these simulations based on all-atom forcefield can help enormously by bridging between the theory and the experiments in structural studies of biomolecules or other organic molecules. On small to medium peptides this algorithm has been shown to work successfully and efficiently both from the computational and the physico-chemical perspective. To understand the biomolecular interactions better, it can be utilized as an insightful tool accompanying the future experimental studies to be done within the research community.

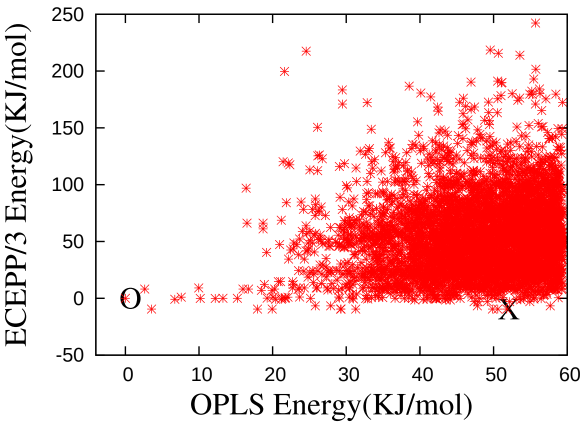


Figure 4.13: $AcPheAla_5Lys + H^+$: Energy correlation between OPLS and ECEPP

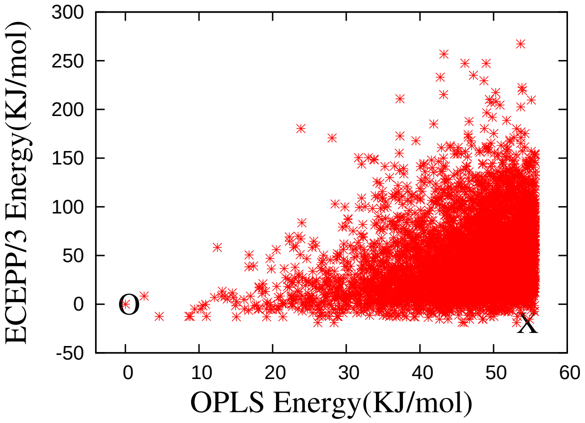


Figure 4.14: $AcPheAla_{10}Lys + H^+$: Energy correlation between OPLS and ECEPP

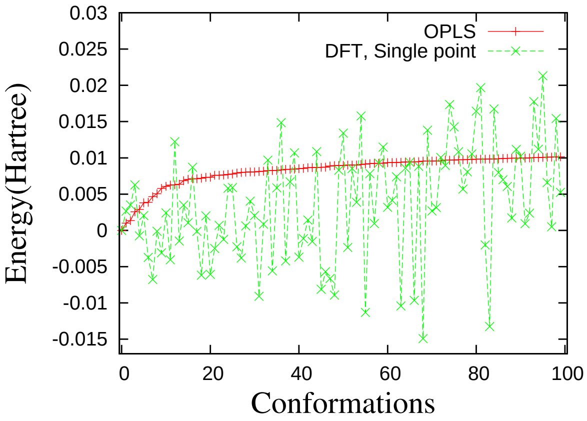


Figure 4.15: Comparison between OPLS and DFT energies of MHOP 100 lowest energy configurations.

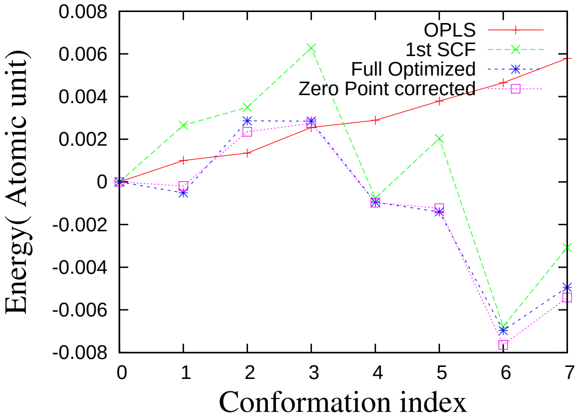


Figure 4.16: Comparison between OPLS and geometry optimized DFT energies of MHOP 10 lowest energy configurations.

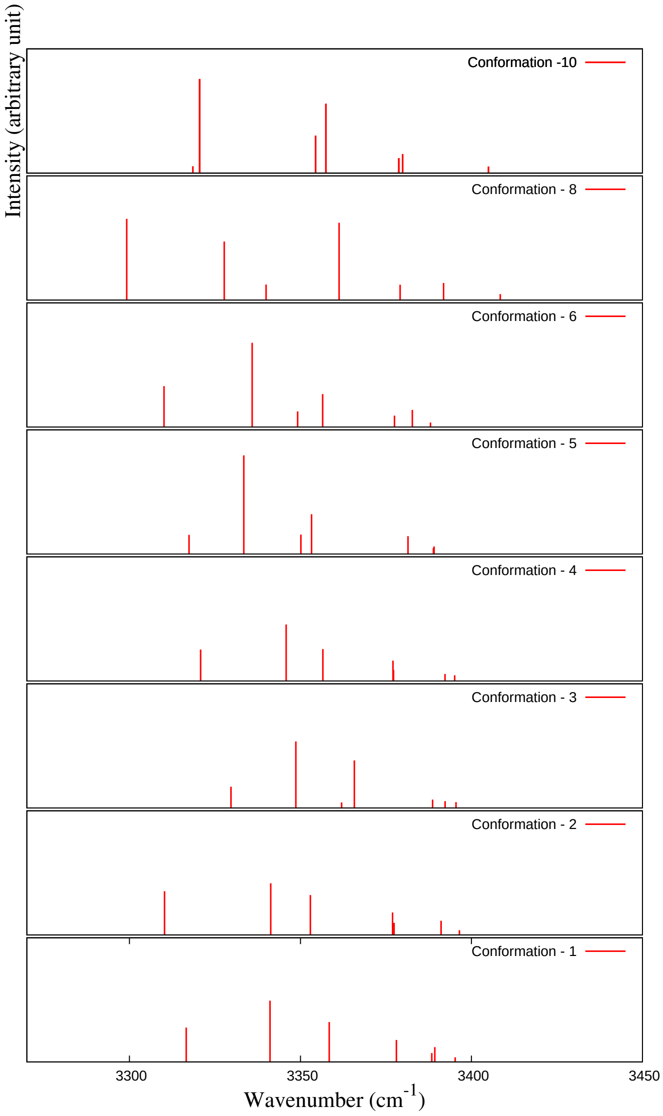


Figure 4.17: The calculated stick spectra of the lowest energy conformers found from MHOP search.

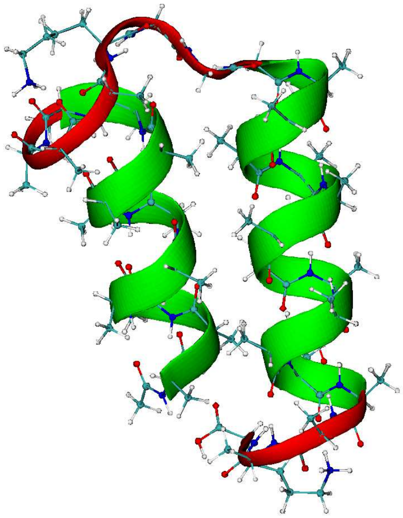


Figure 4.18: Lowest OPLS energy conformation for the peptide having sequence AcA14KG3A14K+2H+ . (The figure is prepared using the program VMD [130])

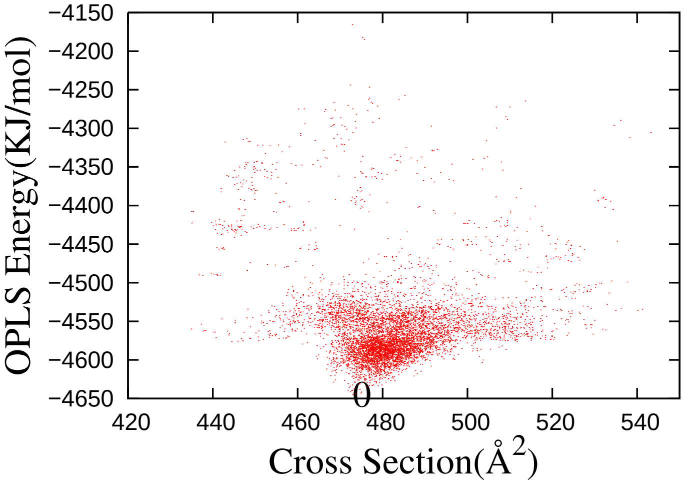


Figure 4.19: OPLS energy vs Cross-section for all the conformations in a minima hopping run.

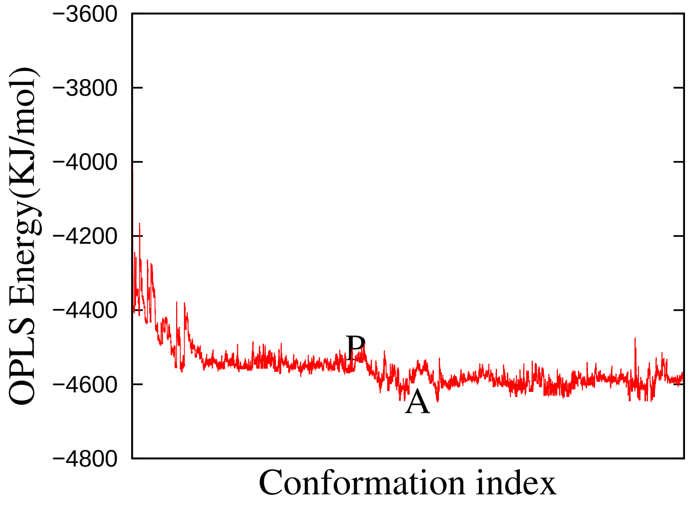


Figure 4.20: OPLS energy along the search trajectory. "A" is where it finds many antiparallel-helices and "P" is where many parallel-helices are found

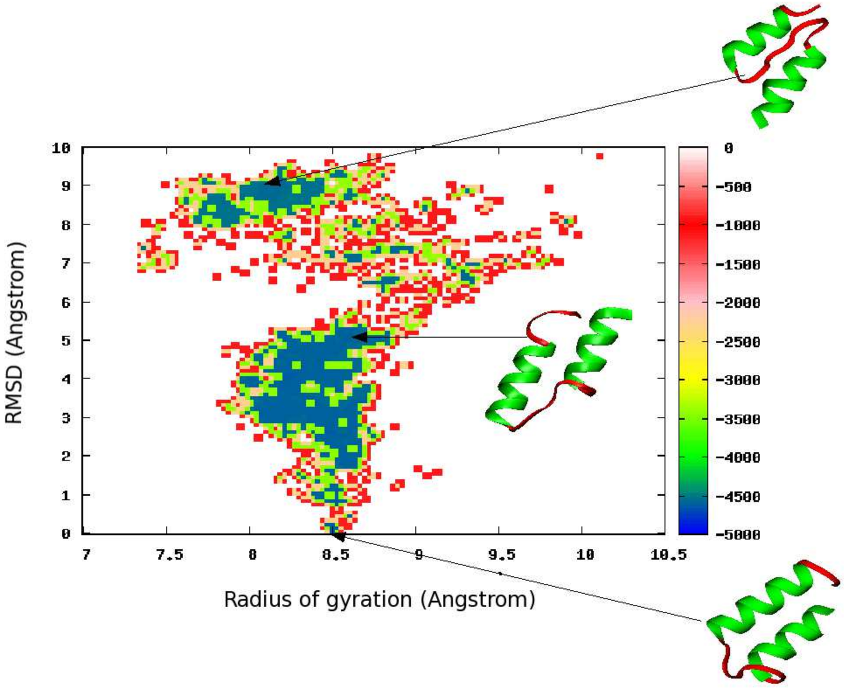


Figure 4.21: Energy map along RMSD and radius of gyration.

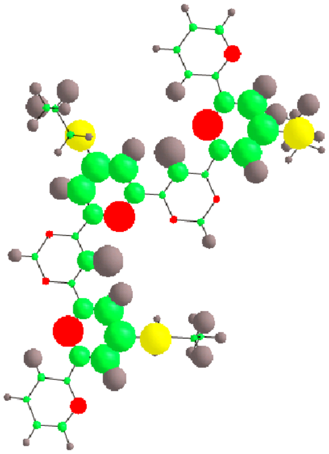


Figure 4.22: The radius of any sphere is proportional to the difference in charge in the corresponding atom from the two sets of parametrization. Carbon is green, nitrogen in red, sulphur is yellow and hydrogen is brown. The maximum variation is 0.11 and minimum is 0.0007 times an electrons charge.

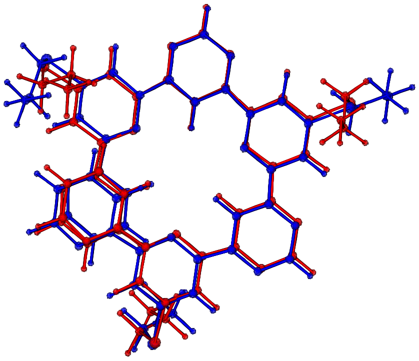


Figure 4.23: The global minima corresponding to two different sets of charge parameters.

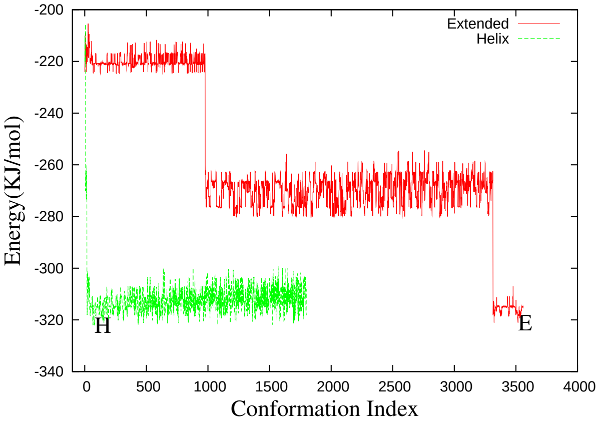


Figure 4.24: The minima hopping search profile starting from the same conformation using two different sets of charge parameters - one is "EXTENDED" which is parametrized on a extended conformation and the other is "HELICAL" which is parametrized on a previously found helical conformation. The global minima for the "EXTENDED" and the "HELICAL" tagged parameter sets were found at the points "E" and "H" respectively.

Chapter 5

Protein Structure Prediction (PSP) within All Atom Framework

Protein folding has attracted researchers from many fields because it is a fundamental and challenging problem. It is of major pharmaceutical importance as many diseases, including Alzheimer's, Bovine Spongiform Encephalopathy, Huntington's and Parkinson's, are believed to be caused by mis-folded proteins. Protein folding is complementary to the problem of ab-initio protein structure prediction whose aim is the determination of proteins' three-dimensional structures from their amino acid sequences. Accurate methods of prediction would be of great help in coping with the large number of sequences being produced with modern large-scale DNA sequencing efforts.

The experimental and theoretical study of the folding process by many groups has helped to illuminate many aspects of the kinetics and thermodynamics of protein folding. However, despite the advances made by structure prediction groups and the regular monitoring of the CASP [173] community, de novo structure prediction of 'new folds' remains an unsolved problem. Computer simulation studies are an important supplement to experiment and can help fill the gaps in our knowledge about protein folding.

β-hairpin The simplest protein motif which involves hydrogen bonded pair of beta strands with a loop in the middle is called a β-hairpin. It is one of the most important structural elements in proteins. A β-hairpin is stabilized through a balance among the hydrophobic interaction, interstrand hydrogen bonding force and the force related to loss of chain entropy. Different folding mechanisms of β-hairpin have been proposed[204, 205]. A lot of scientific studies have been devoted in understanding the hairpin formation[206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 127].

The existing notions are that the most probable way of a hairpin formation is through the zipper mechanism or the hydrophobic collapse mechanism. Apart from these, there have been mentions of several other mechanisms, like simultaneous zipping and collapse [215, 230] or an initially formed α-helix mediated mechanism[239] or reptation[183]. Munoz et al.[204, 205] suggested a 'zip-out' model from the experiments of the C-terminal β-hairpin of the B1 domain of protein G - where it was assumed that the folding initiates at the turn and propagates toward the tails by forming the interstrand hydrogen bonds sequentially and the hydrophobic cluster being formed later. This mechanism was later supported by both experiment and simulation evidences (Du et al., 2004; Kolinski et al., 1999; Mousseau et al., 2004; Zhang et al., 2006). However, Munoz et al. (Munoz et al., 1997) indicated that this is just the most probable way to form hairpin structure and other mechanisms could play a role too. They mentioned a 'zip-in' mechanism with two ends approaching each other to form a loop. The zip-out mechanism has been reported to be more probable than a zip-in mechanism. Dinner et al. (Dinner et al., 1999) proposed a 'middle-out' model. This model suggests that the folding proceeds by forming a partial hydrophobic cluster and then the hairpin hydrogen bonds propagate outwards in both directions from the partial cluster.

We can see that there are multiple opinions about the folding mechanism of β-hairpins. Experiments strongly support the zip-out model, while most simulations prefer the hydrophobic or zip-in model. One of the reasons may be that all-atom level simulations usually did not observe enough folding events. Another reason is that the experiments mainly observed the most probable pathway. Due to the requirement of cooperation of side chains of residues, it is difficult to simulate the complete folding of β-hairpin at all-atom level with standard molecular dynamics (MD) simulation method [227] alternate approaches have to adopted.

In the section 5.2 the Minima Hopping Algorithm (MHOP) to find global minima on potential energy surfaces is used for protein structure prediction. The energy surface of the protein is represented with an all-atom OPLS forcefield and an implicit free-energy solvation term. The system we studied here is the small 10-residue β-hairpin mini-protein, chignolin. Starting from a completely extended structure we found minima with < 0.5 Å RMS coordinate deviation from the geometry-optimized native experimental conformation. A few lowest energy conformations were used for the calculation of NMR-restraint violations and chemical shifts and the local minima found during each run leading to the global minimum were connected to trace out a search pathway of the folding process.

Next, in the section 5.3 one of the tryptophan zippers has been studied under the regular schemes of minima hopping method. We could find conformational minima which are very close to the experimental structure. The next attempt was to compare the minima hopping search process with a regular molecular dynamics simulation of folding. We investigated how pertinent is such a global optimization search trajectory in understanding a realistic folding pathway and a folding mechanism. The results show that one of the mechanisms suggested by minima hopping search for folding this β-hairpin is very similar to the one observed in our molecular dynamics simulation. An early intermediate state with a helical part between THR:3 to TRP:9 is prevalent in both the studies.

5.1 Model and Method

5.1.1 Forcefield

All-atom forcefield models with accurate representations of hydrogen-bonding and hydrophobic interactions are indispensable for high-resolution 3D-structure prediction [190] and for the refinement of structural models obtained from comparative modelling and fragment assembly-based structure prediction. Here, we used the OPLS all-atom force field (OPLS-AA) [102] for the potential energy of the biomolecular system in conjunction with the Generalized Born/Surface Area free-energy implicit solvation model of Still and co-workers (GB/SA) [192] as implemented in the DYNAMO modeling library [103, 104].

The free energy of the solvated system (without the protein backbone entropy contribution) is defined as follows:

$$\begin{aligned} E_{\text{config}} &= E_{\text{bond}} + E_{\text{angle}} + E_{\text{torsion}} + E_{\text{impropers}} \\ &\quad + E_{\text{nonbond}} + E_{\text{solvation}}, \end{aligned}$$

where:

$$E_{\text{bond}} = \sum_{\text{bonds } \ell} K_\ell (\ell - \ell_{\text{eq}})^2,$$

$$E_{\text{angle}} = \sum_{\text{angles } \theta} K_\theta (\theta - \theta_{\text{eq}})^2,$$

$$E_{\text{torsion}} = \sum_{\text{dihedrals } \Phi} \sum_{n} \frac{V_n}{2} [1 + \cos(n\Phi - \gamma_n)],$$

$$E_{\text{impropers}} = \sum_{\text{Impropers } \omega} \sum_{n} \frac{V_n}{2} [1 + \cos(n\omega - \gamma_n)],$$

$$E_{\text{nonbond}} = \sum_{i<j} \left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\epsilon r_{ij}} \right].$$

Here, E_{bond}, E_{angle}, E_{torsion} and $E_{\text{impropers}}$ represent the bond-stretching term, the bond-bending term, the torsion-energy term and the improper dihedral term, respectively. The non-bonded energy in Eq. (5.1.1) is represented by Lennard-Jones and Coulomb terms between pairs of atoms, i and j, separated by the distance r_{ij}. The parameters A_{ij} and

B_{ij} in Eq. (5.1.1) are the coefficients for the Lennard-Jones term, q_i is the partial charge of the i-th atom in the electrostatic term, and ϵ is the dielectric constant.

$$E_{\text{solvation}} = E_{vdW} + E_{cavity} + E_{pol}.$$

$E_{\text{solvation}}$ is the solvation free-energy term, in which E_{pol} is the solute-solvent electrostatic polarization term, E_{cavity} is the solvent-solvent cavity term and E_{vdW} is the solute-solvent van der Waals term. E_{pol} is calculated using the generalized Born equation :

$$E_{pol} = -166.0(1 - \frac{1}{\epsilon}) \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{q_i q_j}{(r_{ij}^2 + \alpha_{ij}^2 e^{D_{ij}})^{0.5}}$$

where $\alpha_{ij} = (\alpha_i \alpha_j)^{0.5}$, $D_{ij} = \frac{r_{ij}^2}{(2\alpha_{ij})^2}$ and the double sum runs over pairs of atoms, i and j. α_i is the Born radius of atom i which is calculated analytically using the approximation in reference [192]. Non-bonding interactions between all pairs of eligible atoms are calculated without cutoffs.

Because saturated hydrocarbons are non-polar molecules (for which $E_{pol} \sim 0$) and their $E_{\text{solvation}}$ in water is approximately linearly related [193] to their solvent accessible surface areas (SA), the GB/SA method approximates the remaining two terms as:

$$E_{vdW} + E_{cavity} = \sum_{k=1}^{N} \sigma_k SA_k$$

Here SA_k is the solvent-accessible surface area for atom k and σ_k is an empirically determined atomic solvation parameter. The latter were taken to have the value 30.5 J mol^{-1} Å^{-2} for all atom types and the probe radius for calculating the solvent-accessible surface was 1.4 Å.

Removal of the dihedral isomers The value of the energy function should not depend on the labeling or numbering of atoms which are of the same element and are equivalently placed in the molecule, e.g. the O and OXT atoms in the GLY10 residue in 1UAO. However, due to the usual way of defining improper dihedral terms, an interchange of the index of the atoms can give rise to two isomers with a small energy difference (of $\sim 10^{-3}$ kJ mol^{-1}) upon a tight geometry optimization [194, 195]. To avoid this problem, we replaced each improper term, I–J–K–L, between a planar atom (K) and its three bound atoms (I, J, L), with three terms which are a cyclic permutation of I, J and L (i.e I–J–K–L, L–I–K–J and J–L–K–I). Each improper permutation term is the same as the other two and has a force constant one third that of the case in which a single improper term is used per planar atom.

5.1.2 Simulation Method

We used Minima hopping [124] algorithm with the run parameters described in the section 3.1. As the potential energy surface of biomolecular systems is complex and very rugged, we adopted a few additional modifications to our original algorithm. These are as follows:

- *Checking for the chirality changes.* A very high temperature MD simulation can change the chirality of certain protein groups whereas folding preserves chirality. To prevent this, we devised a filter which checks the chirality of all the residues and rejects chirality-changing moves immediately. In this way the global minimum search is restricted to the biologically-relevant free energy surface of the atomistic system.
- *Searching along slow vibrational modes.* We used here a scheme described in section 3.1.2 for searching in slow vibrational modes taking out the component of the initial random velocity which is in the space of all bond stretching and bond bending vibrations.

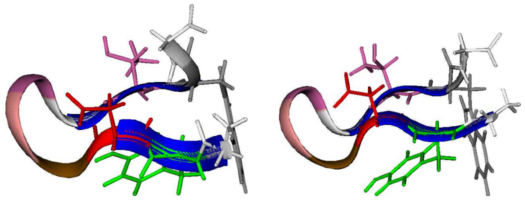
5.2 Chignolin (1UAO)

As a target for our optimizations, we employ chignolin (PDB code 1UAO) [188, 189] which forms a stable β-hairpin in aqueous condition and is composed of 10 amino-acid residues (GYDPETGTWG). Chignolin is a de novo protein that has been designed to have a unique, stable structure and to fold quickly so that it can be studied efficiently both experimentally and theoretically. Our simulations employed the all-atom OPLS forcefield with an implicit free-energy solvation model and a parallelized version of the minima hopping program. A long parallel run was used for our analysis. From the minima search, we found the lowest free energy structure, which has a RMS coordinate deviation (RMSD) less than 0.5 Å from the geometry-optimized experimental conformation, and determined a pathway for the folding process. The hydrogen-bonding network and NOE-violations of the lowest energy conformations were compared to those of the native conformation and the conformations were also clustered using a fixed radius method based on mutual RMSD and a fixed cluster radius of 3 Å. Finally we re-calculated the energies of the whole conformational ensemble using all-atom, distance-dependent, pairwise statistical energy functions based on a Distance-scaled, Finite-Ideal gas REference(DFIRE) and compared them with the OPLS energies.

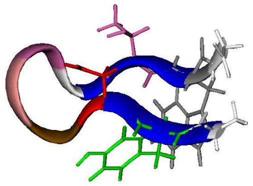
5.2.1 Results of MHOP Simulation

We performed a number of MHOP simulations. In a preparatory stage, five independent simulations starting from the stretched conformation of chignolin and a single simulation starting from the native conformation were performed. Each of these were a parallel run on several processors. After this, we performed a single large simulation, starting from the stretched conformation for further analysis. The results from all simulations were broadly consistent and each found the same lowest-energy structure which is shown in fig. 5.19. It consists of a β-hairpin having a C_alpha-RMSD with respect to the native conformation of only $0.37\AA$. The minimum RMSD structure found during the simulation had a C_alpha-RMSD of 0.11 Å, whereas the geometry-optimized native conformation had an energy 47 kJ mol^{-1} higher than the lowest-energy conformation.

Due to the similarity of the results of the different simulations, we concentrate our analysis on the results of the final long run in what follows.



Native Native Optimized



Lowest Energy

Figure 5.1: Images of the native, native optimized and MHOP lowest energy configurations. (The figures were prepared using the program VMD [130])

MHOP simulations consist of a series of short MD simulations — in this case ∼ 500 steps or 0.5 ps — followed by a geometry-optimization step. We considered the optimizations converged when the RMS gradient tolerance fell below 10^{-4} kJ mol^{-1} Å^{-1} and it is a step that took most of the CPU time during the simulation. On average ∼ 50% of the geometry optimizations gave rise to new local minima of which ∼ 50% were accepted by the MHOP algorithm — i.e. ∼ 25% of the total number of geometry optimizations. In each run there were ∼ 5000 geometry optimizations per processor meaning each processor generated ∼ 1000 distinct local minima. In each MHOP simulation, the lowest-energy structure was found by only a small number of processors (usually one). This is because the algorithm employs the shared search history of all processors to penalize the exploration of already visited parts of the PES and also perhaps because we need more efficient ways of generating new protein conformations so as to be able to escape trapping in specific regions of conformational space. These results emphasize the importance of several independent MHOP simulations so that the nature of the lowest-energy structure can be verified. For completeness, the remaining parameters for our MHOP simulations were (definitions may be found in reference [124]): $\beta_1 = \beta_2 = \frac{1}{\beta_3} = 1.05$; $\alpha_2 = \frac{1}{\alpha_1} = 1.02$ and mdmin $= 2$.

During the MHOP simulation, the initial temperature of each MD simulation is tuned according to the search history and whether previously unvisited minima are found or not. Fig. 5.2 is a plot of the initial temperature of the MD simulations and the energies of the local minima that the simulations lead to.

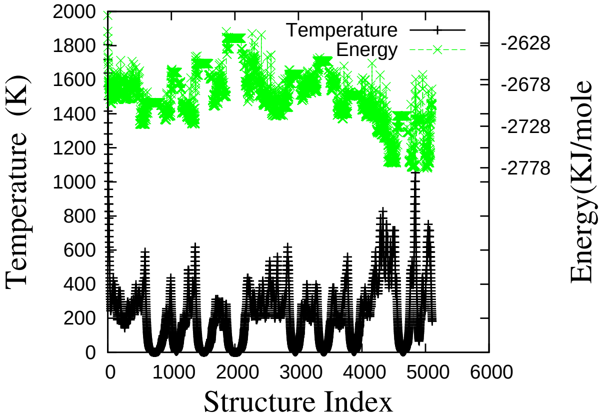


Figure 5.2: The initial MD simulation temperatures and the energies of the resulting local minimum during the search process.

A total of $\sim 33\,000$ distinct local minima was found by all the processors. The *C_alpha*-RMSD of the whole ensemble of structures was calculated and plotted against the energy of the structures in fig. 5.3. From the conformational minima the energy-based structural ranking and the RMSD-based ranking are not identical, i.e. being lower in RMSD does not ensure being lower in free energy. However, the minimum free energy conformation among all the native-like conformations is lower in free energy than all other conformations so the funnel containing the native-like conformations contains the free energy minimum.

Energy profile

From the conformational ensemble the free energy landscape was drawn as a 2-D contour map using two different pairs of collective variables as abscissa and ordinate. The first set of variables are the *C_alpha*-RMSD($\mathring{A}$) and the radius of gyration (R_g in $\mathring{A}$) of all residues (see fig. 5.5) whereas the second set consists of the CM–to–CM (center of mass) distances between the residue pairs TYR2–TRP9 and PRO4–GLY10 (see fig. 5.4). To determine each map, a 2-D grid was constructed and the minimum free energy of all structures falling within each grid point was assigned to be the grid point's free energy.

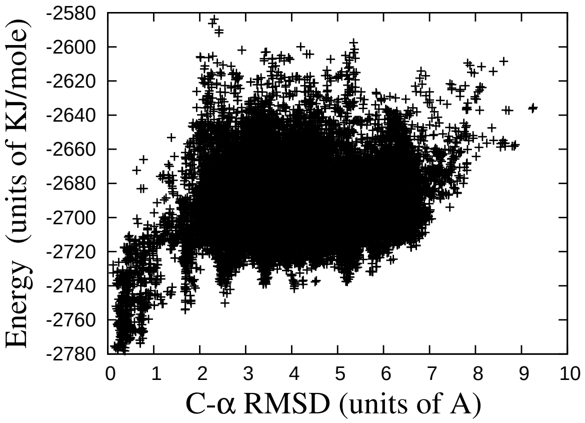


Figure 5.3: Energy versus C_alpha-RMSD with respect to the native conformation.

The native ensemble (dark blue color) is clearly visible in all the contour plots. Other prominent free energy wells also exist (e.g. in fig. 5.5 at R_g ~5.3$\mathring{A}$ and RMSD ~1.7$\mathring{A}$), some of which might serve as stable intermediates during the folding process.

The protein's free energy landscape very much depends on the forcefield and solvation model used. To test this variation, we employed two other energy functions to characterize the energy landscape further. These were: (i) DFIRE, an all-atom distance dependent, pairwise energy function based on a Distance-scaled, Finite-Ideal gas REference(DFIRE) [197]; and (ii) dDFIRE [198] (dipolar DFIRE). Surfaces recalculated with these functions are shown in fig. 5.6. Although there are differences, the overall features of the two surfaces are broadly similar to that determined with the OPLS-AA potential.

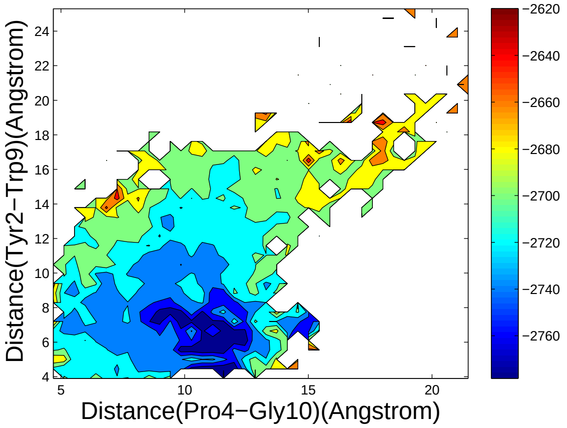


Figure 5.4: Energy landscape as a function of the distances between the residue pairs TYR2-TRP9 and PRO4-GLY10

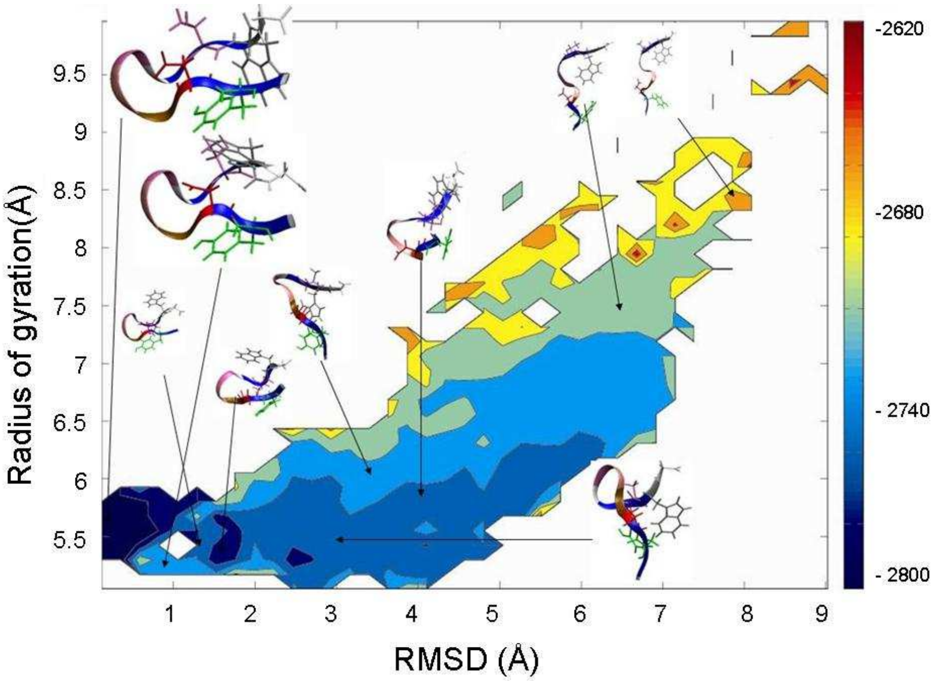
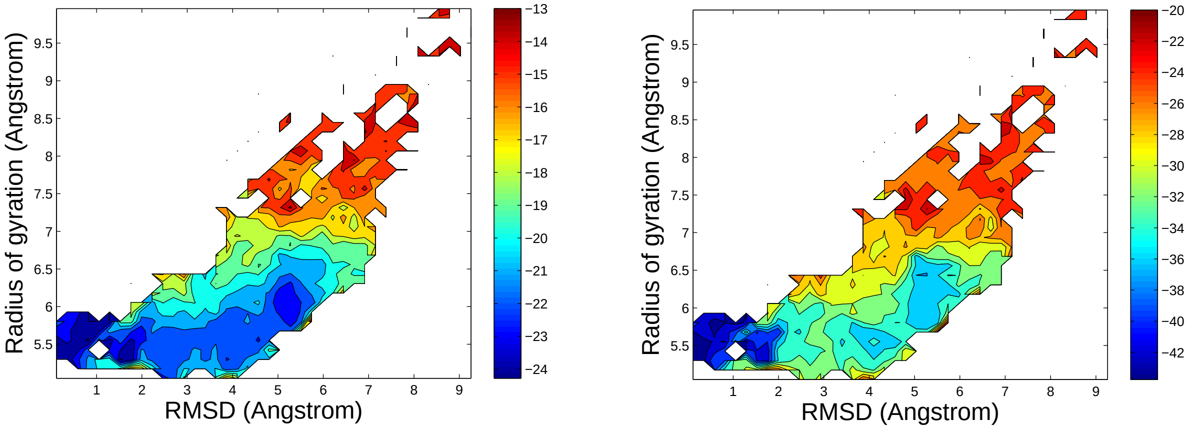


Figure 5.5: Energy landscape as a function of R_g and C_alpha-RMSD. Certain specific conformations are shown on the landscape.



DFIRE energy dDFIRE energy

Figure 5.6: Redrawn landscapes

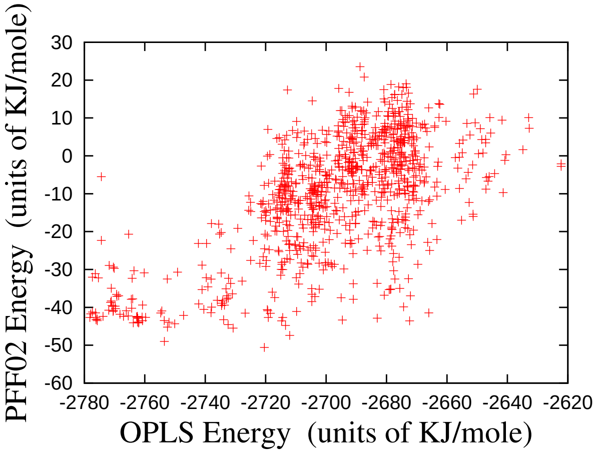


Figure 5.7: Energetic cross-ranking : OPLS vs PFF02

OPLS, AMBER, DFIRE, PFF02 and ECEPP/3 The lowest 1000 of the conformations have been optimized in PFF02 forcefield implemented in POEM software package. The fig.5.7 shows the energetical correlation between these two forcefields. We compared in fig.5.10the structural ranking for the lowest 5000 conformations of our search with ECEPP/3 forcefield energies after we optimized the geometries in that forcefield implemented in SMMP software package. Also in fig. 5.9 we showed the energetic correlation between OPLS forcefield and the DFIRE forcefield. We reoptimized those 5000 lowest OPLS energy local minima using AMBER94 forcefield implemented in NAB package [191]. The SASA parameter was 0.0073 $KCal/\mathring{A}^2$ which is same as was used in the minima hopping runs with OPLS forcefield(section 5.1.1). In fig.5.8 we showed the energetic correlation between OPLS forcefield and the AMBER94 forcefield. The OPLS and ECEPP/3 forcefields seems not to be agreeing with each other. The PFF02 forcefield also indicates some disagreement with OPLS forcefield in terms of the accurate conformational

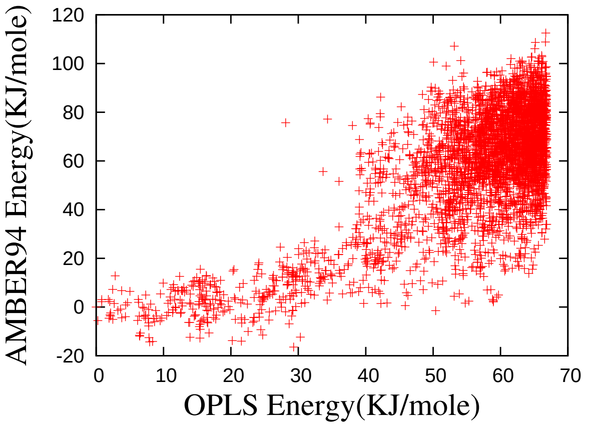


Figure 5.8: Energetic cross-ranking : OPLS vs AMBER94

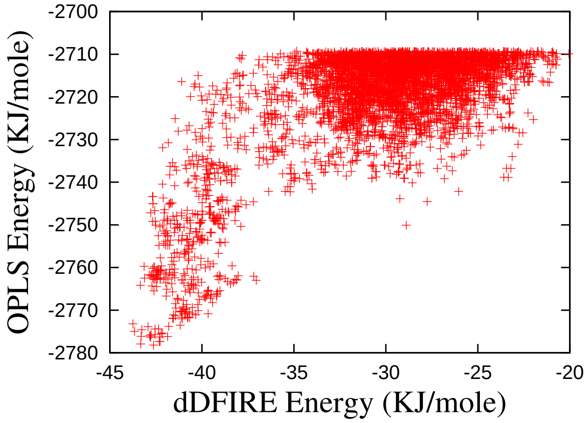


Figure 5.9: Energetic cross-ranking : OPLS vs DFIRE

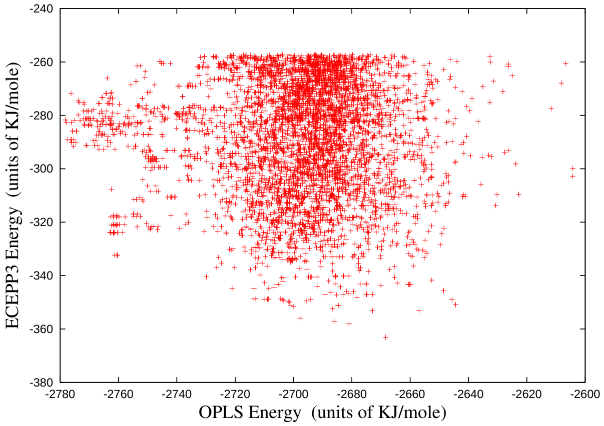


Figure 5.10: Energetic cross-ranking : OPLS vs ECEPP/3

ranking. But on a larger scale DFIRE and PFF02 agrees in defining a energetically low region around the native conformation.

Pathway connecting the local minima

Using all the minima found during a simulation, it is possible to build the shortest path which starts from the initial stretched conformation and ends up at the lowest energy structure. Starting from the final minimum we trace backwards and cut off any part of the search process that forms a closed loop (i.e. due to the occurrence of already visited minima). This path serve as a description of the MHOP search pathway. Energy profiles with the different force fields are plotted in figs. 5.11 and 5.12, and plots of the pathway structures on the free-energy contour maps are shown in figs. 5.13 and 5.14.

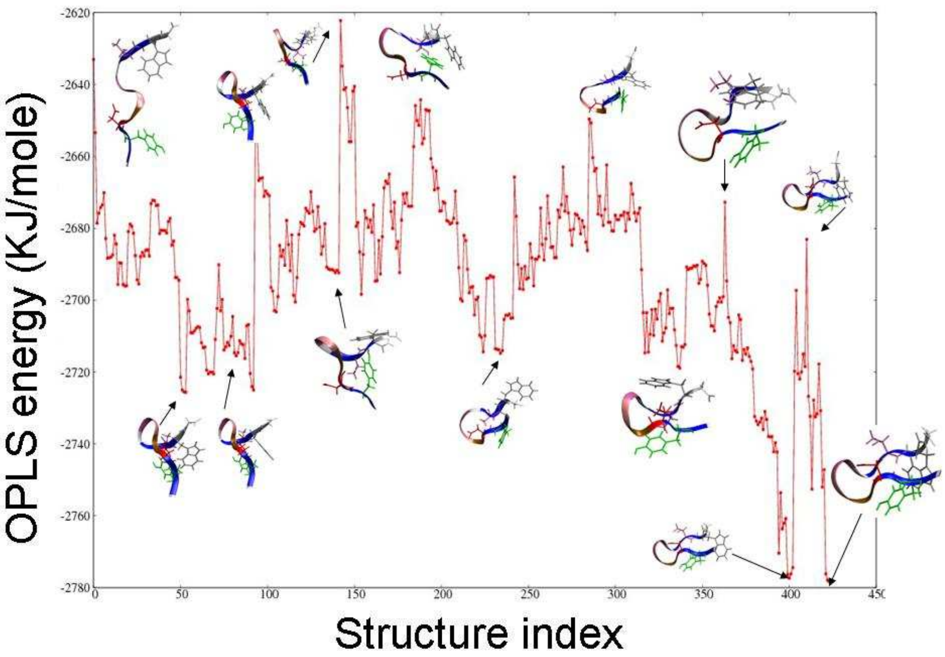


Figure 5.11: OPLS energy profile along the MHOP-search-pathway

Clustering of the conformations

The clustering method employed here is a fixed-radius clustering in which the RMSDs of the conformations within a cluster to its centroid are less than the given radius. The method is such that the number of clusters found cannot be limited beforehand. The

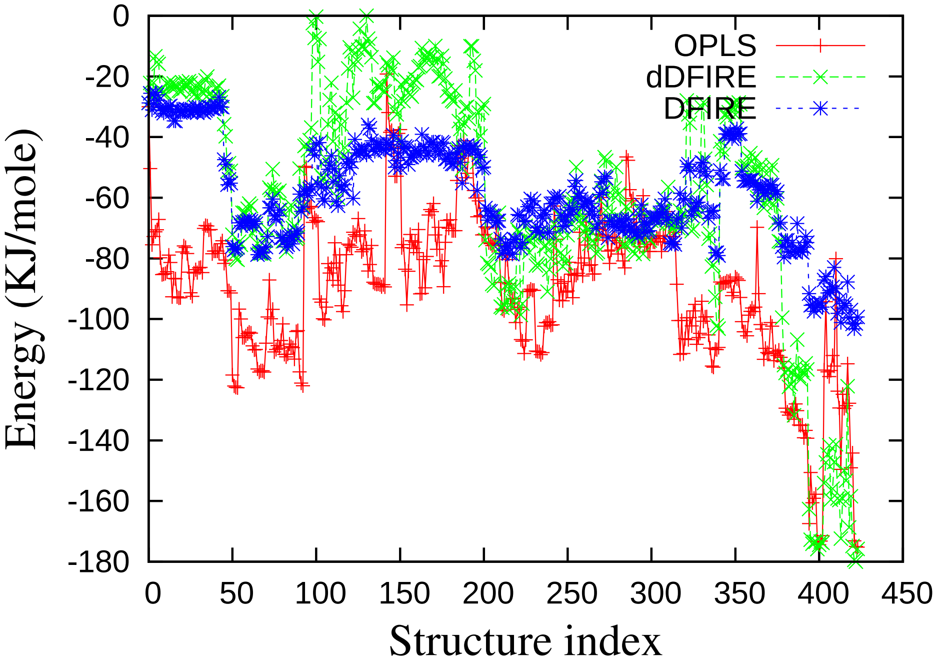


Figure 5.12: Energy profile along the MHOP-search-pathway using different force fields.

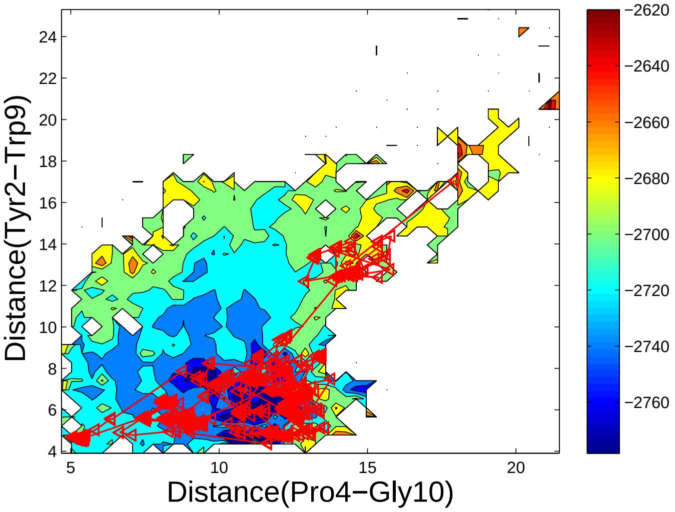


Figure 5.13: The MHOP-search pathway superimposed on the energy landscape obtained as a function of the distances between the residue pairs TYR2-TRP9 and PRO4-GLY10.

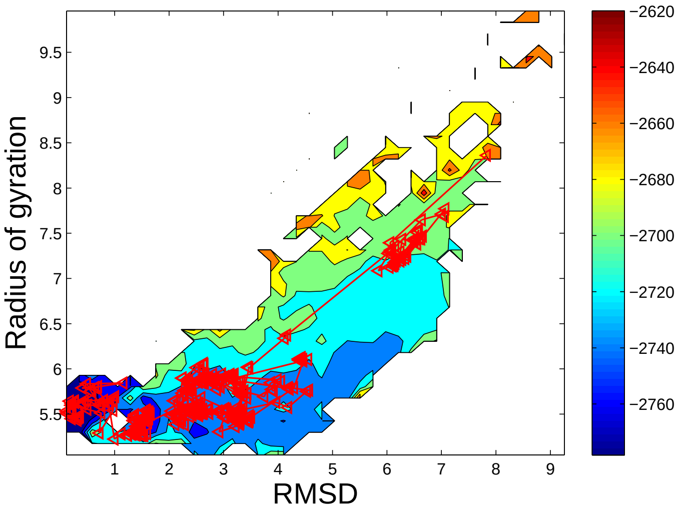


Figure 5.14: The MHOP-search pathway superimposed on the energy landscape obtained as a function of R_g and C_alpha-RMSD.

MMTSB [131] toolset was used for the calculation. We took the top 13 clusters according to the cluster size and another 7 clusters from the ranking based on C_alpha-RMSD to the native conformation (without geometry optimization). The results in table 5.1 were prepared by sorting the 20 clusters according to the C_alpha-RMSD of the cluster centroid with respect to the geometry optimized native conformation. The secondary structure content for the cluster-centroids was calculated using the DSSP program. Cluster.154 contains the lowest energy structure and the centroid of that cluster is at $\sim$ 0.24 Å C_alpha-RMSD to the geometry-optimized native conformation. Cluster.169 has the lowest C_alpha-RMSD ($\sim$ 0.82 Å) to the un-optimized native conformation (C_alpha-RMSD between the native conformation and the optimized native conformation is $0.98\mathring{A}$). Fig. 5.15 shows the minimum energy of the conformations within each cluster plotted against the cluster size.

Comparison with experiment

NOE violation Protein structures determined by NMR spectroscopy and submitted to the Protein Data Bank[132] are actually calculated using a forcefield and a simulation protocol which utilizes experimental data as restraints. An elegant way of checking the quality of a structural model is to calculate how these restraints are violated. The RMS

Cluster name	Cluster size	Energy (kJ mol^{-1})	C-α RMSD ($\mathring{A}$)	Secondary structure (DSSP)
Cluster.154	459	-2778.2	0.24	UEETTTTEEU
Cluster.169	120	-2770.8	0.68	UEETTTTEEU
Cluster.189	52	-2754.0	1.27	UEETTTTEEU
Cluster.72	88	-2718.6	1.36	UUUSSSTTUU
Cluster.199	62	-2727.3	1.58	UBUSSSSSBU
Cluster.149	61	-2718.9	1.61	UUBTTTBSUU
Cluster.212	130	-2753.6	1.64	UEETTTEEUU
Cluster.241	23	-2663.0	2.09	UUUTTTUUUU
Cluster.95	484	-2717.6	2.48	UUUUSSSSUU
Cluster.145	580	-2728.1	2.83	UUUUSSSUUU
Cluster.65	657	-2728.6	2.84	UUUUSSSSUU
Cluster.55	682	-2710.3	3.04	UUUUUSSUUU
Cluster.129	592	-2725.0	3.17	UUUUSSSSUU
Cluster.120	686	-2739.0	3.32	UUSUTTTTUU
Cluster.9	1092	-2707.8	3.36	UUUUUSSUUU
Cluster.32	509	-2724.7	4.20	UUUUUTTTUU
Cluster.67	863	-2737.6	4.54	UUUUUBTTBU
Cluster.38	1044	-2734.9	5.13	UUUUUBTTBU
Cluster.63	531	-2721.9	5.16	UUUUUBTTBU
Cluster.66	593	-2716.4	6.34	UUUUUUSSUU

Table 5.1: Clusters obtained from cluster analysis. Included are their size, RMSD of their centroids to that of the geometry-optimized native structure, the minimum energy among the conformations and secondary structure details. The notations are as follows : U $\equiv$ not defined , E $\equiv$ extended strand participated in $\beta-$ladder, T $\equiv$ H-bonded turn, S $\equiv$ bend, B $\equiv$ residue in isolated $\beta-$bridge.

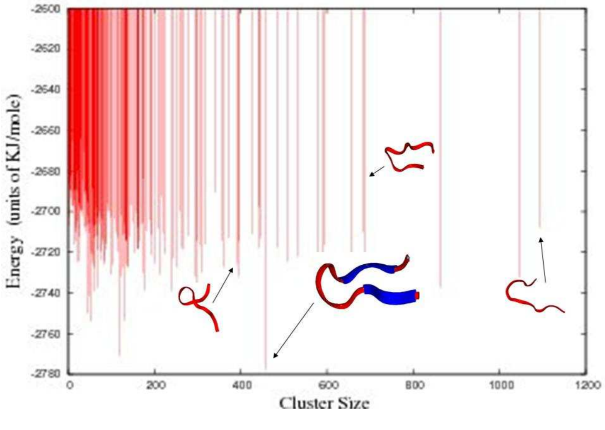


Figure 5.15: Lowest energy of the member conformations against the size among the clusters.

NOE deviation is defined as:

$$\text{RMS}_{\text{NOE}} = \sqrt{\frac{1}{N_r.N_m}\sum_{k=1}^{N_r}\sum_{l=1}^{N_m}(\delta_{kl})^2}$$

Here N_m is the number of structural models, N_r is the number of distance restraints and d_{kl} is the distance of the kth restraint in the lth conformation which is defined as:

$$\delta_{kl} = \begin{cases} d_{kl} - r_k^{upper} & d_{kl} > r_k^{upper} \\ r_k^{lower} - d_{kl} & d_{kl} < r_k^{lower} \\ 0 & \text{otherwise} \end{cases}$$

where r_k^{upper} and r_k^{lower} are the upper and lower bounds on the kth restraint, respectively.

Fig. 5.16 shows the RMS-NOE versus energy profile. The lowest energy conformation has an RMS NOE of $\sim$ 0.9 Å, whereas the conformation having a minimum NOE-violation is $\sim$ 2.1 kJ mol^{-1} higher in energy than the lowest energy conformation.

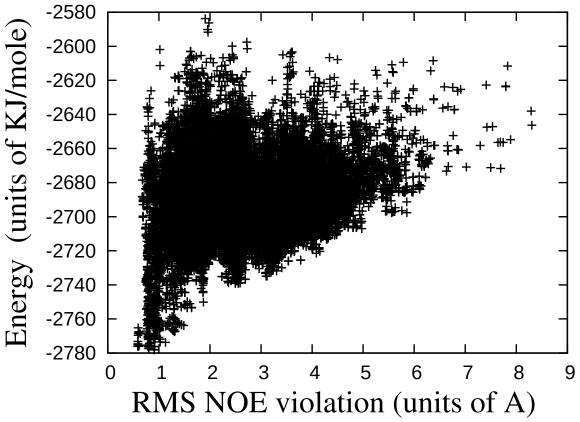


Figure 5.16: Energy of the conformations vs their RMS NOE-violation.

Chemical shift Chemical shift deviations from random-coil values were calculated using the SHIFTS 4.1 web-server [199] for five structures comprising the experimental structure, the lowest-energy conformation, two structures from clusters very close to the native conformation and another from a cluster representing a misfolded state. The results are shown in (in fig. 5.17). The chemical shift deviation for TYR2, TRP9, GLY10 residues for the lowest energy conformation look to be quite different from the experimental values. The ring current effect between TYR2 and TRP9 is not reflected in the conformations resulting from the simulation as prominently as in experiment. This discrepancy is most probably due to the accuracy of the force-field parameters. In contrast, the difference in the value of the chemical shift deviation GLY10 is most likely caused by the flexibility that this C-terminal residue exhibits.

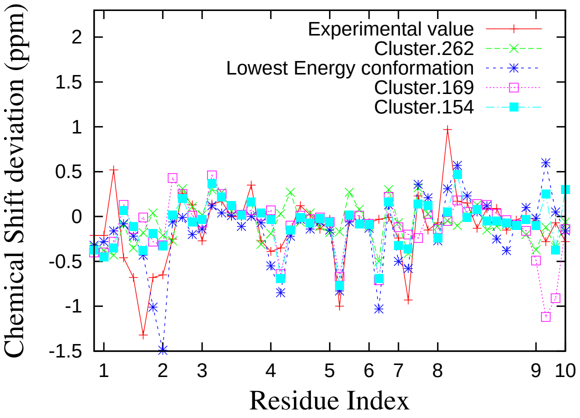


Figure 5.17: Deviation of chemical shift values from those of a random coil.

Hydrogen-bonding analysis The OPLS-AA forcefield represents hydrogen bond interactions as an appropriate balance of electrostatic and Lennard-Jones non-bonding terms and not with an explicit energy term. Therefore, to identify hydrogen bonds, we looked at all possible pairs of hydrogen bond donors and acceptors and selected those which had a length less than 3 Å and a donor–hydrogen–acceptor angle of more than 130°.

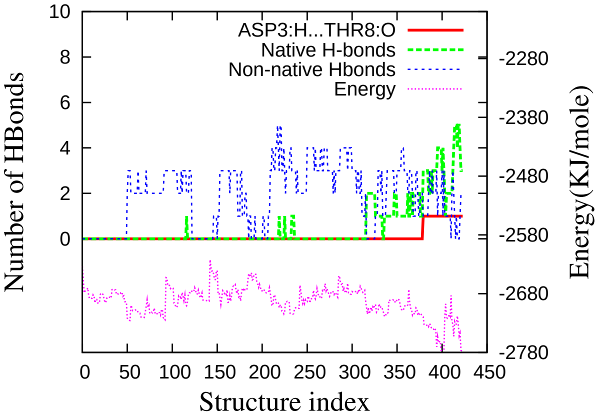


Figure 5.18: Number of Native and Non-native hydrogen bonds along the MHOP-search-pathway.

Fig. 5.18 shows the making and breaking of native and non-native hydrogen bonds along the search pathway. Clearly the ASP3:H–THR8:O hydrogen-bond is the signature of the funnel on the free-energy surface containing the native state. From the same figure, it can be seen that strong non-native hydrogen bonds can form during the early stages of the search pathway. TYR2:H–GLY7:O, GLU5:H–THR8:O and ASP3:H–GLY7:O are some of the non-native interactions that have to be broken before creation of the native hydrogen-bond network.

5.3 Tryptophan Zipper (1LE1)

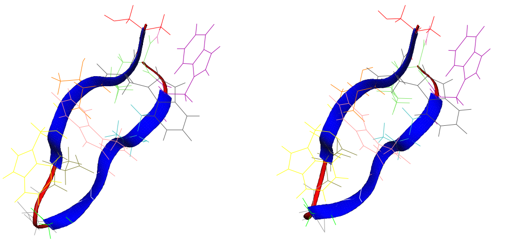
In this study, in the first part, we use the Minima Hopping Algorithm(MHOP) [124] for structure prediction. As a target for our optimizations, we employ trpzip2 (PDB code 1LE1) [188, 189] which forms a stable β-hairpin in aqueous condition and is composed of 12 amino-acid residues (SWTWENGKWTWK). Trpzip2 is a de novo protein that has been designed to have a unique, stable structure and to fold quickly so that it can be studied efficiently both experimentally and theoretically. Its native -hairpin structure has an obvious hydrophobic core composed of two aromatic side-chain pairs and so is similar to the experimentally investigated C-terminal -hairpin of the B1 domain of protein G. The strong hydrophobic interactions of two aromatic side-chain pairs make the trpzip2 more stable and easier to simulate its folding dynamics by simulation. Therefore, trpzip2 is a very ideal model for investigating the general folding mechanisms of -hairpins. Our simulations employed the all-atom OPLS forcefield with an implicit free-energy solvation model and a parallelized version of the minima hopping program. A long parallel run was used for our analysis. From the minima search, we found the lowest free energy structure, which has a RMS coordinate deviation (RMSD) less than 0.79 Å from the experimental conformation and determined a pathway for the folding process. The hydrogen-bonding network of the lowest energy conformations was compared to those of the native conformation and the conformations were also clustered using a fixed radius method based on mutual RMSD and a fixed cluster radius of 3 Å. Next, we re-calculated the energies of the whole conformational ensemble using an all-atom free energy forcefield PFF02[201, 202] and compared them with the OPLS energies. From the MD simulation part of the study, inspite of the short length of MD simulation (40 ns) we observed a few folding events. We reported here one such event where the folding happened after 4 ns through a hydrophobic collapse mechanism. We compared this pathway with minima hopping search trajectories. It was interesting to observe on a comparative note the sequence of microcanonical MD moves of minima hopping against a continuous trajectory of canonical MD both starting from a stretched conformation and leading to the folded state of the peptide.

5.3.1 Results of MHOP Simulation

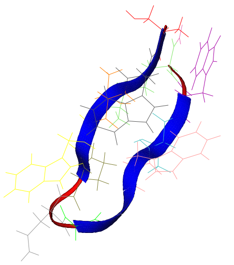
We performed a number of MHOP simulations. In a preparatory stage, five independent simulations starting from the stretched conformation of trpzip2 and a single simulation starting from the native conformation were performed. Each of these were a parallel run on several processors. After this, we performed two large simulations(R1 and R2), starting from the stretched conformation for further analysis. The results from all simulations were broadly consistent and each found the same lowest-energy structure.

Due to the similarity of the results of the different simulations, we concentrate our analysis on the results of the final long run in what follows.

Minimum Energy Conformations



Native Native Optimized



Lowest Energy

Figure 5.19: Images of the native, native optimized and MHOP lowest energy configurations. (The figures were prepared using the program VMD [130])

The lowest energy conformation is shown in fig. 5.19. It consists of a β-hairpin having a *C_alpha*-RMSD with respect to the native conformation of only $0.79\AA$. The minimum RMSD structure found during the simulation had a *C_alpha*-RMSD of 0.48 Å, whereas the geometry-optimized native conformation had an energy 120 kJ mol^{-1} higher than the lowest-energy conformation.

RMSD vs Energy

In the fig.5.20 we have shown the C_α-RMSD from the native conformation vs effective free energy for all the minima during one of the minima hopping runs. The lowest C_α-RMSD conformations have the lowest energies which are actually the native conformational ensemble. The black line in fig.5.20 is connecting each point corresponding to a local minimum found by a successful process of a parallel minima hopping run "R1".Bothe the runs

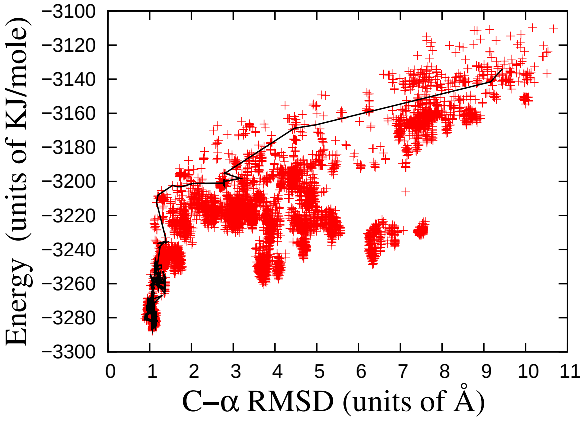


Figure 5.20: C_α-RMSD vs Energy plot. the black line connects all the minima those one of the successful minima hopping processes visited.

“R1” and “R2” found the same lowest minimum - but the intermediate minima and the search trajectories are different from each other.

Intermediate states

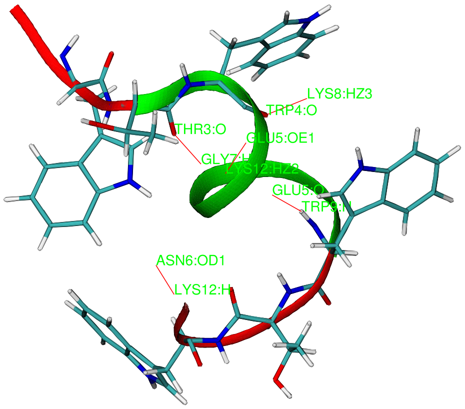


Figure 5.21: A probable intermediate state

In the run “R2” we found a partly-helical intermediate(fig. 5.21) having an energy 59 kJ mol^{-1} higher than the lowest-energy conformation. The helical part is between THR:3

to TRP:9 residues. This intermediate is found at a very early stage of the search. This partly-helical looks very similar to the intermediate found by Wenzel[127] using Basin Hopping Method with PFF02 forcefield.

Conformation energy map

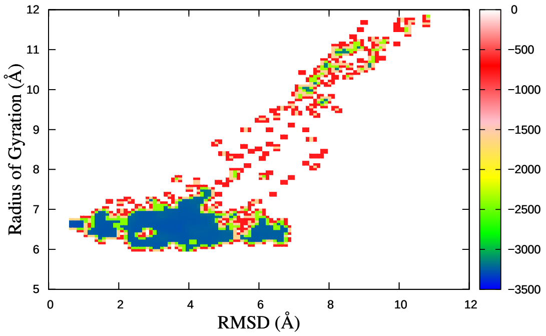


Figure 5.22: The effective free energy(OPLS) map generated from all the conformations during the minima hopping run.

Here in fig.5.22 the Conformational effective free energy map has been shown. The collective variables are C_α RMSD from the native conformation and geometric radius of gyration. The lowest energy conformation is at C_α RMSD $= 0.78\AA$ and Radius of Gyration $= 6.56\AA$ and the partly-helical intermediate is at C_α RMSD $= 5.05\AA$ and Radius of Gyration $= 6.65\AA$.

5.3.2 MHOP and MD

MD simulation Trp-Zipper

With the same forcefield as in section 5.1.1 we conducted Langevin Molecular dynamics at different temperatures ranging between 260 K to 410 K, with a step size of $2fs$, having the collision frequency with the bath to be 25/ps. Each of the simulations was 40 ns long. We collected coordinates every 2 ps. We chose the simulation trajectory corresponding to the temperature 280 K to analyze.

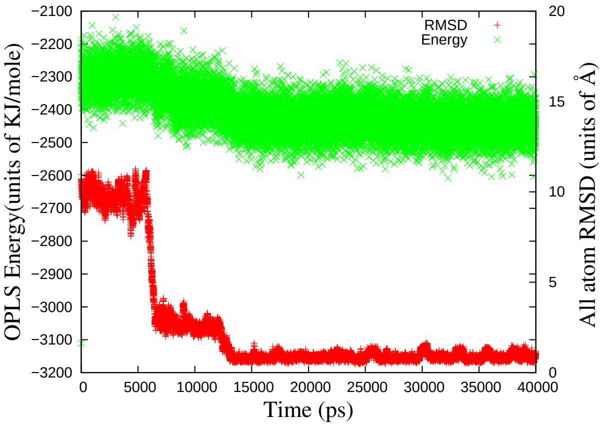


Figure 5.23: MD simulation : Energy and all-atom RMS-deviation from the native conformation are plotted against time.

Fig.5.23 shows the fluctuation in energy and all-atom RMS-deviation from the native conformation along the MD trajectory of temperature 280 K. Between 6 ns to 7 ns both the quantities show a jump which corresponds to a considerable amount of conformational transition - this is the hydrophobic collapse of the peptide chain. From this point it starts to form a stable hydrophobic core and subsequently the native hydrogen bond network to reach the folded state (at around 32 ns). In the fig.5.24 the energy and the all-atom RMS-deviation have been plotted against each other. The native-like conformations (which have a low RMSD) have low energies - it shows that the peptide had almost adirect folding - i.e it entered the native funnel quite fast and explored the low energy part of it quite well within the short period of <40 ns. This run managed to follow a fast folding pathway. The fast folding events are probabilistically rarer than the regular folding events. That explains why most of the other MD runs could not register any folding event - for which one has to sample for long.

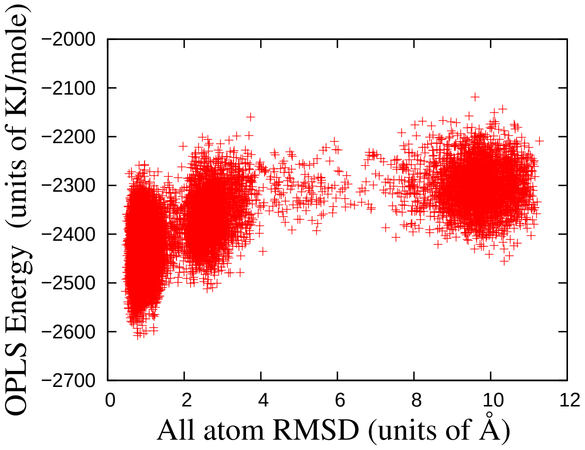


Figure 5.24: MD simulation : Energy vs all-atom RMS-deviation from the native conformation.

Local Minima along MD The energy of a snapshot from the MD trajectory cannot accurately quantify the energetic depth of the basin or the local minimum. To compare the progress along a MD trajectory and a minima hopping search pathway, next, we optimized the coordinates of each of the snapshots taken from the MD trajectory of fig.5.23 upto the force norm 10^{-5}. These energies are then the depth of the basins traversed during the MD simulation. In fig. 5.25 we compared molecular dynamics and minima hopping sampling algorithm in terms of the energies of all the cthe local minima those the trajectory passes through. The C_α-RMSD from the native conformation has also been plotted for all the snapshots from the MD trajectory - the same for the minima hopping search pathway(run of fig.5.20). We estimated that the maximum number of force calls for 1 MD move and 1 geometry optimization for this system in minima hopping would be $\sim$ 10000. From there we roughly calculated the length of a minima hopping trajectory in terms of time.

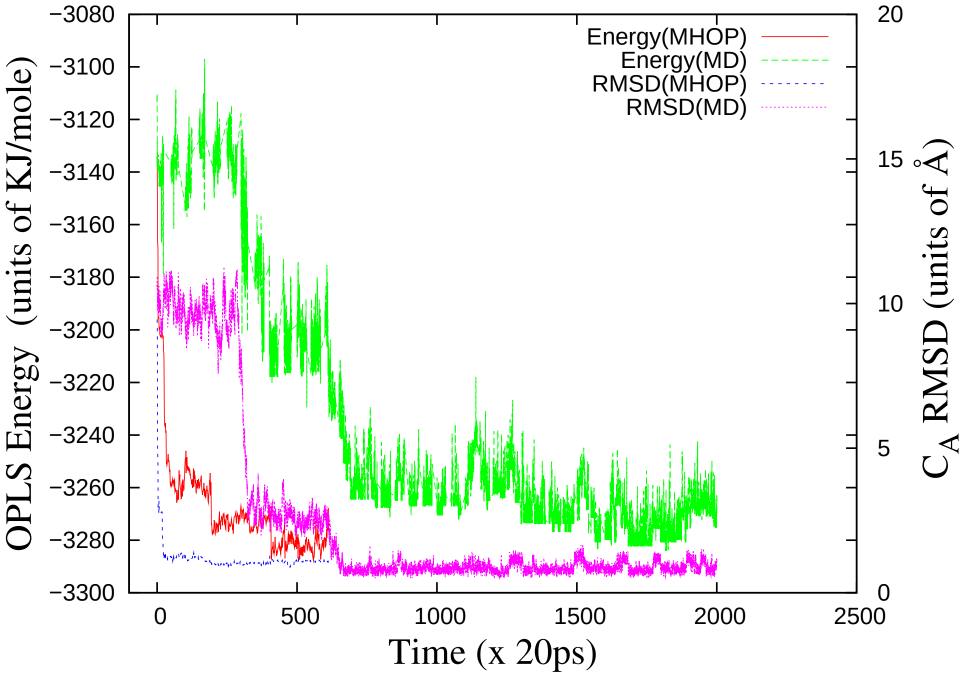


Figure 5.25: MD simulation and Minima hopping simulation compared.

MHOP search trajectory

From the minima hopping runs we could construct two different pathways - each of which suggests a particular mechanism of folding. The successful process of the simulation “R2” points out to the Zipper mechanism, whereas one successful process from the “R1” simulation suggests a “hydrophobic collapse” mechanism.

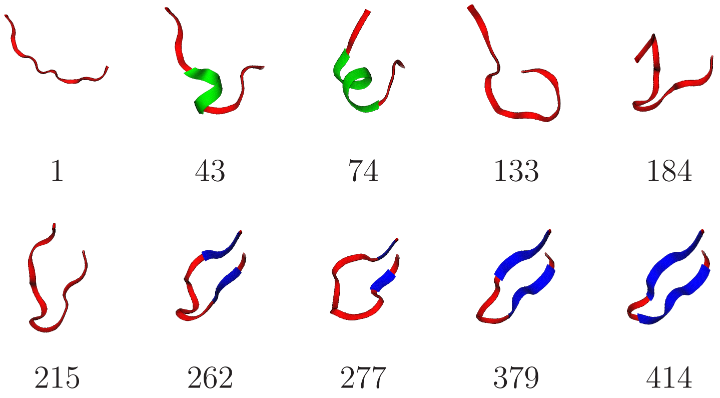


Figure 5.26: Minima hopping search trajectory : Zip-in pathway. This passes through a partly-helical intermediate (The figures were prepared using the program VMD [130])

Helix mediated Zipper mechanism The folding proceeds as follows: the peptide falls into a partly-helical structure first. It breaks subsequently. Next, two aromatic pairs Trp2-Trp11 and Trp4-Trp9 are formed but located between two -strands. This prohibits the formation of inter-strand hydrogen bonds. Next, the trpzip2 tries to adjust the aromatic pairs toward outsides. Next, starting from the most outside hydrogen bond (1-12) other hydrogen bonds form in the order: (1-12),(3-10, 10-3), (5-8, 8- 5). It follows a "zip-in" pathway.

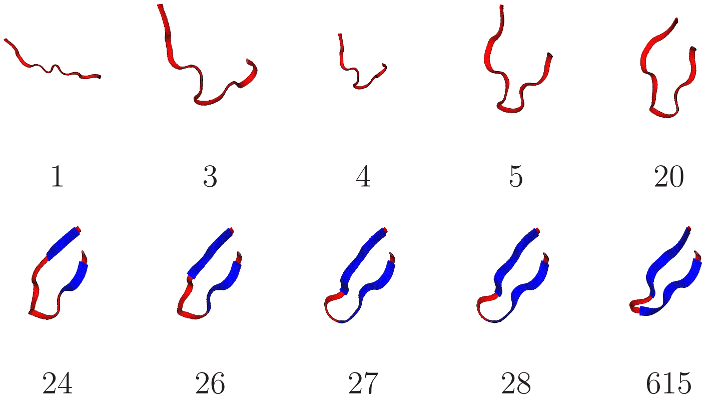


Figure 5.27: Minima hopping search trajectory : A non-zipper pathway - the folding is direct. (The figures were prepared using the program VMD [130])

"Hydrophobic collapse" mechanism In this case, the peptide quickly collapses into a hairpin-like structure similar to the native one. The hydrophobic contact between TRP:2 and TRP:9 is made. There at the center the hydrogen bonds LYS:1:HZ1-GLU:5:OE1 and TRP:9:HE1-THR:3:O are made. The TRP:2 and TRP:9 try to stabilize by changing the orientations of the aromatic rings and the hydrogen bond between LYS:12:O-SER:1:HT2 is made. The peptide adjusts its conformation and folds into the native state. The hydrophobic pairs and the native hydrogen bonds are formed almost simultaneously. This is not a zipper pathway.

MD pathway : Hydrophobic collapse

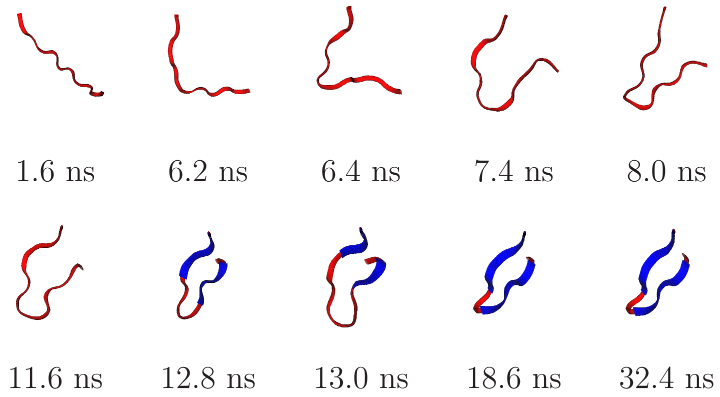


Figure 5.28: Molecular Dynamics trajectory at T=280K : "Hydrophobic Collapse" pathway. (The figures were prepared using the program VMD [130])

The quick folding in molecular dynamics simulation(fig.5.28) is a non-zipper pathway. It is accomplished through a hydrophobic collapse.

From the above observations its clear that the folding through hydrophobic collapse is common in fast folding MD simulations and minima hopping search. The zipper mechanism has been observed in MD simulations before. The partly-helical conformation intermediated folding (of fig. 5.27) has been observed in replica exchange MD simulations of C-terminal peptide from the B1 domain of protein G (a $\beta - hairpin$)[239].

5.4 Conclusions

In this chapter, in section 5.2 we have applied the MHOP algorithm to find the native state, at atomic resolution, of a small peptide which folds into a stable β-hairpin. The structural model predicted was accurate to less than 0.5 Å RMSD and the NOE-violation statistics were satisfactory. The native funnel was deeper than non-native ones, which is a primary condition for the applicability of global optimization methods. A search pathway was extracted from the minima search process and the conformations that were found were used to characterize the free energy surface.

There have been a number of other studies on this peptide using canonical MD, multi-canonical MD [189], REMD [200] and hybrid hamiltonian replica exchange [133] methods, and our results are consistent with those obtained previously. From a purely methodological point of view, MHOP should be a useful approach for protein structure prediction that is complementary to other methods, given that it combines a biased sampling process with a non-thermodynamic search algorithm. Like other methods, though, the utility of the algorithm and its success in ab-initio structure prediction and for studying protein folding are highly dependent on the quality of the forcefield model that is used.

In section 5.3, we could successfully find the lowest energy conformation of a tryptophan-zipper peptide and thus could find the native state of the peptide using minima hopping algorithm. The necessary criteria for a successful structure prediction is that the native funnel of the effective free energy surface has to be deeper than the others - for which the forcefield to be used has to be “correct”. This criteria was fulfiled here. Next, We observed multiple folding pathways for the trpzip2, depending on how the two hydrophobic pairs approach to their native conformations. Our MD simulations support the folding mechanism put forward by the minima hopping search. The probability of folding through a particular pathway cannot be predicted because the search process does not generate any thermodynamical ensemble or truthful kinetics. Moreover, this kind of a biased search has the predilection for any possible fast folding routes. But there could have been a scope on a qualitative level for the minima hopping search trajectories to describe the folding mechanism. This study suggests that a connected sequence of minima searched by the minima hopping stochastic simulator can point towards a realistic physical process and thus describe a mechanism for a conformational reaction. This is due to the micro-canonical MD based moved instead of random and violent moves that was incorporated in minima hopping algorithm and thus always overcoming realistic transition region and avoiding unphysically high barriers.

Chapter 6

Summary

The previous chapters described a number of applications of Minima hopping algorithm on the structural studies of different biomolecules. The major aim of them was to approach towards the protein structure prediction problem. This problem is about finding the global minimum of the free energy surface as per the thermodynamic hypothesis, so given an effective free energy forcefield and an efficient global optimization algorithm one can expect to predict the native state of a protein. So the tasks were : (a) to define an effective free energy forcefield, (b) to select a global optimization algorithm for sampling, (c) to test the whole scheme for systems with simplified interactions, (d) finally to apply this on proteins. We started with minima hopping algorithm which has been described in chapter 3. The work described in ref [126] shows that minima hopping algorithm has shown better overall success as a global optimization algorithm over evolutionary algorithms and basin hopping algorithm. The effective free energy forcefield is defined as the OPLS all-atom forcefield and a GB/SA solvation scheme which incorporates the solvation free energy. For studying proteins in gas phase, no free energy term has been added to the expression of the OPLS energy function. In course of the structure prediction studies, a few allied problems have been work on. Among them are - (a) understanding the energy landscape and (b) making the conformational search algorithm more efficient.

In the chapter 2 we have shown that the BEP principle can be extended to molecular dynamics trajectories which cross from one basin of attraction into another one far from the transition state. This extended principle (MDBEP) says that MD trajectories with lower energy are more likely to lead into basins of attraction of low energy configurations than very high energy trajectories. We have found that in the context of global optimization this principle can be used to improve the efficiency of existing MD based methods by properly tuning the energy of the MD trajectories. In the next section in the chapter, we have conducted a comparative study of two kinds of energy landscapes, one of the OPLS forcefield and the the other one of density functional interactions through the energetic and structural correspondence. Our observations clearly showed that there are a significant amount of unstable and fake local minima present in the landscape corresponding to the forcefield. The nonbonded interactions of the forcefields needs to be corrected. The present set Lennard-Jonnes paramemters have been observed to represent a van der Waal force which has a considerably hard repulsive region - so the forcefield with the current form or the parameters are expected not to reproduce the packing and the compactness of real proteins.

In the chapter 3 several algorithmic developments and implementations have been discussed. Following the intuition the chapter 2 brought, we devised a couple of schemes to direct the the MD moves in the minima hopping algorithm to cross low barriers. The *dimersoft* scheme and the force decomposition scheme were described. Here, an enhanced feedback scheme for minima hopping algorithm have been described. The application of the force decomposition scheme as a preconditioner is briefly shown. A new preconditioned-DIIS geometry optimizer has been tested and shown to be $3-5$ times faster than our best implementation of conjugate gradient. Coupling this preconditioned-DIIS with steepest descent and conjugate gradient a hybrid geometry optimizer is introduced. All of these developments have in some way or the other been utilized in the minima hopping runs and different tests.

To reduce the complexity of protein folding in aqueous medium we studied biomolecular structures in gas phase for a number of systems. In the chapter 4 it has been shown that minima hopping algorithm remains successful in finding the putative global minimum and many low energy conformations efficiently which in turn helped in understanding some aspects of the experiments done on peptides in gas phase like ion-mobility experiments or different spectroscopic experiments. From the comparative study of the predicted structures in different forcefields, it is apparent that its not easy to choose a ”correct” forcefield. It is also seen in this study that the conformation-specific charge parametrization can make a considerable impact on the energy landscapes.

It is believed that β-hairpins are more complex structures as compared to α-helices which have got some translational symmetry along the axis of the helix whereas the local environments along the β-hairpins are different from the others.The inter-strand hydrogen bonds in a β-hairpin have different sequence separations and not being as simple and regular as those in a helix. So, it is more difficult to find those structures in a (Monte Carlo) search process as compared to helices. In Chapter 5 we successfully folded two β-hairpins and have seen that the native funnel is deeper than the non-native ones. The search pathway being a sequence of short Molecular dynamic segments prepended with a soft-mode search of initial directions which ensures the crossing of low-barriers is shown to be almost as realistic as a molecular dynamics trajectory, still being quicker than MD.

The success of applying minima hopping algorithm for biomolecular structure prediction up to now was limited because of a number of reasons.

- The problem of initially not having a fast geometry optimizer. Spending 90% of the CPU time for geometry optimization is something which could not be coped up with - and for exploring a significant amount of the landscape for bigger proteins this step of the algorithm is a real bottleneck.

- We used the all-atom framework of simulation which although being a great scientific challenge to build a solution strategy against, is not quite a practical approach for protein structure prediction. For a significant amount of sampling for a system as small as a 20 residue TRP-Cage protein, e.g. searching 10^5 local minima using the combination of the hardware and the libraries used for the present work, taking one geometry optimization to be the average time which it takes $\sim 20min$, we would have to spend $10^5 * 20(60 * 24 * 365) \approx 4yrs$, that is $\sim$ the whole time frame of the PhD work being reported here !!

- A reliable forcefield in which the native state would always be both within the deepest funnel in the energy landscape and at the bottom of the native funnel. If it is shifted to a high energy region within the native funnel it is possible for the global optimizer to miss them as it is designed to reach the bottom of the funnel very fast. Related to this, is the issue of correct implicit solvation model. This affects the simulation in two ways - first, by influencing the hydrophobic forces and thus influencing the energetic, second, by influencing the ruggedness of the landscape which make the geometry optimization step affected.

Outlook

While there is always a scope for improvements for the sampling algorithms, it seems that the foremost thing to be done is the framework for the structure prediction to be dealt with realism. The strength of minima hopping can be exercised at the points where fast conformational search is needed. The following approaches would be interesting.

- For having a thorough and faster sampling of the landscape an efficient search algorithm should be accompanied with some coarse grained forcefield and for each set of distinct conformations (assuming them to be representative of different funnels) to run all atom “within-the-funnel” refinement simulation. Or perhaps, using comparative modeling based tools to generate many of the crude models, and then using the global optimization strategy for neighbourhood search.

- After the advent of NVIDIA GPUs and the GPU computing it is being possible to get a speedup of 100-1000 times for a single force evaluation. The implementation of minima hopping algorithm and the forcefield library on such a platform can thus overcome the CPU-time bottleneck and be a viable option to work on biomolecular structure prediction in future.

- Comparative modeling in blind structure prediction has been quite successful. But in homology modeling one encounters many unaligned regions in sequence alignment. These loop regions tend to be located at the solvent-exposed surface of globular proteins and thus are very flexible. To predict the loops or in so-called Protein loop modeling one can perhaps use a global optimization based approach like minima hopping, to find many low energy conformations of a loop. Even if a loop might not adopt the global minimum of its conformational surface, it should in principle be located by a thorough sampling.

Appendix A

Programs and Analysis tools used

DYNAMO

Dynamo modular library has been extensively used for the work present here. It is an open source program library that has been designed for the simulation of molecular systems. The most recent version of this can be found from http://www.pdynamo.org/mainpages/.

VMD

VMD has been used for the visualizaion of proteins. Many of the figures present in the previous chapters were prepared using this program. It can be found from
http://www.ks.uiuc.edu/Research/vmd/ .

TINKER

This is a open-source library for biomolecular simulation. It is a versatile and easy-to-use library. It served a number of purposes starting from constructing a conformation with a given set of backbone dihedral angles upto a full global optimization simulation after coupling it with the minima hopping program. It can be found from
http://dasher.wustl.edu/tinker/ .

POEM

POEM is a package where PFF02 forcefield is implemented. It can be found from
http://iwrwww1.fzk.de/biostruct/ .

DFIRE

http://sparks.informatics.iupui.edu/hzhou/dfire.html is a web server which can be used for calculating the DFIRE free energy of a protein conformation.

NAB

This is a open-source library for biomolecular simulation. It was used for calculating conformational energies in Amber forcefield. It can be found from
http://casegroup.rutgers.edu/casegr-sh-2.2.html .

MMTSB Toolset

This toolsent has been used for the analysis of the protein conformations and the trajectories. It can be found from http://www.pdynamo.org/mainpages/ .

GNUPLOT

This is an excellent plotting program. It was used in numerous occasions. It can be found from http://www.gnuplot.info/ .

V_SIM

This is a very useful visualization software. It can be found from
http://www-drfmc.cea.fr/sp2m/L_Sim/V_Sim/index.en.html .

DSSP

Dictionary of Secondary Structure of Proteins (DSSP)was used to identify the secondary elements of proteins. This is a practical tool for expressing the protein in terms of its secondary structure. It can be found from http://swift.cmbi.ru.nl/gv/dssp/ .

SMMP

This package was used for calculating conformations energies in ECEPP forcefield. It can be found from http://www.smmp05.net/ .

Bibliography

[1] Branden, C. and Tooze, J., Introduction to Protein Structure,(1991) Garland Publishing, New York.

[2] Schulz, G.E. and Schirmer, R.H., Principles of Protein Structure, (1979) Springer-Verlag, New York.

[3] D.Wales, Energy landscapes (Cambridge University Press, Cambridge, 2003).

[4] Frenkel D, Smit B. Understanding Molecular Simulation. From Algorithms to Applications. San Diego: Academic Press; 2002.

[5] A. R. Leach, Molecular Modelling, Principles and Applications (AddisonWesley, Essex, 1996).

[6] Harrison S. C. and Durbin R., Proc. Natl. acad. sci. USA **82**, 12, 4028-4030 (1985).

[7] Baldwin R. L., Nature **346**, 409 (1990).

[8] Baldwin R. L., Curr. Opin. Struct. Biol. **3**, 84 (1993).

[9] Scheraga HA, Lee J, Pillardy J, Ye Y-J, Liwo A, Ripoll D., J Glob Optimiz. 1999;15:235260.

[10] Liwo A, Khalili M, Scheraga HA., Proc Natl Acad Sci USA. 2005;102:23622367.

[11] Lei H, Duan Y. Curr Opin Struct Biol. 2007;17:187191.

[12] Christen M, van Gunsteren WF. J Comput Chem. 2008;29:157166.

[13] Androulakis IP, Maranas CD, Floudas CA., J Glob Optimiz. 1995;11:337363.

[14] Lee J, Scheraga HA, Rackovsky S., J Comput Chem. 1997;18:12221232.

[15] Odziej S, Czaplewski C, Liwo A, Chinchio M, Nanias M, Vila JA, Khalili M, Arnautova YA, Jagielska A, Makowski M, et al. Proc Natl Acad Sci USA. **102**, 75477552 (2005).

[16] Joo K, Lee J, Lee S, Seo J-H, Lee SJ, Lee J., Proteins. 2007;69(Suppl 8):8389.

[17] Yamashita H, Endo S, Wako H, Kidera A., Chem Phys Lett. 2001;342:382386.

[18] Ulmschneider JP, Ulmschneider MB, Di Nola A., J Phys Chem B. 2007;110:1673316742.

[19] Duan V, Kollman PA., Science. 1998;282:740744.

[20] Chen JH, Im W, Brooks CL., J Comput Chem. 2005;26:15651578.

[21] Lazaridis, T.; M. Karplus, Proteins: Structure, Function, and Genetics 35: 133152 (1999)

[22] Ferrara, P., P.; J. Apostolakis, and A. Caflisch , Proteins: Structure, Function, and Genetics 46: 2433. (2002)

[23] D.T., Eisenberg; L. Wesson , Protein Sci. 1: 227235 (1992)

[24] Zhou R.,Proteins 53(2):148-61 (2003)

[25] Ho BK, Dill KA., PLoS Comput Biol 2(4):e27 (2006)

[26] Kabsch W., Sander C.,Biopolymers. 22(12):2577-637 (1983).

[27] Chen, J. H., Brooks, C. L. and Khandogin, J. Curr. Opin. Struct. Biol., 18, 140-148 (2008).

[28] Cornell WD, et al. J. Am. Chem. Soc (1995) 117:51795197

[29] MacKerell AD Jr., J. Phys. Chem. B (1998) 102:35863616

[30] Soares TA, et al., J. Comput. Chem (2005) 26:725737

[31] Yoda T, et al. Chem. Phys. Lett (2004) 386:460467

[32] Wang J, et al., J. Comput. Chem (2000) 21:10491074

[33] Duan Y, et al., J. Comput. Chem (2003) 24:19992012

[34] Lwin TZ, Luo R., Prot. Sci (2006) 15:26422655

[35] Blanco FJ, et al., Nat. Struct. Biol (1994) 1:584590

[36] Fesinmeyer RM, et al., J. Am. Chem. Soc (2004) 126:72387243

[37] A. Schug, "Free-Energy Simulations using Stochastic Optimization Methods for Protein Structure Prediction", PhD Thesis, (2005).

[38] A. Verma, ”Development and Application of a Free-Energy Force-Field for All-Atom Protein Folding”, PhD Thesis, (2007).

[39] K.A.Dill and S.Bromberg, Molecular Driving Forces (Garland Science, New York, 2003).

[40] F.Jensen, Introduction to Computational Chemistry (Wiley, New York, 1999).

[41] N.Brønsted, Chem. Rev. **5**, 231 (1928).

[42] R.P.Bell, Proc. Roy. Soc. London, Ser. A **154**, 414 (1936); M.G.Evans, M.Polanyi, Trans. Faraday Soc. **31**, 875 (1935);

[43] R.A.Marcus, J. Phys. Chem. **72**, 891 (1968).

[44] M.R. Hoare and P. Pal, Adv. Phys. 20 161 (1971)

[45] J.K.Nørskov, T.Bligaard, A.Logadottir, S.Bahn, M.Bollinger, L.B.Hansen, H.Bengaard, B.Hammer, Z.Sljivancanin, M.Mavrikakis, Y.Xu, S.Dahl, C.J.H.Jacobsen, J. Catal. **209**, 275 (2002); Z.-P.Liu, P.Hu, J. Chem. Phys. **114**, 8244 (2001); A.Logadottir, T.H.Rod, J.K.Nørskov, B.Hammer, S.Dahl, C.J.H.Jacobsen, J. Catal. **197**, 229 (2001); T.Bligaard, J.K.Nørskov, S.Dahl, J.Matthiesen, C.H.Christensen and J.Sehested, Journal of Catalysis **224**, 206 (2004).

[46] S.Goedecker, J. Chem. Phys. **120**, 9911 (2004).

[47] S.Goedecker, W.Hellmann and T.Lenosky, Phys. Rev. Lett. **95**, 055501 (2005).

[48] S. Yoo S and X.C. Zeng , J. Chem. Phys. **123** 164303 (2005).

[49] T.J.Lenosky, J.D.Kress, I.Kwon, A.F.Voter, B.Edwards, D.F.Richards, S.Yang and J.B.Adams, Phys. Rev. B **55**, 1528 (1997).

[50] G. Rossi and R. Ferrando, Chem. Phys. Let. **423**, 17 (2006).

[51] A. Oganov and C. Glass, J. Chem. Phys. **124**, 244704 (2006)

[52] A. Laio and M. Parrinello, Proc. Nat. Acad. of Sc. USA **99** 12562 (2002)

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

[54] G. T. Barkema and Normand Mousseau, Phys. Rev. Lett.**77**, 4358 (1996); S. Santini, G. Wei, N. Mousseau, and P. Derreumaux, Internet Electron. J. Mol. Des. **2**, 564 (2003).

[55] A.F.Voter, Phys. Rev. Lett.**78**, 3908 (1997); A.F. Voter, F. Montalenti, and T.C. Germann, Annu. Rev. Mater. Res. **32**, 321 (2002).

[56] H. Kabrede, Chem. Phys. Lett. **430**, 336 (2006).

[57] Genovese L. et al, J. Chem. Phys. 129, 014109 (2008)

[58] J. Perdew and A. Zunger Phys. Rev. B, 23 , 5048 (1981)

[59] Perdew, J. P., Burke, K., Ernzerhof, M. Phys. Rev. Lett., 77, 3865 (1996)

[60] Becke, A. D. J. Chem. Phys., 98, 5648 (1993)

[61] Lee, C., Yang, W., Parr, R. G. Phys. Rev. B, 37, 785 (1998)

[62] Lynch, B. J., Fast, P. L., Harris, M., Truhlar, D. G. J. Phys. Chem. A, 104, 4811 (2000)

[63] Zhao, Y., Tishchenko, O., Truhlar, D. G. J. Phys. Chem. B, 109, 19046 (2005)

[64] Tsuzuki, S., Luthi, H. P. J. Chem. Phys., 114, 3949 (2001)

[65] Duncan, J. A., Spong, M. C. J. Phys. Org. Chem., 18, 462 (2005)

[66] Wodrich M. D., Corminboeuf C. and Schleyer P. V. R., Org. Lett., 8 (17), 36313634 (2006)

[67] Zhechkov, L., Heine, T., Patchkovskii, S., Seifert, G., Duarte, H. A. J. Chem. Theory Comput., 1, 841 (2005)

[68] Lein, M., Dobson, J. F., Gross, E. K. U. J. Comput. Chem., 20, 12 (1999)

[69] Valdes, H., Sordo, J. A. J. Comput. Chem., 23, 444. (2002)

[70] S. Goedecker, F. Lancon, and T. Deutsch, Phys. Rev. B 64, 161102 (2001).

[71] K. Németh, O. Coulaud, G. Monard, and J. G. ngyan, J. Chem. Phys. 113, 5598 (2000).

[72] Schlegel, H. B. Geometry Optimization on Potential Energy Surfaces. In Modern Electronic Structure Theory; Yarkony, D. R., Ed.; World Scientific: Singapore, 1995; p 459.

[73] Schlegel, H. B. Geometry Optimization. In Encyclopedia of Computational Chemistry; Schleyer, P. v. R., Allinger, N. L., Kollman, P. A., Clark, T., Schaefer, H. F., III, Gasteiger, J., Schreiner, P. R., Eds.; Wiley: Chichester, U. K., 1998; Vol. 2, p 1136.

[74] Fletcher, R. Practical Methods of Optimization; Wiley: Chichester, U. K., 1981.

[75] Broyden, C. G. J. Inst. Math. Appl.1970, 6, 76.

[76] Fletcher, R. Comput. J. (Switzerland)1970, 13, 317.

[77] Goldfarb, D. Math. Comput.1970, 24, 23.

[78] Shanno, D. F. Math. Comput.1970, 24, 647.

[79] Murtagh, B.; Sargent, R. W. H. Comput. J. (Switzerland)1972, 13, 185.

[80] Powell, M. J. D. Nonlinear Programing; Academic: New York, 1970.

[81] Powell, M. J. D. Math. Program.1971, 1, 26.

[82] Banerjee, A.; Adams, N.; Simons, J.; Shepard, R. J. Phys. Chem.1985, 89, 52.

[83] Simons, J.; Nichols, J. Int. J. Quantum Chem.1990, 24, 263.

[84] Murray, W.; Wright, M. H. Practical Optimization; Academic: New York, 1981.

[85] Nonlinear Optimization; Powell, M. J. D., Ed.; Academic: New York, 1982.

[86] Dennis, J. E.; Schnabel, R. B. Numerical Methods for Unconstrained Optimization and Nonlinear Equations; Prentice Hall: Upper Saddle River, New Jersey, 1983.

[87] Scales, L. E. Introduction to Nonlinear Optimization; Macmillam: Basingstoke, England, 1985.

[88] Fletcher, R. Practical Methods of Optimization; Wiley: Chichester, U. K., 1981.

[89] Pulay, P. Chem. Phys. Lett.1980, 73, 393.

[90] Pulay, P. J. Comput. Chem.1982, 3, 556.

[91] Cancès, E.; Le Bris, C. Int. J. Quantum Chem.2000, 79, 82.

[92] Kudin, K. N.; Scuseria, G. E.; Cancès, E. J. Chem. Phys.2002, 116, 8255.

[93] Li, X.; Millam, J. M.; Scuseria, G. E.; Frisch, M. J.; Schlegel, H. B. J. Chem. Phys.2003, 119, 7651.

[94] Csaszar, P.; Pulay, P. J. Mol. Struct.1984, 114, 31.

[95] Farkas, Ö.; Schlegel, H. B. Phys. Chem. Chem. Phys.2002, 4, 11.

[96] Frisch, M. J. et al., Development Version Rev. D01 ed.; Gaussian, Inc.: Pittsburgh, PA, 2005.

[97] Li. X., Frisch M. J., J. Chem. Theory Comput., 2 (3), 835839 (2006).

[98] Bofill, J. M. J. Comput. Chem.1994, 15, 1.

[99] Farkas, Ö.; Schlegel, H. B. J. Chem. Phys.1999, 111, 10806.

[100] Baker, J. J. Comput. Chem.1993, 14, 1085.

[101] Xiaosong Li, Michael J. Frisch, Journal of Chemical Theory and Computation 2006 2 (3), 835-839,

[102] W. L. Jorgensen, D. S. Maxwell, and J. Tirado-Rives, J. Am. Chem. Soc. **118** 11225 (1996).

[103] M. J. Field, M. Albe, C.Bret, F. Proust-De Martin, and A. Thomas, J. Comp. Chem. **21**, 1088-1100 (2000); DYNAMO is free software and can be obtained from the website `http://www.pdynamo.org`.

[104] M. J. Field, *A Practical Introduction to the Simulation of Molecular Systems* (Cambridge University Press, Cambridge, 1999).

[105] T.J.Lenosky, B.Sadigh, E.Alonso, V.Bulatov, T.Diaz de la Rubia, J.Kim, A.F.Voter adn J.D.Kress, Modelling Simul. Mater. Sci. Eng. **8**, 825 (2000).

[106] D. Caliste, L. Billard and O. D'Hastier URL: http://www-drfmc.cea.fr/sp2m/L_Sim/V_Sim/

[107] Hudgins RR, Ratner MA, Jarrold MF. , J. Am. Chem. Soc. **120**, 12974–12975 (1998).

[108] R. R. Hudgins and M. F. Jarrold, J. Am. Chem. Soc., **121**, 3494-3501 (1999).

[109] M. Kohtani, T. C. Jones, J. E. Schneider and M. F. Jarrold, J. Am. Chem. Soc.,**126** , 7420-7421 (2004).

[110] Stearns J. A., Boyarkin O. V. and Rizzo T. R. , J. Am. Chem. Soc., **129**, 13820-13821 (2007).

[111] Stearns J. A., Seaiby C., Boyarkin O. V.and Rizzo T. R. Phys. Chem. Chem. Phys., **11**, 125-132 (2009).

[112] Boyarkin O. V., Mercier S. R., Kamariotis A. and Rizzo T. R., J. Am. Chem. Soc., **128**, 2816-2817 (2006).

[113] Stearns J. A., Mercier S., Seaiby C., Guidi M., Boyarkin O. V. and Rizzo T. R., J. Am. Chem. Soc.,**129**, 1181411820 (2007).

[114] P. Dugourd, R. R. Hudgins, D. E. Clemmer, and M. F. Jarrold, Rev. Sci. Instrum. **68**, 1122 (1997)

[115] R. R. Hudgins, J. Woenckhaus, and M. F. Jarrold, Int. J. Mass Spectrom. Ion Process.**165/166**, 497 (1997).

[116] M. Kohtani, J. E. Schneider, T. C. Jones, and M. F. Jarrold, J. Am. Chem. Soc. **126**, 16981 (2004).

[117] Zilch LW, Kaleta DT, Kohtani M, Krishnan R, Jarrold MF. , J Am Soc Mass Spectrom. **18**(7), 1239-48 (2007).

[118] Y. Peng, U. H. E. Hansmann, and N. A. Alves, J. Chem. Phys. **118**, 2374 (2003).

[119] Wei Y., Nadler W., Hansmann U. H. E., J. Chem. Phys. 126, 204307 (2007).

[120] M. Kohtani and M. F. Jarrold, J. Am. Chem. Soc. **126**, 8454 (2004).

[121] Ohkita, M.; Lehn, J.-M.; Baum, G.; Fenske, D. Chem. Eur. J. **5**, 34713481 (1999).

[122] Hill D. J.,Mio M. J. ,Prince R. B., Highes T. S. and Moore J. S., Chem. Rev., **101** (12), pp 38934012 (2001).

[123] Anfinsen C. B., Science 181 (1973) 223.

[124] S. Goedecker, J. Chem. Phys **120**, 9911 (2004).

[125] W.Hellmann, R.G.Hennig, S.Goedecker, C.J.Umrigar, B.Delley and T.Lenosky, Phys. Rev. B **75**, 085411 (2007).

[126] Schoenborn S. E., Goedecker S., Roy S.and Oganov A. R. , J Chem Phys. **130** (14),144108 (2009).

[127] Wenzel, W. , Europhys. Lett., 76 (1), p. 156 (2006).

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

[129] Roy, S., Goedecker S. and Hellmann V. , Phys. Rev. E **77**, 056707 (2008).

[130] Humphrey, W., Dalke, A. and Schulten, K., ”VMD - Visual Molecular Dynamics”, J. Molec. Graphics, **14**, pp. 33-38 (1996).

[131] M. Feig, J. Karanicolas and C.L. Brooks III, J. Mol. Graph. Model. **22**, pp. 377–395 , (2004).

[132] http://www.rcsb.org/

[133] W.X. Xu, T.F. Lai, Y. Yang and Y.G. Mu, J. Chem. Phys. **128**, 175105 (2008).

[134] Ismer L., Ireta J., and Neugebauer J., J. Phys. Chem. B **112**, 4109-4112 (2008).

[135] Ismer L., Ireta J., Boeck S. and Neugebauer J., Phys. Rev. E **71**, 031911 (2005).

[136] Feig M., MacKerell A. D. Jr. and Brooks C. L. III, J. Phys. Chem. B **107**, 2831-2836 (2003).

[137] J. Hughes, T.W. Smith, H.W. Kosterlitz, L.A. Fothergill, B.A. Morgan and H.R. Morris, Nature **258**, pp. 577-579 (1975).

[138] W.H. Graham, E.S. Carter II and R.P. Hicks, Biopolymers **32**, pp. 1755-1764 (1992).

[139] Li and H.A. Scheraga, J. Mol. Struct. THEOCHEM. **179** pp. 333-352 (1988).

[140] U.H.E. Hansmann and J.N. Onuchic, J. Chem. Phys. **115** , pp. 1601-1606 (2001).

[141] U.H.E. Hansmann, Y. Okamoto and J.N. Onuchic, Proteins Struct. Funct. Genet. **34** , pp. 472-483 (1999).

[142] D.A. Evans and D.J. Wales, J. Chem. Phys. **119** , pp. 9947-9955 (2003).

[143] B. von Freyberg and W. Braun, J. Comput. Chem. **12**, pp. 1065-1076 (1991).

[144] F. Eisenmenger and U.H.E. Hansmann, J. Phys. Chem. B **101**, pp. 3304-3310 (1997).

[145] H. Meirovitch, E. Meirovitch, A.G. Michel and M. Vsquez, J. Phys. Chem. **98**, pp. 6241-6243 (1994).

[146] I.P. Androulakis, C.D. Maranas and C.A. Floudas, J. Global Opt. **11**, pp. 1-34 (1997).

[147] Z. Li and H.A. Scheraga, Proc. Natl. Acad. Sci. USA **84** pp. 6611-6615 (1987).

[148] A.G. Michel, C. Ameziane-Hassani and N. Bredin, Can. J. Chem. **70** pp. 596-603 (1992).

[149] T. Montcalm, W. Cui, H. Zhao, F. Guarnieri and S.R. Wilson, J. Mol. Struct. THEOCHEM. **308**, pp. 37-51(1994).

[150] Zhan L, Chen JZY, Liu WK. Biophys J **91** 2399-2404 (2006).

[151] Wang Z.-X., Duan Y. J., Comput. Chem. **25**, 1699-1716 (2004).

[152] Vargas R., Garza J., Hay B. P., Dixon D. A. J. Phys. Chem. A **106** (13), 3213-3218 (2002).

[153] Rossky, P. J.; Karplus, M. J. Am. Chem. Soc., **101**, 1913 (1979).

[154] Brooks, C. L., III; Case, D. Chem. Rev., **93**, 2487 (1993).

[155] Smith, P. E.; Pettitt, B. M.; Karplus, M. J. Phys. Chem., **97**, 6907 (1993).

[156] Apostolakis, J.; Ferrara, P.; Caflisch, A. J. Chem. Phys., **110**, 2099 (1999).

[157] Anderson, A. G.; Hermans, J. Proteins: Struct., Funct., Genet., **3**, 262 (1998).

[158] Tobias, D. J.; Brooks, C. L., III. J. Phys. Chem., **96**, 3864 (1992).

[159] Smith, P. E. J. Chem. Phys., **111**, 5568 (1999).

[160] Wu, X.; Wang, S. J. Phys. Chem. B, **102**, 7238 (1998).

[161] Derreumaux, P.; Schlick, T. Proteins: Struct., Funct., Genet., **21**, 282 (1995).

[162] Bolhuis, P.; Dellago, C.; Chandler, D. Proc. Natl. Acad. Sci. U.S.A., **97**, 5877 (2000).

[163] Hummer, G.; Kevrekidis, I. G. J. Chem. Phys., **118**, 10762 (2003).

[164] Drozdov, A. N.; Grossfield, A.; Pappu, R. V. J. Am. Chem. Soc., **126**, 2574 (2004).

[165] Hu, H.; Elstner, M.; Hermans, J. Proteins: Struct., Funct., Genet., **50**, 451 (2003).

[166] Weise, C. F.; Weisshaar, J. C. J. Phys. Chem. B, **107**, 3265 (2003).

[167] Lavrich, R. J.; Plusquellic, D. F.; Suenram, R. D.; Fraser, G. T.; HightWalker, A. R.; Tubergen, M. J. J. Chem. Phys., **118**, 1253 (2003).

[168] F. Eisenmenger, U.H.E. Hansmann, S. Hayryan and C.-K. Hu, Comp. Phys. Comm., **138** 192-212 (2001).

[169] F. Eisenmenger, U.H.E. Hansmann, S. Hayryan and C.-K. Hu, Comp. Phys. Comm., **174** 422-429 (2006).

[170] J.H. Meinke, S. Mohanty, F. Eisenmenger and U.H.E. Hansmann, Meinke, J. H., Mohanty, S., Eisenmenger, F., Hansmann, U. H. E. Comput Phys Commun, **178**, 459 (2008).

[171] D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, B. Wang, D.A. Pearlman, M. Crowley, S. Brozell, V. Tsui, H. Gohlke, J. Mongan, V. Hornak, G. Cui, P. Beroza, C. Schafmeister, J.W. Caldwell, W.S. Ross, and P.A. Kollman (2004), AMBER 8, University of California, San Francisco.

[172] M. J. Frisch et al, GAUSSIAN 03, Revision B.01, GAUSSIAN, INC., Pittsburgh, PA, 2003.

[173] http://predictioncenter.org/ .

[174] Anfinsen C. B., Science 181 (1973) 223.

[175] J.N. Onuchic, Z.A. Luthey-Schulten and P.G. Wolynes, Annu. Rev. Phys. Chem. 48, 545600 (1997)

[176] Y. Sugita, Y. Okamoto , Chem. Phys. Lett. **314**, 141–151 (1999). ; K. Hukushima, K. Nemoto , J. Phys. Soc. Jpn. **65**, 1604–1 608 (1996). ; K. Hukushima, H. Takayama, K. Nemoto , Int. J. Mod. Phys. C **7**, 337–344 (1996).

[177] R. H. Swendsen and J.-S. Wang, Phys. Rev. Lett., **57**, 2607 (1986).

[178] A.P. Lyubartsev, A.A. Martinovski, S.V. Shevkunov, P.N. Vorontsov-Velyaminov , J. Chem. Phys. **96**, 1776–1783 (1992). ; E. Marinari, G . Parisi , Europhys. Lett. **19**, 451–458 (1992).

[179] A. Schug, T. Herges, and W. Wenzel, Phys. Rev. Lett. **91**, 158102 (2003). ; W. Wenzel and K. Hamacher , Phys. Rev. Lett. **82**, 3003 (1999).

[180] Y. Sugita, Y. Okamoto, Chem. Phys. Lett. **329**, 261–270 (2000). ; A. Mitsutake, Y. Sugita, Y. Okamoto , J. Chem. Phys. **118** , 6664–6675 (2003). ; A. Mitsutake, Y. Sugita, Y. Okamoto , J. Chem. Phys. **118**, 6676–6688 (2003).

[181] Laio, A., Parrinello, M. Proc. Natl. Acad. Sci. U.S.A., **99**, 12562-12566 (2002).

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

[183] Wei, G., Mousseau, N. and Derreumaux, P. Proteins, **56**, 464–474 (2004).

[184] Verma A., Schug A., Lee K. H. and Wenzel W. , J. Chem. Phys., **124**, 044515 (2006).

[185] Verma A. and Wenzel W. , J. Phys. Cond. Matt., **19**, 285213 (2007).

[186] Gopal S. M., and Wenzel W., J. Phys. Cond. Matt., **19**, 285210 (2007).

[187] M.C. Prentiss, D.J. Wales and P.G. Wolynes, J. Chem. Phys.,**128**, 225106 (2008).

[188] S. Honda, K. Yamasaki, Y. Sawada, and H. Morii, Structure (London) 12, 1507 (2004).

[189] D. Satoh, K. Shimizu, S. Nakamura and T. Terada, FEBS Lett. 580 (2006), p. 3422.

[190] P. Bradley, K.M. Misura and D. Baker, Science **309**, pp. 1868–1871 (2005).

[191] http://www.scripps.edu/case/nab.html .

[192] Di Qiu, P. S. Shenkin, F. P. Hollinger, and W. Clark Still, J. Phys. Chem A **101**, 3005 (1997)

[193] Hermann, R. B. J. Phys. Chem. 1972, 76, 2754. ; Amidon, G. L.; Yalkowsky, S. H.; Anik, S. T.; Valvani, S. C, J. Phys. Chem. 1975, 72, 2 239. ; Floris, F.; Tomasi, J. J. Comput. Chem. 1989, 10, 616.

[194] D. A. Evans and D.J. Wales, J. Chem. Phys., **119**, 9947-9955 (2003).

[195] D.A. Evans and D.J. Wales, J. Chem. Phys., **121**, 1080-1090 (2004).

[196] D. J. Wales and J. P. K. Doye, J. Phys. Chem. A, **101**, 5111 (1997).

[197] H. Zhou and Y. Zhou, Protein Science, **11**, 2714-2726 (2002).

[198] Y. Yang and Y. Zhou, Proteins, **72**, 793-803 (2008).

[199] http://www.scripps.edu/mb/case/qshifts/qshifts.htm .

[200] Seibert MM, Patriksson A, Hess B, van der Spoel D., J Mol Biol **354**, 173-183 (2005).

[201] Herges T. and Wenzel W., Biophys. J., **87**, 3100 (2004).

[202] Verma A. and Wenzel W., Biophys. J., **96**, 3483-3494 (2009).

[203] Cochran, A.G., Skelton, N.J., Starovasnik, M.A. Proc.Natl.Acad.Sci.USA 98: 5578-5583 (2001).

[204] Muoz V., Ghirlando R., Blanco F. J., Jas G. S., James Hofrichter J.,and William A. E., Biochemistry, 2006, 45 (23), pp 70237035 (2006)

[205] Muoz V., Annual Review of Biophysics and Biomolecular Structure 2007, 36 (1), 395-412 (2007).

[206] Pande, V. S., and Rokhsar, D. S., Proc. Natl. Acad. Sci. U.S.A. 96, 9062-9067 (1999).

[207] Roccatano, D., Amadei, A., Di Nola, A., and Berendsen, H. J. C., Protein Sci. 8, 2130-2143 (1999).

[208] Kolinski, A., Ilkowski, B., and Skolnick, J., Biophys. J. 77, 2942-2952 (1999).

[209] Dinner, A. R., Lazaridis, T., and Karplus, M., Proc. Natl. Acad. Sci. U.S.A. 96, 9068-9073. (1999).

[210] Klimov, D. K., and Thirumalai, D. , Proc. Natl.Acad. Sci. U.S.A.97, 25442549 (2000).

[211] Ma, B. Y., and Nussinov, R., J. Mol. Biol.296, 10911104 (2000).

[212] Bryant, Z., Pande, V. S., and Rokhsar, D. S., Biophys. J.78, 584589 (2000).

[213] Lee, J., and Shin, S. M., Biophys. J.81, 25072516 (2001).

[214] Garcia, A. E., and Sanbonmatsu, K. Y. , Proteins 42, 345354 (2001).

[215] Zhou, R. H., Berne, B. J., and Germain, R., Proc. Natl. Acad. Sci. U.S.A.98, 1493114936 (2001).

[216] Eastman, P., Gronbech-Jensen, N., and Doniach, S., J. Chem. Phys.114, 38233841 (2001).

[217] Tsai, J., and Levitt, M., Biophys. Chem.101, 187201 (2002).

[218] Jang, S., Shin, S., and Pak, Y., J. Am. Chem. Soc.124, 49764977 (2002).

[219] Cieplak, M., Hoang, T. X., and Robbins, M. O., Proteins 49, 104113 (2002).

[220] Zhou, R. H., and Berne, B. J., Proc. Natl. Acad. Sci. U.S.A.99, 1277712782 (2002).

[221] Lee, J., and Shin, S., J. Phys. Chem. B106, 87968802 (2002).

[222] Klimov, D. K., Newfield, D., and Thirumalai, D., Proc. Natl. Acad. Sci. U.S.A.99, 80198024 (2002).

[223] Wei, G. H., Derreumaux, P., and Mousseau, N. , J. Chem. Phys. 119, 64036406 (2003).

[224] Ma, B. Y., and Nussinov, R., Protein Sci. 12, 18821893 (2003).

[225] Zhou, R. H., Proteins 53, 148161 (2003).

[226] Bolhuis, P. G., Proc. Natl. Acad. Sci. U.S.A.100, 1212912134 (2003).

[227] Pande, V. S., Baker, I., Chapman, J., Elmer, S. P., Khaliq, S., Larson, S. M., Rhee, Y. M., Shirts, M. R., Snow, C. D., Sorin, E. J., and Zagrovic, B., Biopolymers 68, 91109 (2003).

[228] Zhou, Y. Q., Zhang, C., Stell, G., and Wang, J., J. Am. Chem. Soc.125, 63006305 (2003).

[229] Evans, D. A., and Wales, D. J., J. Chem. Phys.121, 10801090 (2004).

[230] Felts, A. K., Harano, Y., Gallicchio, E., and Levy, R. M., Proteins 56, 310321 (2004).

[231] Swope, W. C., Pitera, J. W., Suits, F., Pitman, M., Eleftheriou, M., Fitch, B. G., Germain, R. S., Rayshubski, A., Ward, T. J. C., Zhestkov, Y., and Zhou, R. , J. Phys. Chem. B108, 65826594. (2004).

[232] Krivov, S. V., and Karplus, M. , Proc. Natl. Acad. Sci. U.S.A.101, 1476614770 (2004).

[233] Paschek, D., and Garcia, A. E., Phys. Rev. Lett.93, 238105-1238105-4 (2004).

[234] Irback, A., and Sjunnesson, F., Proteins 56, 110116 (2004).

[235] Yoda, T., Sugita, Y., and Okamoto, Y., Chem. Phys.307, 269283 (2004)

[236] Lee, I. H., Kim, S. Y., and Lee, J., Chem. Phys. Lett.412, 307312 (2005).

[237] Carr, J. M., Trygubenko, S. A., and Wales, D. J., J. Chem. Phys.122, 234903 (2005).

[238] Irback, A., J. Phys.: Condens. Matter17, S1553S1564 (2005).

[239] Andrec, M., Felts, A. K., Gallicchio, E., and Levy, R. M., Proc. Natl. Acad. Sci. U.S.A.102, 68016806 (2005).

[240] Bolhuis, P. G., Biophys. J.88, 5061 (2005).

[241] Yang L., Shao Q. and Gao Y. Q., J. Phys. Chem. B, 2009, 113 (3), pp 803808 (2009).

[242] C. Chen and Y. Xiao, Bioinformatics 24, pp. 659665 (2008).

[243] Xiao Y., Chen C., and He Y., Int J Mol Sci. 10(6), 28382848 (2009).

[244] Nymeyer H., J. Phys. Chem. B, 113 (24), pp 82888295 (2009).

[245] Narayanana R., Pelakha L. and Hagen S. J., J. Mol. Biol. 390 (3), 538-546 (2009).

[246] Roy S., Goedecker S., Field MJ and Penev E., J. Phys. Chem. B, **113** (20), 7315 (2009).

Acknowledgment

First of all, I want to express my gratitude to Prof. Stefan Goedecker for giving me the opportunity to work for my PhD under his supervision. I am thankful to him for assigning me an interesting research topic, for providing the possibility of working within an easy environment, for the continuous encouragement and for the stimulating discussions as well as for his helpful remarks and critical comments.

I thank Dr. Martin Field first for providing technical and scientific help on numerous occassions. I thank him once more for kindly accepting being the co-referee of this thesis. I sincerely thank Prof. Torsten Schwede for the interesting discussions and also for kindly accepting to be an additional expert in the thesis committee.

Special thanks goes to Dr. Arthur Voter, LANL for his critical comments and suggessions on a part of my work. Many thanks to Prof Wolfgang Wenzel, FZK for all the help that his group provided.

I acknowledge the Swiss National Science Foundation for the financial support of my research work and Swiss National Supercomputing Centre (CSCS) for providing the computational resources.

I thank all my labmates, specially Vladimir and Alireza for all the interesting academic and non-academic discussions. I thank Frau Kammermann and Frau Kalt of the Department of Physics, Uni Basel for their friendly gestures and help.

Finally, I want to express that nothing could have been possible without the constant support and encouragement of my parents and my wife. I feel fortunate that they were there and my brothers were there. I am indebted to them...

Resume

Shantanu Roy

Date of birth : 28th March , 1982.
Nationality : Indian
Gender : Male
Marital Status : Married
Address :
Chemin du Grand-Pré 10
Epalinges 1066
Switzerland
Email: shantanu.roy@epfl.ch
shantanu.roy@unibas.ch

Office:
SV IBI-SV UPDALPE
EPFL
AAB 0 15 (Bâtiment AAB), Station 15
CH-1015 Lausanne
Switzerland
Phone : +41789224291

Education

- **University of Basel** Basel, Switzerland
  PHD in BioPhysics *Oct 2005 - Sept 2009*
- **University of Pune** Pune, India
  M.Sc(PHYSICS) *July, 2002 - July. 2004*
- **University of Calcutta** Kolkata, India
  B.Sc(HONS. in PHYSICS) *1999-2002*

Publications

- **S Roy** , S Goedecker and V Hellmann – Bell-Evans-Polanyi principle for molecular dynamics trajectories and its implications for global optimization, Phys. Rev. E **77**, 056707 (2008).
- **S Roy**, S Goedecker, MJ Field and E Penev – A minima-hopping study of all-atom protein folding and structure prediction, J. Phys. Chem. B, **113**, (20), pp 73157321 (2009).
- SE Schoenborn, S Goedecker, **S Roy** and AR Oganov – The performance of Minima Hopping and Evolutionary Algorithms for cluster structure prediction, J Chem Phys. **130**,144108(2009).
- SA Ghasemi, M Amsler, R Hennig, **S Roy**, S Goedecker, CJ Umrigar, L Genovese, TJ Lenosky, T Morishita and K Nishio – The energy landscape of silicon systems and its description by force fields, tight binding schemes, density functional methods and Quantum Monte Carlo methods, Phys. Rev. B **81**, 214107 (2010).

Work Experience

- **Prof. Matteo Dal Peraro, EPFL** Lausanne, Switzerland
  PostDoc *Oct 2009 -*
  - # Computational modeling of DNA-transcription factor interactions
- **Prof. Stefan Goedecker, University of Basel** Basel, Switzerland
  Doktorand *2005-2009*
  - # Research Topic in PhD : Minima Hopping within an all-atom framework for Biomolecular Structure Prediction
- **Prof. Indira Ghosh, Bioinformatics Centre , University of Pune** Pune, India
  Research Fellow *2004-2005*
  - # Developemnt of hybrid system of forcefields (QM/MM scheme) for application to molecular dynamics of biomolecules