

University of Heidelberg

Molecular Biotechnology (Summer 2004)

Bachelor Thesis:

**Basic Principles  
in  
Molecular Modeling**

By Barmak Mostofian (2185360) and Filipp Frank (2203674)

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Classical Mechanics</b>	<b>2</b>
2.1	Newton's Second Law . . . . .	2
2.2	Integration Algorithms . . . . .	3
2.2.1	Verlet Algorithm . . . . .	3
2.2.2	Leap-Frog Algorithm . . . . .	4
2.2.3	Velocity Verlet Algorithm . . . . .	5
<b>3</b>	<b>Statistical Mechanics</b>	<b>6</b>
3.1	Definitions . . . . .	6
3.2	Ensemble Averages and Time Averages . . . . .	6
<b>4</b>	<b>CHARMM</b>	<b>8</b>
4.1	Force Fields . . . . .	8
4.1.1	General Features of Molecular Mechanics Force Fields . . . . .	8
4.1.2	Bond Stretching . . . . .	9
4.1.3	Angle Bending . . . . .	11
4.1.4	Torsional Terms . . . . .	12
4.1.5	Improper Torsions / Out-of-Plane Bending . . . . .	13
4.1.6	Electrostatic Interactions . . . . .	14
4.1.7	Van der Waals Interactions . . . . .	15
4.1.8	The CHARMM Force Field – A Simple Molecular Me- chanics Force Field . . . . .	16
4.2	Data Structures . . . . .	17
4.2.1	Residue Topology File (RTF) . . . . .	18
4.2.2	Parameter File (PARAM) . . . . .	18
4.2.3	Protein Structure File (PSF) . . . . .	19
4.2.4	Coordinate File (CRD) . . . . .	19
<b>5</b>	<b>Energy Minimization</b>	<b>20</b>
5.1	Energy Minimization: Statement of the Problem . . . . .	20
5.2	Derivative Minimization Methods . . . . .	20
5.2.1	First-Order Minimization Methods . . . . .	21
5.2.2	A Second-Order Minimization Method - The Newton-Raphson Method (NR) . . . . .	23

<b>6</b>	<b>Molecular Dynamics (MD) and Normal Mode Analysis (NMA)</b>	<b>25</b>
6.1	Molecular Dynamics - Running A Molecular Dynamics Simulation	25
6.1.1	Starting Structure . . . . .	25
6.1.2	Modification of the Starting Structure . . . . .	25
6.1.3	Energy Minimization . . . . .	26
6.1.4	Heating Dynamics . . . . .	26
6.1.5	Equilibration and Rescaling Velocities . . . . .	28
6.1.6	Production Dynamics . . . . .	29
6.2	Normal Mode Analysis . . . . .	31
<b>7</b>	<b>What Else is Possible</b>	<b>37</b>
<b>A</b>	<b>Acknowledgements</b>	<b>39</b>

# List of Figures

4.1	Variation in bond energy with interatomic separation . . . . .	10
4.2	Comparison of the simple harmonic potential (Hooke's Law) with the Morse curve. . . . .	11
4.3	Bond Angle $\theta$ . . . . .	12
4.4	Variation in energy with rotation of the carbon-carbon bond in ethane. . . . .	13
4.5	A torsion angle (dihedral angle) A-B-C-D is defined as the angle Phi between the planes (ABC) and (BCD). A torsion angle can vary through 360 degrees. . . . .	13
4.6	The improper dihedral term is designed to maintain planarity about certain atoms. The potential is described by a harmonic function. $\alpha$ is the angle between the plane formed by the central atom and two peripheral atoms and the plane formed by the peripheral atoms only.) . . . . .	14
4.7	The Lennard-Jones potential. The collision parameter, $\sigma$ , is shown along with the well depth, epsilon. $r_{null}$ is the point of minimum energy. The dashed curves represent Paulirepulsion and van der Waals attraction. . . . .	16
4.8	Example of the RTF for the Alanine residue. . . . .	18
5.1	Quadratic Approximation at the Minimum. . . . .	21
5.2	Steepest Descent . . . . .	22
5.3	Line Search . . . . .	22
5.4	SD on a narrow valley . . . . .	23
5.5	Minimization No. 4 . . . . .	24
6.1	Aligned BPTI Structures shown in Cartoon Representation (Blue: Before Minimization; Red: After Minimization) . . . . .	26
6.2	Temperature vs. Time. . . . .	27
6.3	Total Energy vs. Time. . . . .	27
6.4	Kinetic Energy vs. Time . . . . .	28
6.5	Potential Energy vs. Time . . . . .	28
6.6	Total Energy vs. Time during Equilibration . . . . .	28
6.7	C-S-S Angle of Disulfide Bridge between CYS 5 and CYS 14 . . . . .	29
6.8	C-S-S-C Dihedral Angle of Disulfide Bridge between CYS 5 and CYS 14 . . . . .	29
6.9	RMSD during Production Run for different Parts of the Protein . . . . .	30
6.10	Summary of the proceeding of a Molecular Dynamics simulation . . . . .	31

6.11	As a molecule consisting of three atoms, water has three normal modes which are presented here. Experimental (and calculated) frequencies are shown. The first one has a significantly lower frequency $\nu$ than the others. Quantity $\nu$ is proportional to $\sqrt{\frac{k}{\mu}}$ with force constant $k$ and reduced mass $\mu$ . . . . .	32
6.12	The motion of a molecule around an energy minimum can be approximately described by a parabolic energy profile. This is the reason why one has to generate the energy-minimized structure (green ball), which is located around a minimum of the energy surface, before starting a normal mode calculation . . . . .	33
6.13	Vibrations in 2-dimensional space. In reality one more dimension comes into play . . . . .	34
6.14	A linear triatomic molecule like CO <sub>2</sub> . The vectors (here: scalars) $x_1, x_2, x_3$ define displacements of the corresponding atoms. . . .	35
6.15	Results of normal mode calculation for a linear triatomic molecule. $\lambda_i, \eta_i$ and $x_i$ describe eigenvalues, eigenvectors and amplitudes. .	36
6.16	Protein molecules are the most examined type of molecule with respect to vibrational motion. Obtaining the normal modes of motion, one can notice a difference between the motional frequency of bigger (global) parts and smaller (local) parts of the molecule, e.g. a whole domain or even just an atomic link between two distinct atoms. Global motions of a protein are often specific to it and can be related to its function. . . . .	36
7.1	A saddle point of the multidimensional reaction path corresponds to the transition state between reactants and products. . . . .	37

# Chapter 1

## Introduction

Molecular modeling is concerned with ways to describe the behavior of molecules and molecular systems. Computational techniques have revolutionized molecular modeling to the extent that calculations could not be performed without the use of a computer – molecular modeling is invariably associated with computer modeling or computational chemistry. This discipline encompasses not only quantum mechanics (QM) but also molecular mechanics (MM), conformational analyses and other computer-based methods for understanding and predicting the behavior of molecular systems.

In this article we describe the approach to molecular modeling with means of classical mechanics and the usefulness of empirical energy functions for investigating the physical and chemical properties of a wide variety of molecules. The CHARMM program [1] is explained in more detail since it presents a state-of-the-art computer program which can efficiently handle all aspects of computations with energy functions (force fields). In fact, MM-methods are also called force field methods. The potential energy of a molecular system is calculated and the changes of the system in time can be determined by means of integration algorithms. These computations are the core of a molecular dynamics (MD) simulation. After such a simulation, investigating the structures can start, e.g. calculations of deviation or fluctuation.

We would like to present the mathematical background of a MD simulation like calculations of the potential energy, of a minimized value for this energy, of the systems proceeding during a dynamic simulation or of normal modes of the molecular systems vibration. Some results of these operations dealing with a small protein called bovine pancreatic trypsin inhibitor (BPTI) are also presented as far as discussed before. Our goal is to make the way these methods work easier to understand with the help of our results on BPTI. We also explain the physical assumptions that help setting up a force field and the most important definitions and equations of statistical mechanics which mark a correlation between the analyses at a microscopic level and thermodynamic properties of a system. Finally, we will provide an overview of some other techniques that could be performed within CHARMM but we did not apply on BPTI.

# Chapter 2

# Classical Mechanics

## 2.1 Newton's Second Law

Molecular Dynamics simulations are based on Newton's second law, the equation of motion [2, 3]:

$$F_i = m_i \cdot a = m_i \cdot \frac{dv_i}{dt} = m_i \cdot \frac{d^2v_i}{dt^2}. \quad (2.1)$$

It describes the motion of a particle of mass  $m_i$  along the coordinate  $x_i$  with  $F_i$  being the force on  $m_i$  in that direction. This is used to calculate the motion of a finite number of atoms or molecules, respectively, under the influence of a *force field* that describes the interactions inside the system with a *potential energy function*,  $V(\vec{x})$ , where  $\vec{x}$  corresponds to the coordinates of all atoms in the system. The relationship of the potential energy function and Newton's second law is given by

$$F(\vec{x}_i) = -\nabla_i V(\vec{x}), \quad (2.2)$$

with  $F(\vec{x}_i)$  being the force acting on a particle due to a potential,  $V(\vec{x})$ . Combining these two equations gives

$$\frac{dV(\vec{x})}{dx_i} = -m_i \cdot \frac{d^2x_i}{dt^2}, \quad (2.3)$$

which relates the derivative of the potential energy to the changes of the atomic coordinates in time. As the potential energy is a complex multidimensional function this equation can only be solved numerically with some approximations.

With the acceleration being  $a = -\frac{1}{m} \cdot \frac{dV}{dx}$  we can then calculate the changes of the system in time by just knowing (i) the potential energy  $V(\vec{x})$ , (ii) initial coordinates  $x_{i,0}$  and (iii) an initial distribution of velocities,  $v_{i,0}$ . Thus, this method is deterministic, meaning we can predict the state of the system at any point of time in the future or the past.

The initial distribution of velocities is usually randomly chosen from a Gaussian or Maxwell-Boltzmann distribution [3], which gives the probability of atom  $i$  having the velocity in the direction of  $x$  at the temperature  $T$  by:

$$p(v_{i,x}) = \left( \frac{m_i}{2\pi k_b T} \right)^{\frac{1}{2}} \cdot \exp \left( -\frac{1}{2} \frac{m_i v_{i,x}^2}{k_b T} \right). \quad (2.4)$$

Velocities are then corrected so that the overall momentum of the system equals a zero vector:

$$P = \sum_{n=1}^N m_i \vec{v}_i = \vec{0}. \quad (2.5)$$

## 2.2 Integration Algorithms

The solution of the equation of motion given above is a rather simple one which is only sufficiently good over a very short period of time, in which the velocities and accelerations can be regarded as constant. So algorithms were introduced repeatedly performing small time steps, thus propagating the system's properties (positions, velocities and accelerations) in time. Time steps are typically chosen in the range of 1 fs [2]. It is necessary to use such a small time step, as many molecular processes occur in such small periods of time that they cannot be resolved with larger time steps. A time series of coordinate sets calculated this way is referred to as a *trajectory* and a single coordinate set as a *frame*. Today trajectories of about 10 ns, thus consisting of  $10^2$  frames, can be calculated.

### 2.2.1 Verlet Algorithm

All algorithms assume that the system's properties can be approximated by a Taylor series expansion [4]:

$$\vec{x}(t + \delta t) = \vec{x}(t) + \delta t \cdot \vec{v}(t) + \frac{1}{2} \delta t^2 \cdot \vec{a}(t) + \dots \quad (2.6)$$

$$\vec{v}(t + \delta t) = \vec{v}(t) + \delta t \cdot \vec{a}(t) + \frac{1}{2} \delta t^2 \cdot \vec{b}(t) + \dots \quad (2.7)$$

$$\vec{a}(t + \delta t) = \vec{a}(t) + \delta t \cdot \vec{b}(t) + \frac{1}{2} \delta t^2 \cdot \vec{c}(t) + \dots, \quad (2.8)$$

with  $\vec{x}$ ,  $\vec{v}$  and  $\vec{a}$  being the positions, the velocities and the accelerations of the system. The series expansion is usually truncated after the quadratic term. Probably the most widely used algorithm for integrating the equations of motion in MD simulations is the *Verlet algorithm* (Verlet 1967)[2, 3]. It can be derived by simply summing the Taylor expressions for the coordinates at the time  $(t + \delta t)$  and  $(t - \delta t)$ :

$$\vec{x}(t + \delta t) = \vec{x}(t) + \delta t \cdot \vec{v}(t) + \frac{1}{2} \delta t^2 \cdot \vec{a}(t) + \dots \quad (2.9)$$

$$\vec{x}(t - \delta t) = \vec{x}(t) - \delta t \cdot \vec{v}(t) + \frac{1}{2} \delta t^2 \cdot \vec{a}(t) - \dots \quad (2.10)$$

$$\Rightarrow \vec{x}(t + \delta t) = 2\vec{x}(t) - \vec{x}(t - \delta t) + \delta t^2 \cdot \vec{a}(t). \quad (2.11)$$

Thus, it uses the position  $\vec{x}(t)$  and acceleration  $\vec{a}(t)$  at time  $t$  and the positions from the previous step  $\vec{x}(t - \delta t)$  to calculate new positions  $\vec{x}(t + \delta t)$ . In this algorithm velocities are not explicitly calculated but can be obtained in several ways. One is to calculate mean velocities between the positions  $\vec{x}(t + \delta t)$  and  $\vec{x}(t - \delta t)$ .

$$\vec{v}(t) = \frac{1}{2\delta t} \cdot [\vec{x}(t + \delta t) - \vec{x}(t - \delta t)]. \quad (2.12)$$

The advantages of this algorithm are that it is straightforward and has modest storage requirements, comprising only two sets of positions  $[\vec{x}(t)$  and  $\vec{x}(t - \delta t)]$  and the accelerations  $\vec{a}(t)$ . The disadvantage, however, is its moderate precision, because the positions are obtained by adding a small term  $[\delta t^2 \cdot \vec{a}(t)]$  to the difference of two much larger terms  $[\vec{x}(t + \delta t) - \vec{x}(t - \delta t)]$ . This results in rounding errors due to numerical limitations of the computer.

Furthermore, this is obviously not a self-starting algorithm. New positions  $\vec{x}(t + \delta t)$  are obtained from the current positions  $\vec{x}(t)$  and the positions at the previous step  $\vec{x}(t - \delta t)$ . So at  $t = 0$  there are no positions for  $t - \delta t$  and therefore it is necessary to provide another way to calculate them. One way is to use the Taylor expansion truncated after the first term:

$$\vec{x}(t + \delta t) = \vec{x}(t) + \delta t \cdot \vec{v}(t) + \frac{1}{2} \delta t^2 \cdot \vec{a}(t) + \dots \quad (2.13)$$

$$\Rightarrow \vec{x}(t - \delta t) = \vec{x}(t) - \delta t \cdot \vec{v}(t) + \frac{1}{2} \delta t^2 \cdot \vec{a}(t) + \dots \quad (2.14)$$

## 2.2.2 Leap-Frog Algorithm

There are several variations of the Verlet algorithm trying to avoid its disadvantages. One example is the *leap-frog algorithm* [2, 3]. It uses the following equations:

$$\vec{v}(t + \frac{1}{2}\delta t) = \vec{v}(t - \frac{1}{2}\delta t) + \delta t \cdot \vec{a}(t) \quad (2.15)$$

$$\vec{x}(t + \delta t) = \vec{x}(t) + \delta t \cdot \vec{v}(t + \frac{1}{2}\delta t), \quad (2.16)$$

where  $\vec{a}(t)$  is obtained using

$$\vec{a}(t) = -\frac{1}{m} \cdot \frac{dV}{d\vec{x}}. \quad (2.17)$$

First, the velocities  $\vec{v}(t + \frac{1}{2}\delta t)$  are calculated from the velocities at  $(t - \frac{1}{2}\delta t)$  and the accelerations  $\vec{a}(t)$ . Then the positions  $\vec{x}(t + \delta t)$  are deduced from the velocities just calculated and the positions at time  $t$ . In this way the velocities first ‘leap-frog’ over the positions and then the positions leap over the velocities. The leap-frog algorithm’s advantages over the Verlet algorithm are the inclusion of the explicit velocities and the lack of the need to calculate the differences between large numbers.

An obvious disadvantage, however, is that the positions and velocities are not synchronized. This means it is not possible to calculate the contribution of the kinetic energy (from the velocities) and the potential energy (from the positions) to the total energy simultaneously.

### 2.2.3 Velocity Verlet Algorithm

The *velocity Verlet algorithm* (Swope *et al.* 1982)[2, 3] yields positions, velocities and accelerations at time  $t$  and does not compromise precision:

$$\vec{x}(t + \delta t) = \vec{x}(t) + \delta t \cdot \vec{v}(t) + \frac{1}{2} \delta t^2 \vec{a}(t) \quad (2.18)$$

$$\vec{v}(t + \delta t) = \vec{v}(t) + \frac{1}{2} \delta t [\vec{a}(t) + \vec{a}(t + \delta t)]. \quad (2.19)$$

For this algorithm more than two calculations have to be done for a single time step. This is due to the fact that calculation of the velocities  $\vec{v}(t + \delta t)$  requires acceleration values at  $(t)$  and  $(t + \delta t)$ . So first the positions at  $(t + \delta t)$  are calculated; then the velocities at time  $(t + \delta t)$  are computed using

$$\vec{v}(t + \delta t) = \vec{v}(t) + \frac{1}{2} \delta t \cdot \vec{a}(t + \delta t). \quad (2.20)$$

### Summary

We have now learned four examples of integration algorithms. But what could make us prefer one over another? As for any other computer algorithm the ideal method should be fast, which means computationally efficient, require as little memory as possible and be easy to program. These however are not the main features you should examine. They are of rather secondary interest for most MD simulations because most algorithms do not demand significant storage amount and calculations for the integration are rather fast in comparison to other calculations in a simulation such as the calculation of the force acting on every single atom in the system. Thus, other features are considered first: The algorithm should conserve overall momentum and energy, be time-reversible and permit a long time step without great loss of precision.

The choice of the integration step size, in fact, is very important. One must weigh the increased accuracy of using a small step size against the longer real time that can be simulated when a larger step size is used.

## Chapter 3

# Statistical Mechanics

MD simulations provide information at the microscopic level. Statistical mechanics are then required to convert this microscopic information to macroscopic observables such as pressure, energy, heat capacities, etc. Statistical mechanics relate these macroscopic observables to the distribution of molecular positions and motions. Therefore, time independent statistical averages are introduced. For a better understanding some definitions are reviewed here [3, 5]:

### 3.1 Definitions

**The *mechanical or microscopic state*** of a system is defined by the atomic positions  $x_i$  and the momenta  $p_i = m_i \cdot v_i$ . They can be considered as a multidimensional space with  $6N$  coordinates, for which they both contribute  $3N$  coordinates. This space is called *phase space*.

**The *thermodynamic or macroscopic state*** of a system is defined by a set of parameters that completely describes all thermodynamic properties of the system. An example would be the temperature  $T$ , the pressure  $P$ , and the number of particles  $N$ . All other properties can be derived from the fundamental thermodynamic equations.

**An *ensemble*** is the collection of all possible systems which have different microscopic states but have the same macroscopic or thermodynamic state. Ensembles can be defined by fixed thermodynamic properties as already stated before. Examples for ensembles with different characteristics are:  
 $NVE, NVT, NPT, \mu VT$ ,  
( $E$  = total energy,  $P$  = pressure,  $V$  = volume,  $\mu$  = chemical potential)

### 3.2 Ensemble Averages and Time Averages

In an experiment one examines a macroscopic sample with an enormously high number of atoms or molecules respectively. So the measured thermodynamic properties reflect an extremely large number of different conformations of the system, representing a subset of the ensemble. We have to say subset, because

an ensemble is the complete collection of microscopic systems and a macroscopic sample can only consist of a finite number of systems. A sufficiently big sample, however, can be seen as good approximation to an ensemble. That is why statistical mechanics defines averages corresponding to experimentally measured thermodynamic properties as *ensemble averages* [3, 5].

The ensemble average is given by:

$$\langle A \rangle_{ensemble} = \iint d\vec{p}^N d\vec{x}^N A(\vec{p}^N, \vec{x}^N) \rho(\vec{p}^N, \vec{x}^N), \quad (3.1)$$

where  $\langle A \rangle$  is the measured observable, which is stated as a function of the momenta  $\vec{p}_i$  and the positions  $\vec{x}_i$ . Quantity  $\rho(\vec{p}^N, \vec{x}^N)$  is the probability density for the ensemble and the integration is performed over all momenta and positions of the system  $d\vec{p}^N, d\vec{x}^N$ . So, the ensemble average is the average value of an observable weighted with its probability.

This integral is extremely difficult to calculate as it involves calculating all possible states of the system.

In an MD simulation an extremely large number of conformations is generated sequentially in time. To calculate an ensemble average the simulation has to cover all possible conformations corresponding to the ensemble, at which the simulation takes place. The *time average* [3] is given by

$$\langle A \rangle_{time} = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A(\vec{p}^N, \vec{x}^N) dt, \quad (3.2)$$

where  $\tau$  is the simulation time.

This expression is well approximated by an average over a simulation performed over a sufficiently long period of time and so representing a sufficient amount of phase space:

$$\langle A \rangle_{time} = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A(\vec{p}^N, \vec{x}^N) dt = \frac{1}{M} \sum_{i=1}^M A(\vec{p}_M^N, \vec{x}_M^N), \quad (3.3)$$

where  $M$  is the number of frames and  $A(\vec{p}_M^N, \vec{x}_M^N)$  the values of the observable  $A$  in frame  $M$ .

The idea is based on the *Ergodic Hypothesis* [5], one of the most fundamental axioms of statistical mechanics, which states that the time average equals the ensemble average:

$$\langle A \rangle_{ensemble} = \langle A \rangle_{time}. \quad (3.4)$$

The idea, as already indicated above, is to simulate the system for a relatively long time, so that it will pass through an extremely high number of conformations, which can then be referred to as a representative subset of an ensemble.

# Chapter 4

## CHARMM

The CHARMM (Chemistry at HARvard Molecular Mechanics) program is a general purpose molecular mechanics simulation program. Besides energy minimization, dynamics simulation, vibrational analysis and thermodynamic calculations which are all performed with the use of the empirical potential energy function (the force field), there are interfaces to several quantum mechanical programs allowing mixed QM and MM calculations. The program can treat systems ranging in size from small individual organic molecules to large proteins and DNA molecules either isolated, in solutions or in crystalline solids. In this section we explicitly describe the core of the CHARMM program, the force field, and some important files which the program works with. We start with an introduction into characteristics of force fields before dealing with the single portions of the function.

### 4.1 Force Fields

#### 4.1.1 General Features of Molecular Mechanics Force Fields

Force fields enable the potential energy of a molecular system  $V$  to be calculated rapidly. The energy is a function of the atomic positions of all the atoms in the system which are usually expressed in term of Cartesian coordinates. Unlike quantum mechanical methods that deal with the electrons in a system, force field techniques calculate the energy of a system only as a function of the nuclear positions. This is legitimated by the Born-Oppenheimer-Approximation [6]. Thus, Molecular Mechanics are invariably used to perform calculations on systems containing a significant number of atoms which would bring enormous quantum mechanical calculations with them.

A typical force field represents each atom in the system as a single point and energies as a sum of two-, three-, and four-atom interactions such as bond stretching and angle bending. Although, simple functions (e.g. Hooke's Law) are used to describe these interactions, the force field can work quite well. The potential energy of a certain interaction is described by an equation which involves the positions of the particles and some parameters (e.g force constants or reference values) which have been determined experimentally or by quantum mechanical calculations.

Several types of force fields exist. Two of those may use an identical functional form yet have very different parameters and thus bring about different energies for the same system. Moreover, force fields with the same functional form but different parameters, and force fields with different functional forms, may give close results. A force field should be considered as a single entity; it doesn't need to be correct to divide the energy into its individual components or even to take some of the parameters from one force field and mix them with parameters from another one.

An important point that one shouldn't forget is that no 'correct' form for a force field exists. If one functional form performs better than another, that form will be favored. Most of the force fields commonly used do have a very similar form – we will discuss this particular form in more detail later on – but it should always be kept in mind that there may be better functional forms, particularly when developing a force field for new classes of molecules. molecular mechanics force fields are often a compromise between accuracy and computational efficiency; the most accurate ones may often be unsatisfactory for efficient computation. As the performance of computers increases, it becomes possible to incorporate more sophisticated models.

A concept that is common to most force fields is that of an *atom type*. For a quantum mechanics calculation it is usually necessary to specify the charge of the nuclei, together with the geometry of the system and the overall charge and spin multiplicity. For a force field calculation, however, the overall charge and spin multiplicity are not explicitly required, but it is usually necessary to assign an atom type to each atom in the system. This contains information about its hybridization state and sometimes the local environment. For example, it is necessary to distinguish between carbon atoms which adopt a tetrahedral geometry (sp<sup>3</sup>-hybridized), which are trigonal (sp<sup>2</sup>- hybridized) and carbons which are linear (sp-hybridized). The corresponding parameters are expressed in terms of these atom types, so that the reference angle  $\Theta_0$  for a tetrahedral carbon atom would be about 109.5° and that for a trigonal carbon near 120. For example, the MM2 [7], MM3 [8, 9, 10] and MM4 [11, 12, 13, 14, 15] force fields of Allinger and co-workers, that are widely used for calculations on "small" molecules, distinguish the following types of carbon atoms: sp<sup>3</sup>, sp<sup>2</sup>, sp, carbonyl, cyclopropane, radical, cyclopropene and carbonium ion. The value of the potential energy  $V$  is calculated as a sum of internal or bonded terms, which describe the bonds, angles and bond rotations in a molecule, and a sum of external or non-bonded terms, which account for interactions between non-bonded atoms or atoms separated by three or more covalent bonds. So it is:

$$V(\vec{x}) = V_{bonded}(\vec{x}) + V_{non-bonded}(\vec{x}), \quad (4.1)$$

Let us now discuss the individual contributions to a molecular mechanics force field, giving a selection of the various functional forms that are in common use. We shall then have a look at the CHARMM force field which is used for our calculations on BPTI.

#### 4.1.2 Bond Stretching

The potential energy curve (Morse potential) for a typical bond has the shape shown in Fig.4.1. This potential has the form:

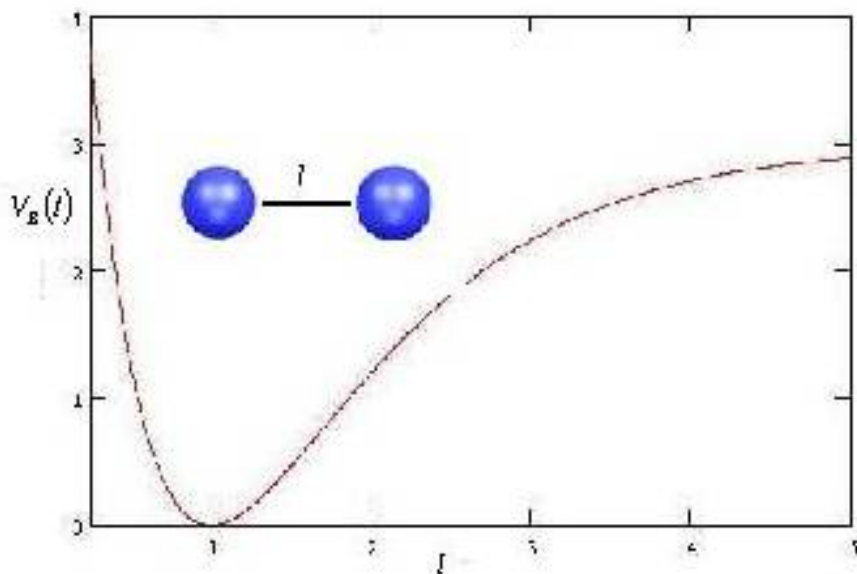


Figure 4.1: Variation in bond energy with interatomic separation

$$V_B(l) = D_e \{1 - \exp[-a(l - l_0)]\}^2. \quad (4.2)$$

$D_e$  is the depth of the potential energy minimum and  $a = \omega \cdot \sqrt{\frac{\mu}{2D_e}}$ , where  $\mu$  is the reduced mass and  $\omega$  is the frequency of the bond vibration. The frequency  $\omega$  is related to the stretching constant of the bond  $k_l$ , by  $\omega = \sqrt{\frac{k_l}{\mu}}$ , where  $k_l$  determines the strength of the bond. The length  $l_0$  is the reference bond length. It is the value that the bond adopts when all other terms in the force field are zero. Both  $l_0$  and  $k_l$  are specific for each pair of bound atom. Values of  $k_l$  are often evaluated from experimental data such as infrared stretching frequencies or from quantum mechanical calculations. Values of  $l_0$  can be inferred from high resolution crystal structures or microwave spectroscopy data.

The Morse potential is not usually used in molecular mechanics force fields. It is computationally demanding the curve describes a wide range of behavior, from the strong equilibrium to dissociation. Normally, this is not necessary for Molecular Mechanics calculations where we are more interested in slight deviations of bonds from their equilibrium values. Consequently, simpler expressions are often used. The most elementary approach is to use a Hooke's Law formula in which the energy varies with the square of the displacement from the reference bond length  $l_0$ :

$$V_B(l) = \frac{1}{2}k_l \cdot (l - l_0)^2 \quad (4.3)$$

The Hooke's Law functional form is a reasonable approximation to the shape of the potential energy curve at the bottom of the potential well, at distances

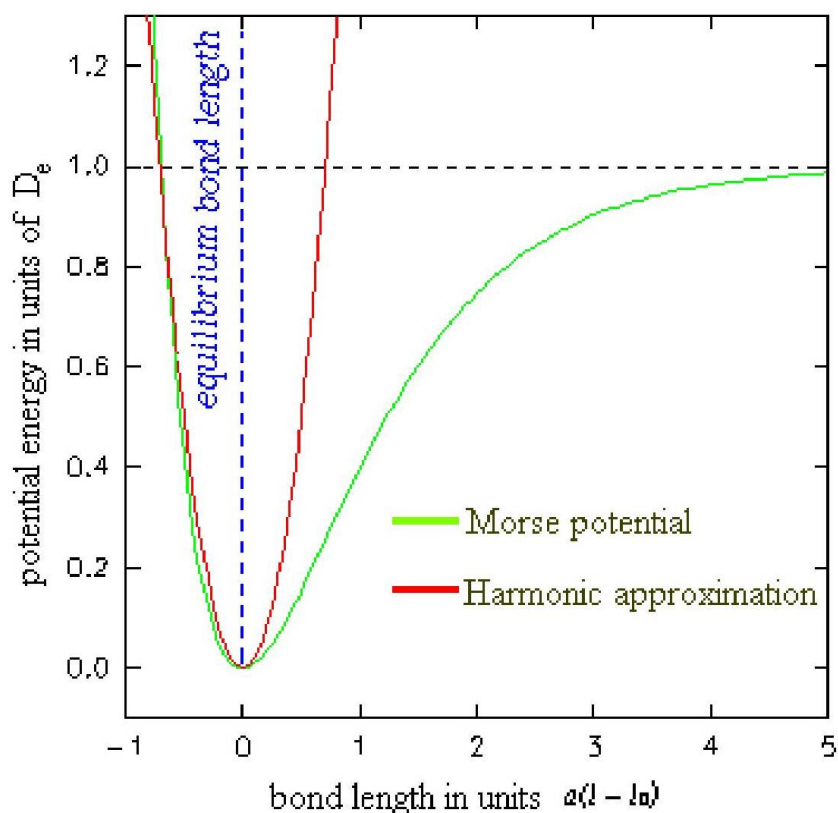


Figure 4.2: Comparison of the simple harmonic potential (Hooke's Law) with the Morse curve.

that correspond to bonding in ground-state molecules. It is less accurate away from equilibrium (Fig.4.2).

To approximate the Morse curve more accurately, cubic and higher terms can be included which give a better model close to the equilibrium structure than the quadratic form but also create a potential passing through maxima further away from  $l_0$ . This can lead to a catastrophic lengthening of bonds.

### 4.1.3 Angle Bending

The deviation of valence angles from their reference values (Fig.4.3) is also frequently described using a harmonic potential:

$$V_A(\Theta) = \frac{1}{2} \cdot k_\Theta \cdot (\Theta - \Theta_0)^2 \quad (4.4)$$

The force constant  $k_\Theta$  and the reference value  $\Theta_0$  depend on the chemical type of atoms constituting the angle. Rather less energy is required to distort an angle away from equilibrium than to stretch or compress a bond, and thus force

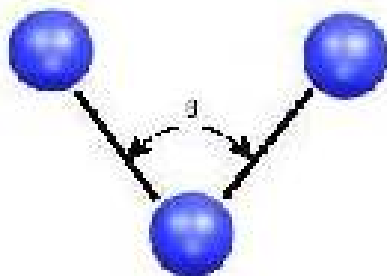


Figure 4.3: Bond Angle  $\theta$

constants are smaller here. As with the bond-stretching terms, the accuracy of the force field can be improved by the incorporation of higher-order terms. These two terms, the bond-stretching and angle-bending, describe the deviation from an ideal geometry; effectively, they are penalty functions and the sum of them should be close to zero in a perfectly optimized structure. These two terms are often regarded as hard degrees of freedom, in that quite substantial energies are required to cause significant deformations from their reference values. Most of the variation in structure and relative energies is due to torsional and non-bonded contributions.

#### 4.1.4 Torsional Terms

This term represents the torsion angle potential function which models the presence of steric barriers between atoms separated by three covalent bonds. The existence of barriers to rotation about chemical bonds is fundamental to understanding the structural properties of molecules and conformational analysis. The three minimum energy staggered conformations and the three maximum energy eclipsed structures of ethane (Fig.4.4) are a classic example of the way in which the energy changes with a bond rotation [16]. Quantum mechanical calculations suggest that the rotation-barrier arises from antibonding interactions between the H-atoms on opposite ends of the molecule; they are minimized when the conformation is staggered. Not all molecular mechanics force fields use torsional potentials; it may be possible to rely on non-bonded interactions between the atoms at the end of each torsion angle. However, most force fields for organic molecules do use explicit torsional potentials with a contribution from each bonded quartet of atoms A-B-C-D in the system (Fig.4.5). Thus, there would be nine individual torsional terms for ethane. Torsional potentials are almost always expressed as a cosine series expansion. One functional form is:

$$V_T(\Phi) = \frac{1}{2}k_\Phi \cdot [1 + \cos(n\Phi - \delta)] \quad (4.5)$$

The quantity  $k_\Phi$  is often referred to as the barrier height, but to do so is misleading, obviously so when more than one term is present in the expansion. Moreover, other terms in the force field equation contribute to the barrier height

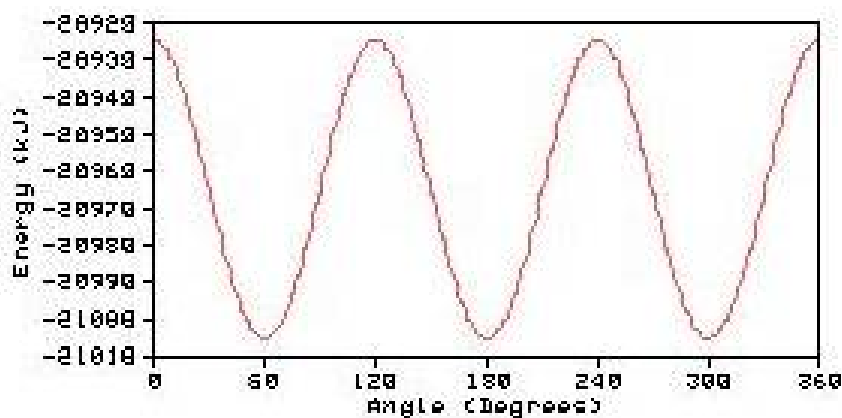


Figure 4.4: Variation in energy with rotation of the carbon-carbon bond in ethane.

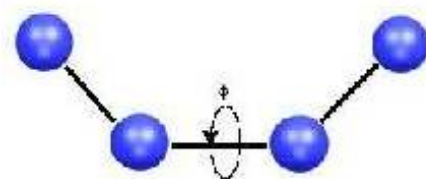


Figure 4.5: A torsion angle (dihedral angle) A-B-C-D is defined as the angle  $\Phi$  between the planes (ABC) and (BCD). A torsion angle can vary through 360 degrees.

as a bond is rotated, especially the non-bonded interactions between atom A and D. The value of  $k_{\Phi}$  does, however, give a qualitative indication of the relative barriers to rotation; for example  $k_{\Phi}$  for an amide bond (A-C=N-D) will be larger than for a bond between two  $sp^3$  carbon atoms (A-C-C-D).  $n$  is the periodicity; its value gives the number of minimum points in the function as the bond is rotated through  $360^\circ$ .  $\delta$  (the phase factor) sets the minimum energy angle.

#### 4.1.5 Improper Torsions / Out-of-Plane Bending

Several chemical groups involve arrangements of 4 or more atoms in a plane. For these groups it is sometimes advantageous to have an additional term enforcing planarity. For example, it is found experimentally that the oxygen atom of cyclobutanone remains in the plane of the ring. This is because the  $\pi$ -bonding energy, which is maximized in the planar arrangement, would be much reduced if the oxygen were bent out of the plane. Using a force field containing just stan-

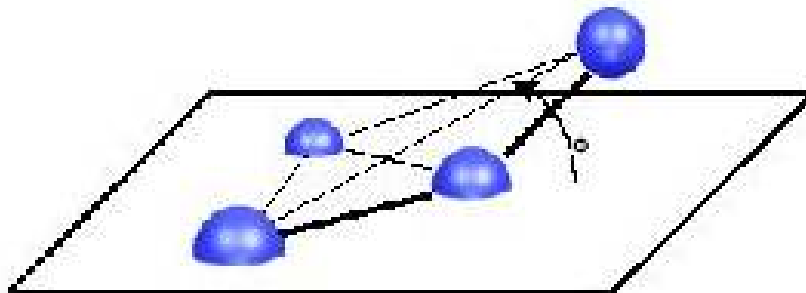


Figure 4.6: The improper dihedral term is designed to maintain planarity about certain atoms. The potential is described by a harmonic function.  $\alpha$  is the angle between the plane formed by the central atom and two peripheral atoms and the plane formed by the peripheral atoms only.)

With the standard terms we already know, the equilibrium structure would have the oxygen atom located out of the plane formed by the adjoining carbon atom and the two carbon atoms bonded to it. The simplest way to achieve the desired geometry is to use an out-of-plane bending term. One approach is to treat the four atoms as an improper torsion angle i.e., a torsion angle in which the four atoms are not bonded in the sequence A-B-C-D. Another way involves a calculation of the angle between a bond from the central atom and the plane defined by the central atom and the other two atoms (Fig.4.6). A value of  $0^\circ$  corresponds to all four atoms being planar. With these definitions the deviation of the out-of-plane coordinate can be modeled using a harmonic potential of the form:

$$\nu(\alpha) = \frac{1}{2}k_\alpha \cdot \alpha^2. \quad (4.6)$$

The improper torsion or improper dihedral definition is more widely used as it can then be easily included with the ‘proper’ torsional terms in the force field. However, the other form may be better to implement out-of-plane bending in the force field.

After learning the most important bonded terms of the energy function we are now going to have a look at the non-bonded terms which consist at least of the electrostatic and the van der Waals interactions in the system.

#### 4.1.6 Electrostatic Interactions

Interactions between atoms due to their permanent dipole moments are described approximately by treating the charged portions as point charges. Then we use the Coulomb potential for point charges to estimate the forces between the charged atoms. The Coulomb potential is an effective pair potential that describes the interaction between two point charges. It acts along the line connecting the two charges. It is given by the equation:

$$V_E(i, j) = \frac{q_i q_j}{4\pi\epsilon_0 r_0}. \quad (4.7)$$

$r_{ij}$  is the distance between  $q_i$  and  $q_j$ , the electric charge in coulombs carried by charge  $i$  and  $j$  respectively, and  $\epsilon_0$  is the electrical permittivity of space. Alternative approaches to the calculation of electrostatic interactions, e.g. the *central multipole expansion* which is based on the electric moments, may provide more exact solutions to the electrostatic interactions [17, 18].

#### 4.1.7 Van der Waals Interactions

The van der Waals interaction between two atoms arises from a balance between repulsive and attractive forces. The repulsive force arises at short distances where the electron-electron interaction is strong (*Pauli repulsion*). The attractive force, also referred to as dispersion force, arises from fluctuations in the charge distribution in the electron clouds. The fluctuation in the electron distribution on one atom gives rise to an instantaneous dipole which, in turn, induces a dipole in a second atom giving rise to an attractive interaction. Each of these two effects is equal to zero at infinite atomic separation and becomes significant as the distance decreases. The attractive interaction is longer range than the repulsion but as the distance becomes short, the repulsive interaction becomes dominant. This gives rise to a minimum in the energy (see Fig.4.7). For a force field we require a means to model the interatomic potential curve accurately, using a simple empirical expression that can be rapidly calculated. The best known of the van der Waals potential functions is the *Lennard-Jones 12-6 function*, which takes the following form for the interaction between two atoms:

$$V_{vdW}(r) = 4\epsilon \cdot \left[ \left( \frac{\sigma}{r} \right)^{12} - \left( \frac{\sigma}{r} \right)^6 \right]. \quad (4.8)$$

The Lennard-Jones 12-6 potential contains just two adjustable parameters: the collision diameter  $\sigma$  (the separation for which the energy is zero) and the well depth  $\epsilon$ . These parameters are graphically illustrated in Fig.4.7.

The need for a function that can be rapidly evaluated is a consequence of the large number of van der Waals interactions that must be determined in many of the systems that we would like to model. The 12-6 potential is widely used, particularly for calculations on large systems, as r-12 can be quickly calculated by squaring the r-6 term. The r-6 term can also be calculated from the square of the distance without having to perform a computationally expensive square root calculation. Different powers have also been used for the repulsive part of the potential [19]; values of 9 or 10 give a less steep curve and are used in some force fields.

The 6-12 term between hydrogen-bonding atoms is replaced by an explicit hydrogen-bonding term in some force fields, which is often described using a Lennard-Jones 10-12 potential. This function is used to model the interaction between the donor hydrogen atom and the heteroatom acceptor atom. Its use is intended to improve the accuracy with which the geometry of hydrogen-bonding systems is predicted.

The most time consuming part of a molecular dynamics simulation is the calculation of the non-bonded terms in the potential energy function. In principle, these energy terms should be evaluated between every pair of atoms; in this

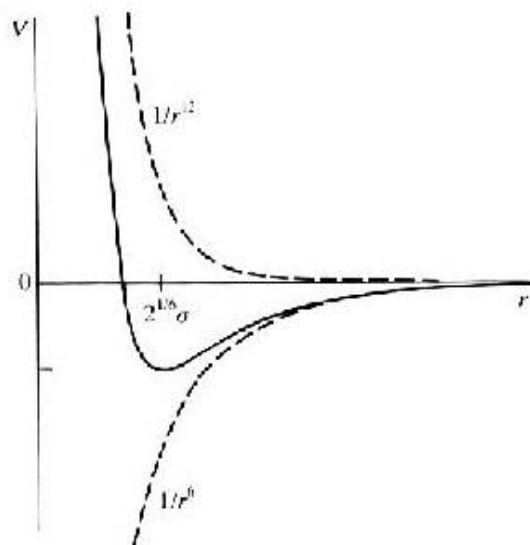


Figure 4.7: The Lennard-Jones potential. The collision parameter,  $\sigma$ , is shown along with the well depth, epsilon.  $r_{null}$  is the point of minimum energy. The dashed curves represent Pauli repulsion and van der Waals attraction.

case, the number increases as the square of the number of atoms for a pair-wise model. To speed up the computation two approaches are applied. In the first approach the interactions between two atoms separated by a distance greater than a pre-defined distance, the *cutoff distance*, are ignored. The interactions are simply set to zero for interatomic distances greater than the cutoff distance (Truncation-Method) or the entire potential energy surface is modified such that at the cutoff distance the interaction potential is zero (Shift-Method). The other way is to reduce the number of interaction sites. The simplest way to do this is to subsume some or all of the atoms (usually just the hydrogen atoms) into the atoms to which they are bonded (United-Atom-Method). Considerable computational savings are possible; for example, if butane is modeled as a four-site model rather than one with twelve atoms, the van der Waals interaction between all the atoms involves the calculation of six terms rather than 78.

#### 4.1.8 The CHARMM Force Field – A Simple Molecular Mechanics Force Field

Now that we know how the different contributions to a force field can be described, let us look at the simplest form such a force field can have, the CHARMM force field:

$$\begin{aligned}
V_{tot} = & \frac{1}{2} \sum_{bonds} k_l \cdot (l - l_0)^2 + \frac{1}{2} \sum_{angles} k_\Theta \cdot (\Theta - \Theta_0)^2 + \\
& + \frac{1}{2} \sum_{torsions} k_\Phi \cdot [1 + \cos(n\Phi - \delta)] + \frac{1}{2} \sum_{impropers} k_\alpha \cdot \alpha^2 + \\
& + \sum_{i,j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} + \sum_{ij} 4\epsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^6 \right].
\end{aligned} \tag{4.9}$$

Let us consider how this simple force field would be used to calculate the energy of a conformation of a simple molecule. Propane is one that is most popular for this task [2] it has more terms than ethane for example, and it is not as complicated as butane (butane has far more non-bonded and torsional energy terms). As it is explained by Leach, propane has ten bonds: two C-C and eight C-H bonds. The C-C bonds are symmetrically equivalent but the C-H bonds fall into two classes, one group corresponding to the two hydrogens bonded to the central methylene (CH<sub>2</sub>) carbon and one group corresponding to the six hydrogens bonded to the methyl carbons. In some sophisticated force fields different parameters would be used for these two different types of C-H bond, but in the CHARMM force field the same bonding parameters (i.e.  $k_l$  and  $l_0$ ) would be used for each of the eight C-H bonds. This is a good example for transferability since the same parameters can be used for a wide variety of molecules. There are 18 different valence angles in propane, comprising one C-C-C angle, ten C-C-H angles and seven H-C-H angles. Note that all angles are included in the CHARMM force field even though some of them may not be independent of the others. There are 18 torsional terms: twelve H-C-C-H torsions and six H-C-C-C torsions. They are modeled with a cosine series expansion having minima at the trans and gauche conformations. The improper dihedral term is dropped out for propane. Finally, there are 27 non-bonded terms to calculate, comprising 21 H-H interactions and six H-C interactions. A sizeable number of terms are thus included in the CHARMM model, even for a molecule as simple as propane. Even so, the number of terms, namely 73 is much less than the number of integrals that would be involved in an equivalent quantum mechanical calculation.

The force field equation given above is only one variant of the many CHARMM force fields. There are other potential energy functions which contain more terms for a more precise calculation, implemented in the CHARMM program, e.g. the extended electrostatics model [20] or the fast multipole method [21] for treating long-range electrostatic interactions with a multipole approximation. Our calculations on BPTI, however, were performed with the force field presented above.

## 4.2 Data Structures

Data Structures include information about the molecule, its composition, its chemical connectivity, certain atomic properties and parameters for the energy function and more. This information for a particular class of molecules, e.g. proteins or nucleic acids, is contained in the topology file and the parameter file.

```

RESI ALA0.00
GROU
ATOM NNH1-0.47 ! |
ATOM HNH 0.31 !HN-N
ATOM CACT1 0.07! | HB1
ATOM HAHB 0.09 ! | /
GROUP !HA-CA--CB-HB2
ATOM CBCT3-0.27! | \
ATOM HB1HA 0.09! | HB3
ATOM HB2HA 0.09! O=C
ATOM HB3HA 0.09! |
GROUP
ATOM CC 0.51
ATOM OO-0.51
BONDCB CA N HN N
BOND C CA C +N CA HA CB HB1 CB HB2 CB HB3
DOUBLE O
IMPR N -C CA HN C CA +N O
DONOR HN
ACCEPTOR O
IC -C CA *N HN 1.3551 126.4900 180.0000 115.4200 0.9996
IC -C N CA C 1.3551 126.4900 180.0000 114.4400 1.5390
IC N CA C +N 1.4592 114.4400 180.0000 116.8400 1.3558
IC +N CA *C O 1.3558 116.8400 180.0000 122.5200 1.2297
IC CA C +N +CA 1.5390 116.8400 180.0000 126.7700 1.4613
IC N C *CA CB 1.4592 114.4400 123.2300 111.0900 1.5461
IC N C *CA HA 1.4592 114.4400 -120.4500 106.3900 1.0840
IC C CA CB HB1 1.5390 111.0900 177.2500 109.6000 1.1109
IC HB1 CA *CB HB2 1.1109 109.6000 119.1300 111.0500 1.1119
IC HB1 CA *CB HB3 1.1109 109.6000 -119.5800 111.6100 1.1114

```

Figure 4.8: Example of the RTF for the Alanine residue.

For a specific molecule, the necessary data is stored in the Protein Structure File and the Coordinate File, respectively.

#### 4.2.1 Residue Topology File (RTF)

The RTF contains local information about atoms, bonds, angles etc. for each possible type of monomer unit (residue) that can be used in building a particular type of macromolecule. For each residue the covalent structure is defined, i.e., how the atoms are connected to one another to form amino acids, DNA bases or lipid molecules. Fig.4.8 depicts the RTF for the amino acid Alanine.

#### 4.2.2 Parameter File (PARAM)

The parameter File is associated with the RTF file as it contains all the necessary parameters for calculating the energy of the molecule(s). These include the equilibrium bond lengths and angles for bond stretching, angle bending and dihedral angle terms in the potential energy function as well as the force constants and the Lennard-Jones 12-6 potential parameters. As already mentioned, these parameters are associated with particular atom types.

### 4.2.3 Protein Structure File (PSF)

The PSF is the most fundamental data structure used in CHARMM. It is generated for a specific molecule or molecules and contains the detailed composition and connectivity of the molecule(s). It describes how molecules are divided into residues and molecular entities (segments), which can range from a single macromolecular chain to multiple chains solvated by explicit water molecules. The PSF must be specified before any calculations can be performed on the molecule. The PSF constitutes the molecular topology but does not contain information regarding the bond lengths, angles, etc., so it is necessary to read in the parameter file to add the missing information.

### 4.2.4 Coordinate File (CRD)

The CRD file contains the Cartesian coordinates of all atoms in the system. Those are most often obtained from x-ray crystal structures or from NMR structures. Missing coordinates can be built within the CHARMM program; in addition, hydrogen atoms which are not present in x-ray crystal structures of proteins can be placed using CHARMM. Two sets of coordinates can also be accommodated in the program which is useful for a variety of calculations, e.g. the *RMSD (Root Mean Square Deviation)* between the experimental structure and a structure from the simulation (see Chapter 6.1).

## Chapter 5

# Energy Minimization

### 5.1 Energy Minimization: Statement of the Problem

Given is a multidimensional function  $V$  which depends on several variables  $x_1, x_2, x_3, \dots, x_i$ . The problem is to find the values of these variables where  $V$  gives minimum values. Minimum values are those points in an energy landscape where you will reach higher positions in any direction. So the first derivative of the function with respect to every single variable is zero:

$$\frac{\partial V}{\partial x_i} = 0, \quad (5.1)$$

with  $x_i$  and  $x_j$  being the coordinates of the simulated atoms; and the matrix  $\frac{\partial^2 V}{\partial x_i^2}$  is positive definite.

According to this, minimization methods can be classified into two main groups: Those algorithms which use derivatives and those which do not [2]. Derivatives provide useful information about the shape of the energy landscape and so can significantly accelerate the search process. That is why methods which use derivatives are more appropriate to minimize the complex functions describing the energy landscape of a molecule.

For this reason we will solely concentrate on these methods in the following.

### 5.2 Derivative Minimization Methods

The direction of the first derivative, the gradient, of a function tells us in which way we will find a minimum and its magnitude indicates the steepness of the slope at a given point of the function. The second derivative provides information about the curvature of the energy landscape.

To describe derivative methods it is useful to write the energy function as a Taylor expansion around point  $x_k$  (that we write first for the 1-dimensional case) [4]:

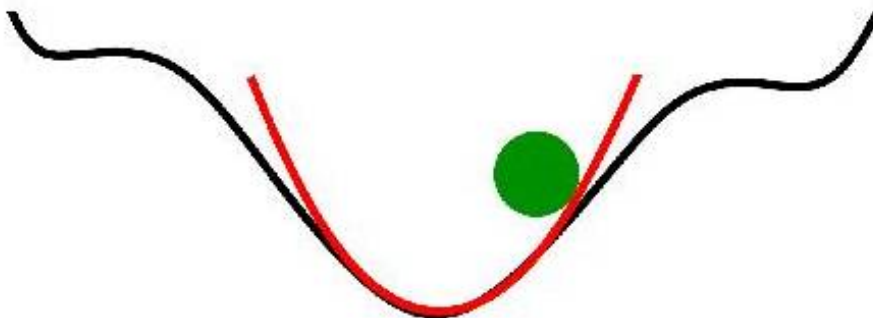


Figure 5.1: Quadratic Approximation at the Minimum.

$$V(x) = V(x_k) + (x - x_k) \cdot V'(x_k) + \frac{1}{2}[(x - x_k)^2 \cdot V''(x_k)] + \dots, \quad (5.2)$$

where  $V'$  is the first derivative and  $V''$  the second derivative of the energy function  $V$ .

In the case of a multidimensional function the variable  $x$  corresponds to a vector  $\vec{x}$  and the derivatives are replaced by matrices: For a system with  $N$  atoms  $V(\vec{x})$  is a function of  $3N$  coordinates [2]. So  $\vec{x}$  has  $3N$  components and the gradient,  $\vec{g} = V'(\vec{x})^T$ , accordingly is a vector with  $3N$  dimensions as well, with each element being the partial derivative of  $V$  with respect to a single coordinate,  $\frac{\partial V}{\partial x_i}$ . The second derivative  $V''(\vec{x})$  is a  $(3N \times 3N)$ -matrix. Every element  $(i, j)$  corresponds to the partial second derivate of  $V$  with respect to the coordinates  $x_i$  and  $x_j$ ,  $\frac{\partial^2 V}{\partial x_i \partial x_j}$ . This is a symmetric matrix which is called Hessian Matrix. Thus the multidimensional Taylor expansion is written as follows [22]:

$$V(\vec{x}) = V(\vec{x}_k) + V'(\vec{x}_k) \cdot (\vec{x} - \vec{x}_k) + \frac{1}{2}[(\vec{x} - \vec{x}_k)^T \cdot V''(\vec{x}_k) \cdot (\vec{x} - \vec{x}_k)] + \dots \quad (5.3)$$

This is a quadratic function and thus can only be seen as an approximation for an energy function. However, the area close to a minimum is well approximated by this Taylor expansion as can be seen in Fig.5.1

In the following we will discuss the most important and most frequently used methods for energy minimization in molecular modelling. They can be classified according to the highest order derivative used.

### 5.2.1 First-Order Minimization Methods

First-Order minimization methods use the information of the first derivative, the gradient  $\vec{g}_k$ , to find local minima of the function of interest.

#### Steepest Descent (SD)

The steepest descent method moves along the negative gradient  $-\vec{g}_k$  downhill the energy landscape beginning from a starting point of interest. As all these

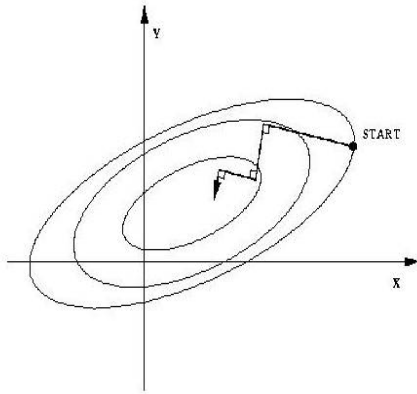


Figure 5.2: Steepest Descent

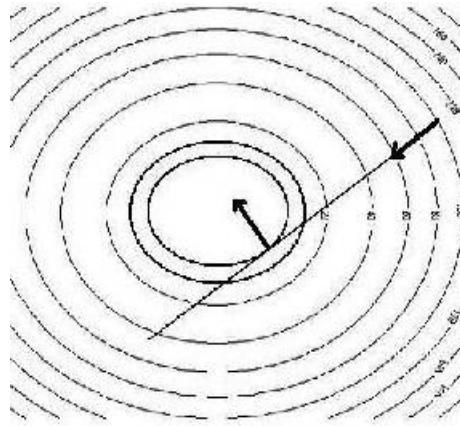


Figure 5.3: Line Search

methods are iterative methods the search is done stepwise:

Calculate the gradient  $\vec{g}_k$  at a starting point  $\vec{x}_k$  and take a step in the direction of the negative gradient  $-\vec{g}_k$ , the point of arrival being the starting point for the next iteration.

Knowing the direction of the gradient the next thing to do is to determine the length of the step to take. This is achieved by performing a line search in the direction of the gradient [2]. Imagine a cross-section through the energy function in direction of  $\vec{g}_k$  and you see that there will be a (one-dimensional) minimum, which is the optimal starting point for the next step,  $\vec{g}_{k+1}$ .

If you take the next step from this minimum point, the directions,  $\vec{v}_k$ , and the gradients,  $\vec{g}_k$ , of successive steps will always be orthogonal [2]:

$$\vec{g}_k \cdot \vec{g}_{k+1} = 0 \quad (5.4)$$

$$\vec{v}_k \cdot \vec{v}_{k-1} = 0 \quad (5.5)$$

Finding the minimum via line search can be achieved in several ways. One possibility, for example, is to perform a one dimensional quadratic approximation in the direction of the gradient.

### Conjugate Gradient Method

The SD approach encounters severe problems when used with functions other than quadratic functions especially when they have the shape of a narrow valley. In this case it gives undesirable behaviour [22] taking only very short steps and thus extremely increasing computation (see Fig.5.4).

That is why another, similar approach has been introduced: the conjugate gradient method. In contrast to SD, in this method the gradients of successive steps are not orthogonal. Instead the directions at each point  $\vec{v}_k$  are orthogonal to gradients of the following steps  $\vec{g}_k$  provided that a line search has been performed:

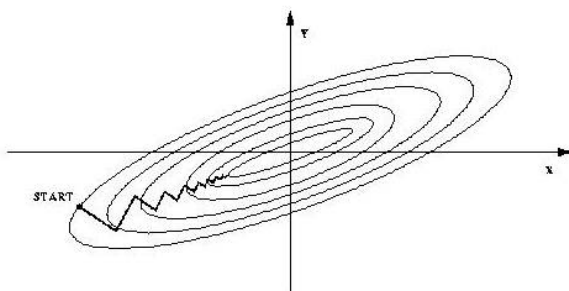


Figure 5.4: SD on a narrow valley

$$\vec{\nu}_k \cdot \vec{g}_{k+1} = 0. \quad (5.6)$$

For every direction  $\vec{\nu}_k$  the direction of the previous step  $\vec{\nu}_{k-1}$  is taken into account:

$$\vec{\nu}_k = -\vec{g}_k + \gamma_k \cdot \vec{\nu}_{k-1}, \quad (5.7)$$

with  $\gamma$  being a scalar constant given by

$$\gamma_k = \frac{\vec{g}_k \cdot \vec{g}_k}{\vec{g}_{k-1} \cdot \vec{g}_{k-1}}. \quad (5.8)$$

Thus, new directions are linear combinations of the current gradient  $\vec{g}_k$  and the previous direction  $\vec{\nu}_{k-1}$ . As there is no previous direction for the first step, the conjugate gradient methods starting direction is the same as for SD.

### 5.2.2 A Second-Order Minimization Method - The Newton-Raphson Method (NR)

For this method lets have a look at the Taylor expansion in the 1-dimensional case again:

$$V(x) = V(x_k) + (x - x_k) \cdot V'(x_k) + \frac{1}{2}[(x - x_k)^2 \cdot V''(x_k)] + \dots \quad (5.9)$$

The first derivative of this function is

$$V'(x) = V'(x_k) + (x - x_k) \cdot V''(x_k). \quad (5.10)$$

At the minimum  $x = x_{Min}$  the first derivative is Zero ( $V'(x_{Min}) = 0$ ) and so

$$x_{Min} = x_k - \frac{V'(x_k)}{V''(x_k)}. \quad (5.11)$$

The expression for a multidimensional function is

Minimization	1. Method	2. Method	Final Energy / kcal-mol <sup>-1</sup>
1.	600 Steps SD	-	-875
2.	600 Steps NR	-	-1142
3.	100 Steps NR	500 Steps SD	-1069
4.	100 Steps SD	500 Steps NR	-1189

Table 5.1: Results of different Energy Minimizations performed on BPTI.

$$\vec{x}_{Min} = \vec{x}_k - (V'')^{-1}(\vec{x}_k)V'(\vec{x}_k), \quad (5.12)$$

where  $(V'')^{-1}(\vec{x}_k)$  is the inverse of the Hessian matrix. Calculation of it can be computationally very expensive, especially for complex energy functions of large molecules [1].

Like the SD approach NR is an iterative method. For a quadratic function this method finds the minimum in one step from any starting point. In practice, however, a quadratic function is just an approximation for an energetic landscape and you have to perform more steps to get to a minimum. In this case  $\vec{x}_{Min}$  is used as the starting point ( $\vec{x}_k$ ) for the next iteration.

Besides the methods covered here, there are a lot of other procedures used to minimize the energy in molecular dynamics simulations. For example, there is a number of methods called quasi-Newton methods, which aim to reduce computation by eliminating the need to calculate the Hessian matrix.

Having discussed several minimization methods used in molecular modelling we now want to know which method should be applied. Since we start minimization from experimentally obtained structures with very high energy, SD or the conjugate method will be first applied as they reach lower energies very fast due to their low demands on computation. However, when reaching regions of lower energy, where slopes are less steep, other, more sophisticated methods such as Newton-Raphson can be applied to speed up the search for low energy points. To examine the performance of the two most frequently used methods, SD and NR, four different minimizations were performed on BPTI the results confirmed the assumptions stated above.

Minimization 4, which had the best performance, is shown in Fig.5.4 where the energy is plotted against the number of time steps. When the minimization method switches to NR, the number of time steps is set to Zero again.

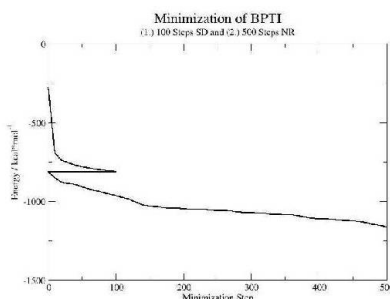


Figure 5.5: Minimization No. 4

## Chapter 6

# Molecular Dynamics (MD) and Normal Mode Analysis (NMA)

### 6.1 Molecular Dynamics - Running A Molecular Dynamics Simulation

#### 6.1.1 Starting Structure

Before starting a molecular dynamics simulation one needs an initial set of coordinates meaning a configuration of the molecule being simulated. In general structures retrieved from X-ray scattering or NMR experiments are used [23]. These can be obtained from the Brookhaven Protein Data Bank (<http://www.rcsb.org/pdb/>). The starting structure is of extreme importance. As molecular dynamics simulations are computationally very expensive, it is only possible to calculate a few nanoseconds within an appropriate period of time. Dynamic molecular processes like protein folding, however, take much longer; protein folding for example can require a few milliseconds up to 10 seconds in vivo. Thus, the starting structure should be close to the state you want to simulate.

#### 6.1.2 Modification of the Starting Structure

Structures retrieved from a protein databank lack hydrogen atoms, because they cannot be resolved properly. So these have to be added to the file. In addition there are water molecules immanent in the structure. A decision has to be made whether to include water molecules in the simulation or not. Then you can either add additional water molecules or remove them. This depends on what kind of data you want to produce. People who want to simulate crystal states will exclude water. Others, examining ligand binding for example, will include water, as they simulate a process that occurs in a biological environment.

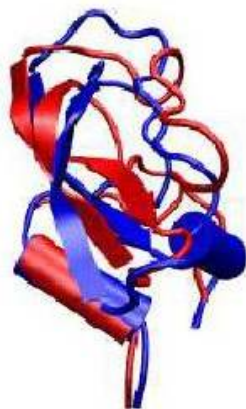


Figure 6.1: Aligned BPTI Structures shown in Cartoon Representation (Blue: Before Minimization; Red: After Minimization)

### 6.1.3 Energy Minimization

As mentioned before NMR structures and X-ray crystal structures tend to have high energy interactions like Pauli repulsions. That is due to the fact that the methods to retrieve molecular structures are not perfect and especially in x-ray-structures there are crystal contacts, which lead to a compaction of the molecules. Furthermore, hydrogen atoms are added to relatively arbitrary positions near their neighbors. Thus there are atoms lying too close together so that the Pauli repulsion outweighs the dispersion attraction and the energy is raised high above natural energy levels. These high energy interactions lead to local distortions which result in an unstable simulation. They can be released by minimizing the energy of the structure before starting a run.

The minimization results in a structure with an energy near the lowest possible energy the system can have. Thus, this state corresponds to a temperature near 0 K, where no motion can be seen [2]. You can easily imagine that there is no motion meaning no forces on the atoms in the system of a minimized structure: In an energy minimum the gradient of the potential energy equals a zero vector,  $\frac{\partial V(\vec{x})}{\partial \vec{x}}$ . This means the derivative of  $V(\vec{x})$  with respect to any coordinate of any atom equals zero. And so, with the force being the negative gradient of  $V(\vec{x})$ , there is no net force acting on any atom in any direction and therefore no motion will be seen inside the system. As you can see in Fig.6.1, the structures before and after minimization can vary extremely.

### 6.1.4 Heating Dynamics

To raise the temperature from 0 K to the desired value, you first have to assign initial velocities to the atoms corresponding to a Gaussian distribution for a certain temperature to provide the system with energy [1].

Then for the first time in the course of the MD procedure Newton's equations

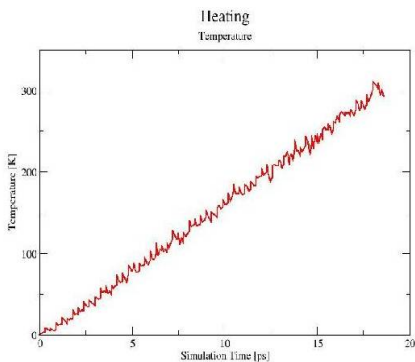


Figure 6.2: Temperature vs. Time.

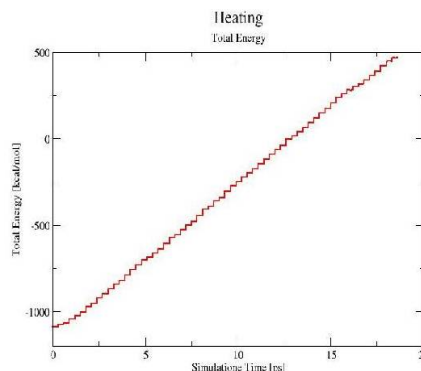


Figure 6.3: Total Energy vs. Time.

of motion are integrated to propagate the system in time. This is done for a certain period of time to let the system equilibrate in the new thermodynamic state, giving the energy time to evenly distribute throughout the system. In the next step the velocities are scaled to values corresponding to a slightly higher temperature and another equilibration phase is carried out. You can reach the desired temperature by simply repeating this process.

Typical steps for raising the temperature are about 5 K and the short equilibration period lasts about 0.3-1.0s depending on the size of the simulated system. So a heating process with 5 K every 0.3 ps, for example, would raise the temperature from 0 K to 300 K in 20 ps (as shown for BPTI in Fig6.2), which is in fact very quick. An even more rapid heating, though, would result in high-energy motions and interactions that are physically not feasible. Energy density for example would be increased locally to a level, when bonds break or, if quantum mechanics were neglected, they would elongate far beyond natural dissociation lengths. So every molecule with a complex three dimensional structure made up by hydrogen bonds or other non-bonded interactions would denature and thus be useless for any further simulation.

In Figures 6.2-6.5 it is shown how some properties of the system vary in time during a heating process. As you can see the kinetic energy and the temperature show exactly the same behavior. This is because they are related to each other by following equation:

$$E_{kin} = \frac{1}{2}m\langle\vec{v}^2\rangle = \frac{3}{2}NkZ, \quad (6.1)$$

where  $N$  is the number of atoms and  $k$  the Boltzmann constant.

Having raised the total energy of the system in a heating step, this amount of energy is evenly distributed throughout the system as already stated before. On the one hand this distribution occurs in space, meaning very high and very low velocities of single atoms get closer to the mean value. On the other hand the distribution occurs in terms of different energies: kinetic and potential energy. As you can see the total energy is the only property that is constant during equilibration. Kinetic and potential energy, however, behave irregularly, showing that the total energy distributes between them in a vibrational kind of way.

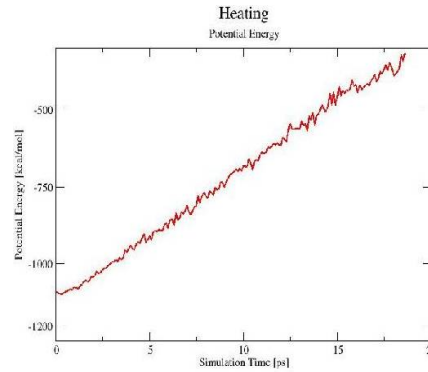
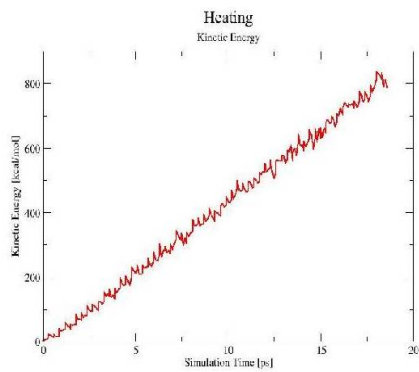


Figure 6.4: Kinetic Energy vs. Time      Figure 6.5: Potential Energy vs. Time

### 6.1.5 Equilibration and Rescaling Velocities

After having heated the system so quickly, the structure might be unstable and the temperature may drop too low [3]. That is why you have to equilibrate your system properly before running the real dynamics simulation, the production run. As stated before, equilibration is the process where the kinetic energy and the potential energy evenly distribute themselves throughout the system.

This is done by simulating the system whilst monitoring important properties such as temperature, the different energy terms, structure etc.. At constant periods of time the velocities are rescaled to the values for the desired temperature. This can be seen as vertical lines in Fig.6.6.

This is done until the simulation becomes stable with respect to time, which means till thermodynamic terms like temperature and energy are retained in a certain, small interval for a sufficiently long time. Only when the average temperature of the system stabilizes one can collect the trajectory information for analysis.

You may wonder now why the energy or temperature drops at all. In natural systems the energy is conserved, of course. In a simulation, however, energy

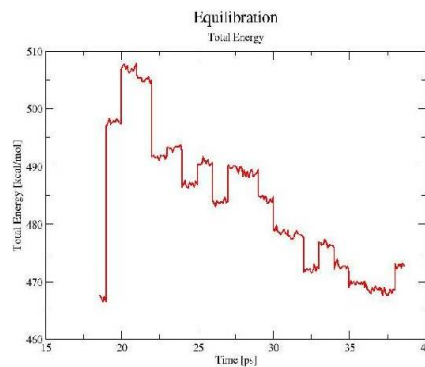


Figure 6.6: Total Energy vs. Time during Equilibration

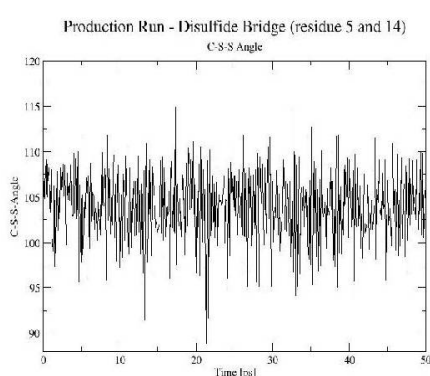


Figure 6.7: C-S-S Angle of Disulfide Bridge between CYS 5 and CYS 14

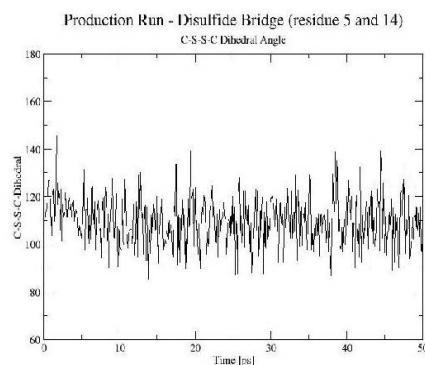


Figure 6.8: C-S-S-C Dihedral Angle of Disulfide Bridge between CYS 5 and CYS 14

conservation may be violated due to several reasons: The force field may neglect some critical effects (for example due to a badly selected cutoff) or the numerical integration method may be not precise enough due to numerical limitations of the computer resulting from the binary coding of numbers.

### 6.1.6 Production Dynamics

After all these preparations, which can take up to a couple of weeks, the actual molecular dynamics simulation can be started by integrating Newton's equations of motion of the system for the desired period of time. This can be from several hundred picoseconds up to some nanoseconds.

All the coordinates, velocities, accelerations and momenta generated during the production run are saved and used for analysis. Thus, for example, average structures can be calculated and compared to experimental structures and even more important time dependent properties such as different energy terms, angles and dihedral angles or distances between atoms or whole selections of atoms can be displayed and interpreted.

For example the time dependent behavior of the C-S-S angles and C-S-S-C dihedral angles in a disulfide bridge of BPTI during a 50 ps simulation is shown in Figures 6.7 and 6.8.

As disulfide bridges are tertiary structure elements and thus play a role in keeping the three-dimensional structure alive, they should be conserved during simulation. A flip between totally different levels of angles would indicate a conformational change of the disulfide bridge, which was obviously not the case during our simulation.

Another important time dependent property of dynamic systems is the *root mean square deviation (RMSD)*. *RMSD* indicates how much two structures vary in terms of differences between the coordinates of the structures and is calculated with

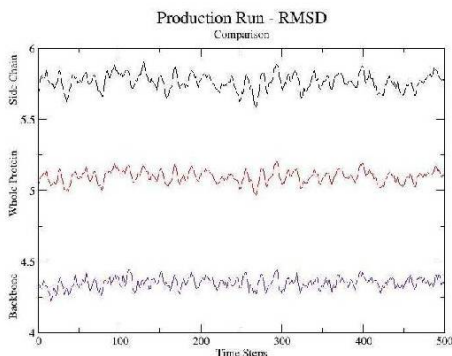


Figure 6.9: RMSD during Production Run for different Parts of the Protein

$$D_{RMS} = \langle (\vec{x}_i^\alpha - \vec{x}_i^\beta)^2 \rangle^{\frac{1}{2}} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\vec{x}_i^\alpha - \vec{x}_i^\beta)^2} \quad (6.2)$$

where  $\vec{x}_i$  is the coordinate of atom  $i$  and  $\alpha$  and  $\beta$  correspond to the different structures. Calculation of  $RMSD$  for a time series of coordinate sets compared to a reference structure yields graphs like shown in Fig.6.9, where  $RMSD$  values of BPTI are displayed for (i) the whole protein, (ii) the side chains and (iii) the backbone of the protein compared to BPTI crystal structure.

As you can see, the side chains  $RMSD$  is higher compared to the backbone, which means that they are more flexible, whereas the backbone seems rather rigid. The whole protein  $RMSD$  obviously has values between these two, because for this calculation both atom selections (side chain and backbone) are taken into account at once.

A special case of  $RMSD$  is the *root mean square fluctuation (RMSF)* where the reference structure is an average structure over the whole trajectory:

$$F_{RMS} = \langle (\vec{x}_i^\alpha - \vec{x}_i^{Average})^2 \rangle^{\frac{1}{2}} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\vec{x}_i^\alpha - \vec{x}_i^{Average})^2}. \quad (6.3)$$

$RMSF$  can be related to experimental data by comparing with so-called B factors,  $B_i$ , which are measured in X-ray experiments:

$$B_i = \frac{8}{3} \pi^2 (F_{RMS,i})^2 \quad (6.4)$$

$RMSF$  calculation can also give better information on how structures fluctuate during a simulation, while  $RMSD$  is more appropriate to show that simulations are performed close to experimental structures to convince scientists with a rather critical view on reliability of molecular dynamics simulations.

Thus, combined with molecular graphics programs which can display molecular structures in a time dependent way, molecular dynamics simulations provide a powerful tool to visualize and understand conformational changes involved in

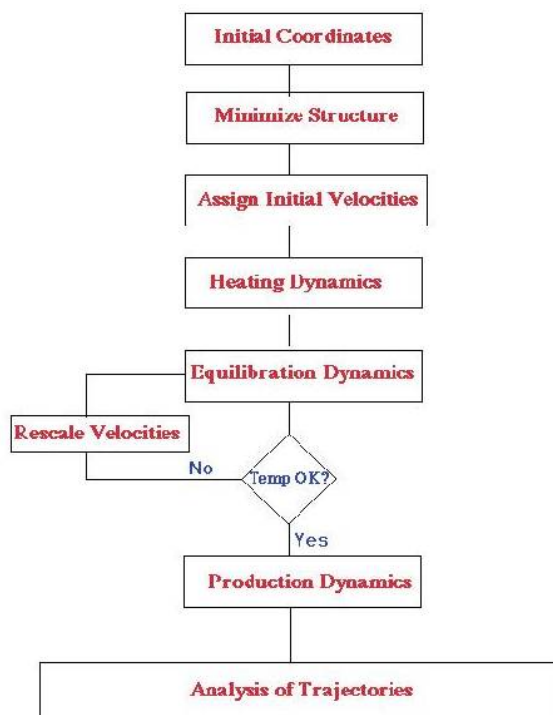


Figure 6.10: Summary of the proceeding of a Molecular Dynamics simulation

ligand binding, catalysis or other functions of biological macromolecules at an atomic level.

To give an overview of the whole procedure described in this chapter, it is summarized in Fig.6.10.

## 6.2 Normal Mode Analysis

The Normal Mode Analysis (NMA) calculates features of the normal vibration modes of molecules. Obtaining them, scientists try to deduce certain properties of molecules (e.g. protein function) just by observing the overall vibrational motion of the examined molecule or by comparing the theoretical to spectroscopic results. Furthermore, comparisons to experiments can be used in the parametrisation of a force field. NMA is performed at a hypothetical, motionless state at 0 K. However, experimental measurements are made on molecules at a finite temperature when the molecules undergo all kinds of motion. To compare theoretical and experimental results it is necessary to make appropriate corrections to allow for these motions. These corrections are calculated using standard statistical mechanics formulae.

The vibrational contribution ( $U_{vib}$ ) to the internal energy at a temperature  $T$  requires knowledge of the actual vibrational frequencies ( $n_i$ ). Constituent  $U_{vib}$

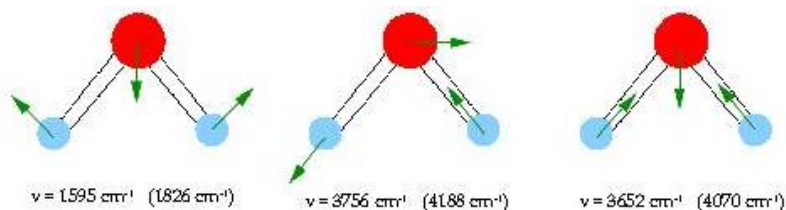


Figure 6.11: As a molecule consisting of three atoms, water has three normal modes which are presented here. Experimental (and calculated) frequencies are shown. The first one has a significantly lower frequency  $\nu$  than the others. Quantity  $\nu$  is proportional to  $\sqrt{\frac{k}{\mu}}$  with force constant  $k$  and reduced mass  $\mu$

equals the difference in the vibrational enthalpy at the temperature  $T$  and at 0 K and is given by:

$$U_{vib}(T) = \sum_{i=1} N_{nm} \left( \frac{h\nu_i}{\exp(h\nu_i/k_bT) + 1} \right) \quad (6.5)$$

$N_{nm}$  is the number of normal vibrational modes for the system. Even the zero-point energy  $U_{vib}(0)$ , obtained by summing  $\frac{1}{2}h\nu_i$  for each normal mode, can be quite substantial, amounting to about 100 kcal/mol for a  $C_6$  molecule. Other thermodynamic quantities such as entropies and free energies may also be calculated from the vibrational frequencies using the relevant statistical mechanics expressions.

Normal modes of vibrations are oscillations about an energy minimum, characteristic of a system's structure and its energy function. For a purely harmonic energy function, any motion can be exactly expressed as a superposition of normal modes (anharmonic potentials can be approximated by harmonic potentials at sufficiently low temperatures of the system). Thus, normal modes are useful because they describe collective motions of the atoms in a coupled system that can be individually excited.

The three normal modes of water are schematically illustrated in Fig.6.11; a nonlinear molecule with  $N$  atoms has  $3N-6$  normal modes [24].

A normal mode calculation is based on the assumption that the energy surface is quadratic in the vicinity of an energy minimum (Fig.6.12).

This means that each normal mode acts as a simple harmonic oscillator with

$$m \frac{d^2x}{dt^2} = F = -kx. \quad (6.6)$$

Since series expansions are useful for approximating functions (4) we can rewrite the energy function in one dimension like this:

$$V(x) = V(0) + V'(0) \cdot x + \frac{1}{2}V''(0) \cdot x^2 + \dots \quad (6.7)$$

Making the function vary in a quadratic fashion the series is truncated after the second derivative. If zero corresponds to a minimum of the energy landscape it is

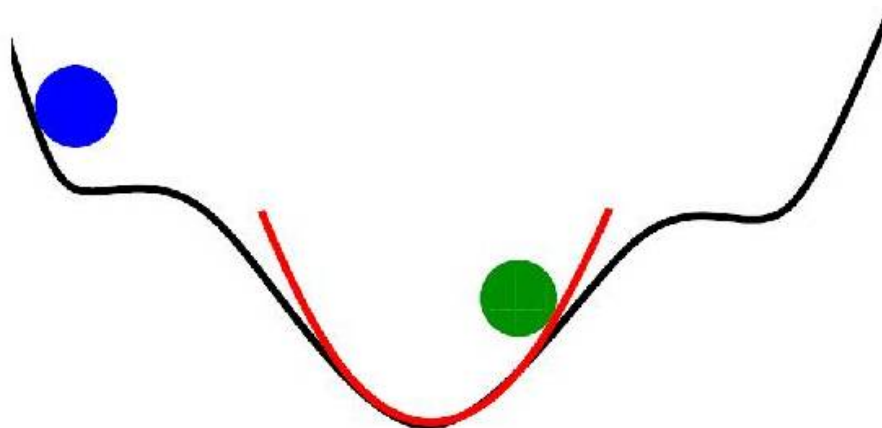


Figure 6.12: The motion of a molecule around an energy minimum can be approximately described by a parabolic energy profile. This is the reason why one has to generate the energy-minimized structure (green ball), which is located around a minimum of the energy surface, before starting a normal mode calculation

$$V(x) = V(0) + \frac{1}{2}V''(0) \cdot x^2, \quad (6.8)$$

and with

$$m \frac{d^2x}{dt^2} = F = -V'(x) \quad (6.9)$$

we obtain

$$F = -V''(0) \cdot x. \quad (6.10)$$

The solution of the second-order differential equation

$$\frac{d^2x}{dt^2} = \frac{V''(0)}{m}x \quad (6.11)$$

is  $x = A \cdot \exp(i\omega t)$  with the angular velocity  $\omega = \frac{k}{m}$  and the amplitude  $A$ . Fig.6.11 actually illustrates the vibrational motion of a particle in one dimension (red path). Fig.6.13 shows this kind of motion in two dimensions. To picture it in three dimensions is more difficult.

In three dimensions the frequencies of the normal modes together with the displacements of the individual atoms may be calculated from a molecular mechanics force field using the Hessian matrix of second derivatives ( $\mathbf{V}$ ). The Hessian must first be converted to the equivalent force-constant matrix in mass-weighted coordinates ( $\mathbf{F}$ ), as follows:

$$\mathbf{F} = \mathbf{M}^{\frac{1}{2}} \mathbf{V}'' \mathbf{M}^{\frac{1}{2}}. \quad (6.12)$$

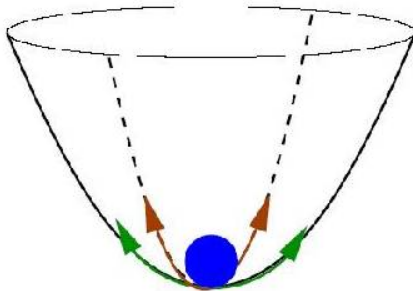


Figure 6.13: Vibrations in 2-dimensional space. In reality one more dimension comes into play

$\mathbf{M}$  is a diagonal matrix of dimension  $3N \times 3N$ , containing the atomic masses. All elements of  $\mathbf{M}$  are zero except those on the diagonal;  $M_{1,1} = M_{2,2} = M_{3,3} = m_1$ ,  $M_{4,4} = M_{5,5} = M_{6,6} = m_2$ , ...,  $M_{3N-2,3N-2} = M_{3N-1,3N-1} = M_{3N,3N} = m_N$ . Each non-zero element of  $\mathbf{M}^{\frac{1}{2}}$  is thus the inverse square root of the mass of the appropriate atom. The masses of the atoms must be taken into account because a force of a given magnitude will have a different effect upon a larger mass than a smaller one. For example, the force constant for a bond to a helium atom is, to a good approximation, the same as to a hydrogen, yet the different mass of the helium gives a different motion and a different zero-point energy. The use of mass-weighted coordinates takes care of these problems. Strictly mathematically, the above term for the force-constant matrix comes from a coordinate transformation where the valid term

$$\mathbf{M}\ddot{x} = -\mathbf{V}''x \quad (6.13)$$

has been matrix multiplied with the square matrix  $\mathbf{M}^{-\frac{1}{2}}$ , thus leading

$$\mathbf{M}^{\frac{1}{2}}\ddot{x} = \mathbf{M}^{\frac{1}{2}}\mathbf{V}''\mathbf{M}^{-\frac{1}{2}}\mathbf{M}^{\frac{1}{2}}x. \quad (6.14)$$

We next have to obtain the eigenvalues and eigenvectors of the matrix  $\mathbf{F}$  which then have to be retransformed to the original coordinate system in order to correspond to the desired modes of motion. Calculating eigenvalues and eigenvectors is usually performed using matrix diagonalization [25]. If the Hessian is defined in terms of Cartesian coordinates, then six of these eigenvalues will be zero as they correspond to translational and rotational motion of the entire system. The frequency of each normal mode is then calculated from the eigenvalues  $\lambda_i$  using the relationship

$$\nu_i = \frac{\sqrt{\lambda_i}}{2\pi}, \quad (6.15)$$

as  $\sqrt{\lambda}$  corresponds to  $\frac{k}{m}$  and since the solution of the second-order differential equation is an oscillating function, the frequency  $\nu$  equals  $\frac{\omega}{2\pi}$ . The ratio of the amplitudes ( $x_i$ ) of each normal mode are given by the eigenvectors  $\eta_i$  which themselves characterize the mode of motion.

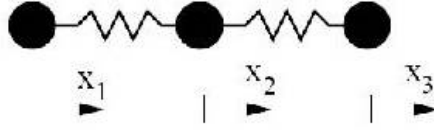


Figure 6.14: A linear triatomic molecule like  $\text{CO}_2$ . The vectors (here: scalars)  $x_1, x_2, x_3$  define displacements of the corresponding atoms.

As a simple example of a normal mode calculation consider the linear triatomic system in Fig.6.14. We shall just consider motion along the long axis of the molecule and that all three atoms have the same masses ( $m_1 = m_2 = m_3$ ).

The displacements of the atoms from their equilibrium positions along this axis are denoted by  $x_i$ . It is assumed that these are small compared to the values of the equilibrium bond lengths and the system is harmonic with bond force constants  $k$ . The potential energy is given by:

$$V = \frac{1}{2}k \cdot (x_1 - x_2)^2 + \frac{1}{2} \cdot (x_2 - x_3)^2. \quad (6.16)$$

We next calculate the first derivatives of the potential energy with respect to the three coordinates  $x_1, x_2, x_3$ :

$$m\ddot{x}_1 = -\frac{\partial V}{\partial x_1} = -k \cdot (x_1 - x_2); \quad (6.17)$$

$$m\ddot{x}_2 = -\frac{\partial V}{\partial x_2} = -k \cdot (x_2 - x_1) - k \cdot (x_2 - x_3); \quad (6.18)$$

$$m\ddot{x}_3 = -\frac{\partial V}{\partial x_3} = -k \cdot (x_3 - x_2); \quad (6.19)$$

And the Hessian

$$\mathbf{V}'' = \begin{pmatrix} k & -k & 0 \\ -k & 2k & -k \\ 0 & -k & k \end{pmatrix}, \quad (6.20)$$

the mass-weighted matrix

$$\mathbf{M} = \begin{pmatrix} m & 0 & 0 \\ 0 & m & 0 \\ 0 & 0 & m \end{pmatrix}, \quad (6.21)$$

and the force-constant matrix

$$\mathbf{F} = \begin{pmatrix} \frac{k}{m} & \frac{-k}{m} & 0 \\ \frac{-k}{m} & \frac{2k}{m} & \frac{-k}{m} \\ 0 & \frac{-k}{m} & \frac{k}{m} \end{pmatrix}. \quad (6.22)$$

The eigenvalues and eigenvectors of  $\mathbf{F}$  dont need to be retransformed in this example because of the shape of  $\mathbf{M}$  (all atoms have the same masses). The eigenvalues, each corresponding to a different mode of motion, are:

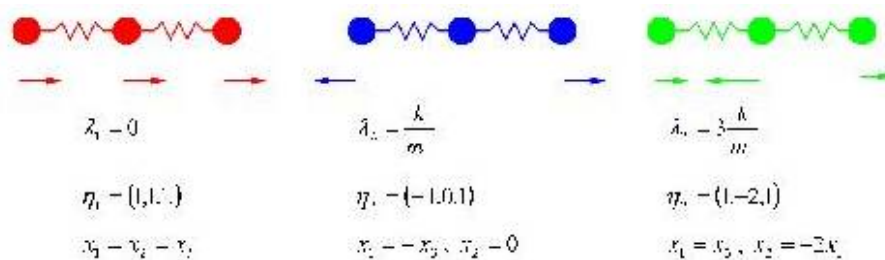


Figure 6.15: Results of normal mode calculation for a linear triatomic molecule.  $\lambda_i$ ,  $\eta_i$  and  $x_i$  describe eigenvalues, eigenvectors and amplitudes.

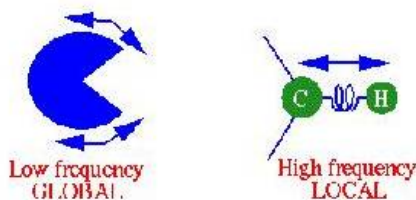


Figure 6.16: Protein molecules are the most examined type of molecule with respect to vibrational motion. Obtaining the normal modes of motion, one can notice a difference between the motional frequency of bigger (global) parts and smaller (local) parts of the molecule, e.g. a whole domain or even just an atomic link between two distinct atoms. Global motions of a protein are often specific to it and can be related to its function.

$$\lambda_1 = 0, \lambda_2 = \frac{k}{m} \text{ and } \lambda_3 = 3\frac{k}{m} \quad (6.23)$$

Now each of the three frequencies can be obtained as shown above. They correspond to modes of a translation, a symmetric stretch and an asymmetric stretch respectively (Fig.6.2). In this triatomic example the three-dimensional eigenvector of each mode determines the displacement of each atom.

The motions of larger segments of an analyzed protein, for example, are adequate for studying the molecules function. These motions are represented by lowfrequency modes. High frequency modes correspond to motions of smaller molecule parts (Fig.6.16).

The harmonic approximation to the energy surface is found to be appropriate for welldefined energy minima such as the intramolecular degrees of freedom of small molecules. For larger systems the harmonic approximation breaks down. Such systems also have an extraordinarily large number of minima on the energy surface. In these cases it is not possible to calculate accurately thermodynamic properties using normal mode analysis. Rather, molecular dynamics simulations or other methods must be used to sample the energy surface from which properties can be derived.

## Chapter 7

# What Else is Possible

Other analyses exploring the energy surface focus on determining reaction pathways or transition structures of molecules. Since the minimum points of the energy landscape may correspond to the reactants or products of a chemical reaction or two important conformations of a molecule (Fig.7.1), the path between those two minima (the '*reaction path*' or '*pathway*') might be of interest. The transition structure is the point of highest potential energy along the reaction pathway.

As one can imagine many methods have been worked out for elucidating re-

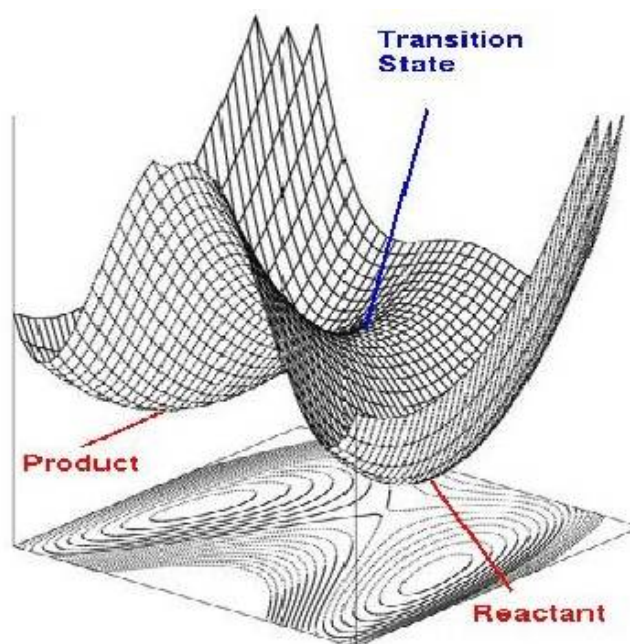


Figure 7.1: A saddle point of the multidimensional reaction path corresponds to the transition state between reactants and products.

action pathways and finding transition structures. These structures correspond to saddle points with one negative eigenvalue of the Hessian matrix, where the energy passes through a maximum for movement along the reaction path. Computational methods for locating transition structures do so by searching along the reaction pathway, e.g. by using minimization algorithms when provided with an initial structure close to the wanted one. Conversely, methods for finding the reaction pathway start from the transition structure and move downwards using minimization. Yet other methods determine both simultaneously from the two minima bordering the reaction path. In particular, the conjugate peak refinement [26] is a good method for locating transition structures for systems with many atoms where a number of such states between two conformations may be.

Besides the molecular dynamics simulation, there is another widely spread computer simulation technique we would like to mention here: The Monte Carlo (MC) method which differs in some ways from the MD method. The Monte Carlo simulation method also provides a picture of the system in different conformations. However, it does not show how the system switches between these since the behavior of atomic and molecular systems cannot be determined with respect to processing time. A Monte Carlo simulation generates configurations of a system by randomly changing the positions of the atoms present and so the outcome of each calculation depends only on its preceding one. Furthermore, the total energy is determined directly from the potential energy function. These two points are in contrast to a molecular dynamic simulation where Newtons equations of motion are the basis. Nonetheless, thermodynamic quantities can be derived using appropriate statistical mechanics formulae.

## Appendix A

# Acknowledgements

We would like to thank the group of ‘Biocomputing’ at the ‘Interdisciplinary Center for Scientific Computing’, Heidelberg, and especially Lars Meinhold and Dr. Vandana Kurkal for organizing the F-Practical (‘Vacuum Simulations of BPTI’) and supporting our work.

Moreover, we would like to express our gratitude towards Dr. Moritz Diehl, who supervised us during the seminar ‘Mathematical Methods in Bioinformatics’ which provided the basis for this report.

# Bibliography

- [1] Brooks B.R., Bruccoleri R.E., CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations, *Journal of Computational Chemistry*, **1983**: 4, 187
- [2] Leach A.R., Molecular Modelling, *Pearson Prentice Hall*, Harlow, GB, **2001**
- [3] [http://www.ch.embnet.org/MD\\_tutorial/index.html](http://www.ch.embnet.org/MD_tutorial/index.html)
- [4] Diehl M., Fischer T., Skript zur Mathematik fuer die Molekulare Biotechnologie, *Universitaet Heidelberg, D*, **2003**, Chapter 3.8
- [5] McQuaerrie D.A., Statistical Mechanics, *University Science Books, Sansalito, USA*, **2000**
- [6] Born M., Oppenheimer R., Zur Quantentheorie der Molekeln, *Annalen der Physik*, **1927**, 84: 457
- [7] Allinger N.L., MM2: A Hydrocarbon Force Field Utilizing V1 and V2 Torsional Terms, *Journal of the American Chemical Society*, **1977**, 99: 8127
- [8] Allinger N.L., Li F., Yan L., The MM3 Force Field for Alkenes, *Journal of Computational Chemistry*, **1990**, 11: 848
- [9] Allinger N.L., Li F., Yan L., Molecular Mechanics (MM3) Calculations on Conjugated Hydrocarbons, *Journal of Computational Chemistry*, **1990**, 11: 868
- [10] Allinger N.L., Yuh Y.H., Lii J.-J., The MM3 Force Field for Hydrocarbons I, *Journal of the American Chemical Society*, **1989**, 111: 8551
- [11] Allinger N.L., Chen K., Lii J.-H., An Improved Force Field (MM4) for Saturated Hydrocarbons, *Journal of Computational Chemistry*, **1996**, 17: 642
- [12] Allinger N.L., Chen K., Katzenelenbogen J.A., Wilson S.R., Anstead G.M., Hyperconjugative Effects on Carbon-Carbon Bond Lengths in Molecular Mechanics (MM4), *Journal of Computational Chemistry*, **1996**, 17: 747

- [13] Nevins N., Chen K., Allinger N.L., Molecular Mechanics (MM4) Calculations on Alkenes, *Journal of Computational Chemistry*, **1996**, 17: 669
- [14] Nevins N., Chen K., Allinger N.L., Molecular Mechanics (MM4) Calculations on Conjugated Hydrocarbons, *Journal of Computational Chemistry*, **1996**, 17: 695
- [15] Nevins N., Chen K., Allinger N.L., Molecular Mechanics (MM4) Vibrational Frequency Calculations for Alkenes and Conjugated Hydrocarbons, *Journal of Computational Chemistry*, **1996**, 17: 730
- [16] Vollhardt K.P.C., Schore N.E., Organic Chemistry, *Freeman*, **1999**, Chap: 2.5-2.7
- [17] Cox S.R., Williams D.E., Representation of Molecular Electrostatic Potential by a New Atomic Charge Model, *Journal of Computational Chemistry*, **1981**, 2: 304
- [18] Fowler P.W., Buckingham A.D., Central or Distributed Multipole Moments? Electrostatic Models of Aromatic Dimers, *Chemical Physics Letters*, **1991**, 176: 11
- [19] Halgren T.A., Representation of van der Waals (vdW) Interactions in Molecular Mechanics Force Fields: Potential Form, Combination Rules, and vdW Parameters, *Journal of the American Chemical Society*, **1992**, 114: 7827
- [20] Stote R.H., States D.J., On the Treatment of Electrostatic Interactions in Biomolecular Simulation, *Journal de Chimie Physique*, **1991**, 88: 2419
- [21] Greengard L., Rokhlin V., A Fast Algorithm For Particle Simulations, *Journal of Computational Chemistry*, **1987**, 73: 325
- [22] Krabs W., Modern Methods of Optimization, *Springer, Heidelberg/Berlin, D*, **1990**
- [23] Bhm H.-J., Klebe G., Kubinyi H., Wirkstoffdesign, *Spektrum Heidelberg, D*, **1996**
- [24] Atkins P.W., Physikalische Chemie, *Wiley-VCH, Berlin, D*, **2001**, Chapter 16.4
- [25] Diehl M., Fischer T., Skript zur Mathematik fuer die Molekulare Biotechnologie, *Universitaet Heidelberg, D*, **2003**, Chapter 5.2, 5.3
- [26] Fischer S., Karplus M., Conjugate Peak Refinement: An Algorithm for Finding Reaction Paths and Accurate Transition States in Systems with Many Degrees of Freedom, *Chemical Physics Letters*, **1992**, 194: 252