UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL ATOMIC ENERGY AGENCY
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

H4.SMR/916 - 38

SEVENTH COLLEGE ON BIOPHYSICS:

Structure and Function of Biopolymers: Experimental and Theoretical Techniques.

4 - 29 March 1996

Molecular Dynamics Simulation of Proteins

Massimo Marchi

Section de Biophysique des Proteines et des Membranes
Centre d'Etudes de Saclay
Gif-sur-Yvette
France

CHAPTER 27

Molecular dynamics simulation of proteins

Massimo Marchi

Section de Biophysique des Protéines et des Membranes, DBCM, DSV, CEA
Centre d'Etudes de Saclay, 91191 Gif-sur-Yvette Cedex, France
E-mail: marchi@job.saclay.cea.fr

Conference Proceedings Vol.49
"Monte Carlo and Molecular Dynamics of Condensed Matter Systems"
K. Binder and G. Ciccotti (Eds.)
SIF, Bologna, 1996

Contents

1. Introduction

The complexity of the systems studied by molecular dynamics (MD) simulations has increased steadily since the early 1960's. From the first pioneering simulations of a few hundreds of Lennard-Jones particles carried out by Rahman[1] and Verlet[2] on large mainframe computers, computer technology has evolved considerably to allow today routine large scale simulations of ten thousand particle systems on personal workstations. Following the early works on atomic liquids, the MD technique has been extended to the study of molecular liquids and molecular solids. In 1977 McCammon and P. Wolynes published the first paper[3] reporting on a short (about 9 ps) MD simulation of bovine pancreatic trypsin inhibitor (BPTI) which is one of the smallest proteins whose high resolution X-ray structure was determined in 1975 by Deisenhofer and Steigemann.[4] Since then a great variety of application to different proteins and other biological systems (DNA and membranes) have been presented in literature. As for non-biological systems. MD simulation has become a tool to interpret and analyze experimental data which do not provide direct information at the microscopic level. Unfortunately, the range of biological phenomena that can be tackled by MD simulation is limited. Today the time-scales accessible by simulation do not go beyond the tens or maybe hundreds of picoseconds while important phenomena such as folding occurs with time constants of microseconds, seconds and beyond.[5] Nevertheless, the use of computer modeling is nowadays well established in some areas of biochemistry and simulations are used routinely in protein structure determination by X-ray and neutron crystallography, [6, 7] and 2-D NMR.[8, 9]

Before we go beyond this introduction, I must point out that indeed MD techniques for simulating proteins or any other biologically relevant system are identical to those used for simpler systems such as atomic or molecular liquids. After all, an attractive feature of computer simulation is that we can increase the size of the system without rewriting the whole code (this, of course, if the computer program has been written in a sensible way!). There are however at least two important aspects of protein simulation deeply entrenched with one another which make the difference. Firstly, the atoms of a protein are strongly bound with their nearest neighbors through covalent bonds. The combination of the complex bond topology with the interactions among the other non-bonded atoms introduces energy and geometric barriers between regions of configurational space. As a consequence of the former point proteins in their native states have a large variety of time scales from picosecond to minutes and days. Thus, not only we have to run much longer simulations than for liquids, but also we are *assured*[1] to sample only a very limited region of phase space around the local potential energy minimum. This is certainly the most important handicap of the MD technique. Solutions to the time-scale problem have not yet been found, even though progress has been made in this direction.[10, 11]

[1] We can run a small protein for at most 1-2 ns!

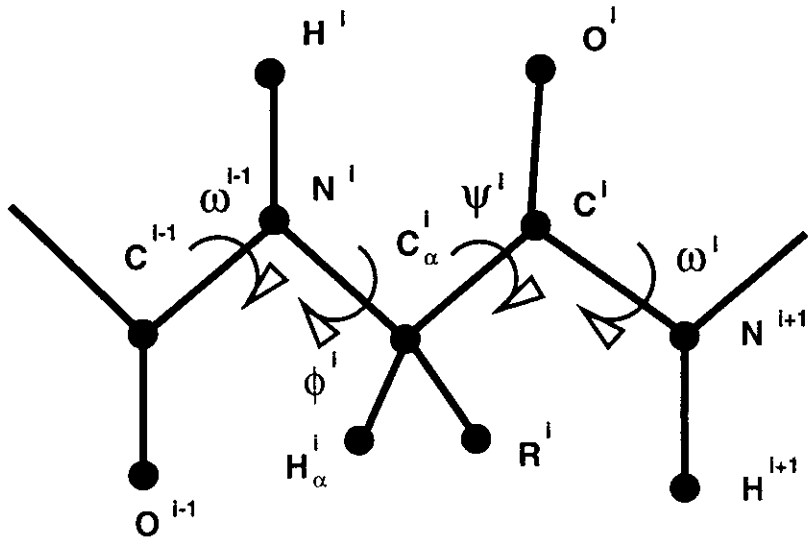


Figure 1. Structure of the Polypeptide Chain

In this lecture, I will not try to give a synoptic view of the entire field of protein simulation with relative applications. I will focus instead on the practical aspects which contrast MD simulation of liquids with MD of proteins. The lecture will cover the following points. What is a protein (Sec. 2)? Which are the most common interaction potentials (Sec. 3)? What are the practical problems encountered in setting up the covalent topology and in assigning the potential parameters to the atom of a protein (Sec. 4)? Which initial conditions and simulation algorithm should one use (Sec. 5)? How to handle long range electrostatic forces(Sec. 6). How to perform simulations in ensembles other than the microcanonical(Sec. 7). Finally, I will present an application of the MD simulation technique in the NPT ensemble to a crystal of the bovine pancreatic trypsin inhibitor (BPTI) (Sec. 8).

2. What is a Protein

Proteins are heteropolymers uniquely characterized by the sequence of their building units, the amino acid residues. This sequence is named the primary structure of the chain. A typical polypeptide is pictured in Fig. 1. All the amino acid residues share a backbone composed of the sequence $-NH-C_\alpha H-CO-$ and can be distinguished by the chemical structure of their side-chain, R. Dihedral angles[2] of the protein backbone associated with the C-N (i.e. the peptide bond), N-C_α and C_α-C are called respectively ω, ψ and ϕ. While the dihedral angles ω are rigid given the partial double bond character of the $C-N$ bonds, ϕ and ψ are variable along the chain. The possible range of values that such angles can assume is primarily limited by steric hindrance. However, this does not mean that all the permitted combinations of dihedral angles are found in proteins. Indeed non-bonded forces such as Coulombic and hydrogen bond interactions among the residues (backbone plus side-chains) determine the relative arrangement of the residues. In native structure of proteins certain regions

[2] See Sec. 4 for the definition of dihedral angle

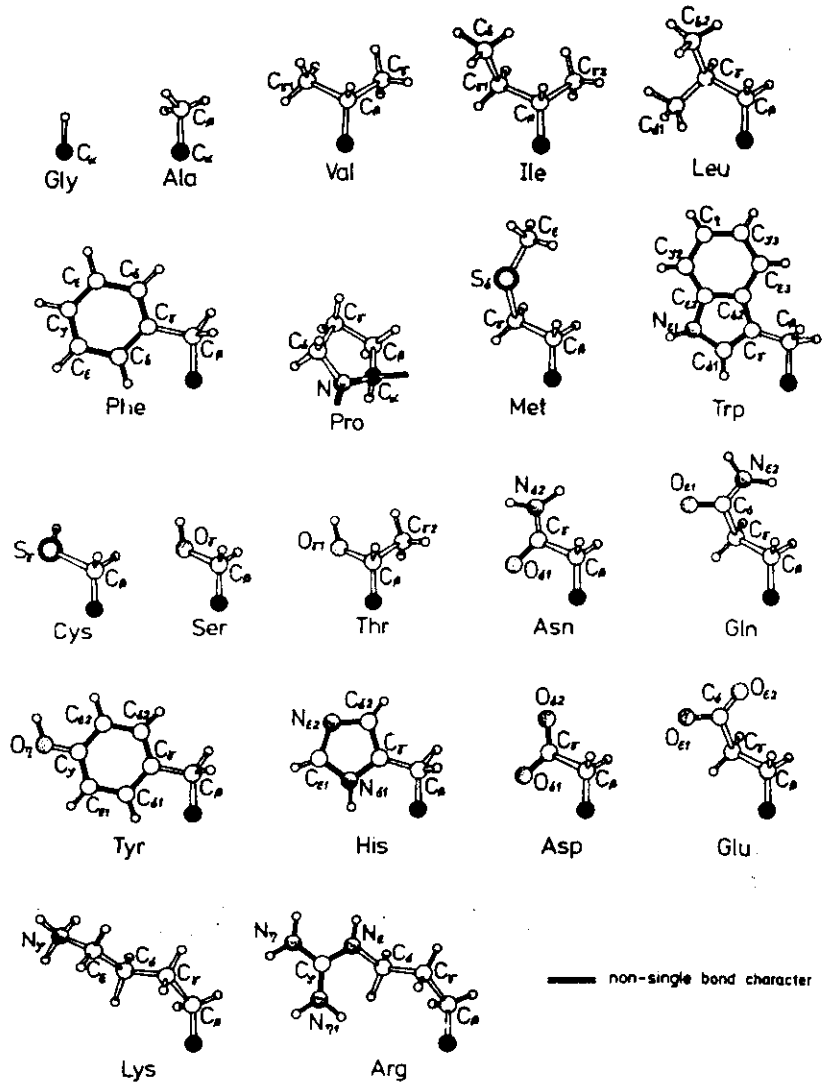


Figure 2. The 20 standard amino acid residues. Ala, Val, Leu, and Ile are aliphatic; Gly, Pro, Cys and Met are non-polar; His, Phe, Tyr and Trp are aromatic; Asn, Gln, Ser and Thr are polar; Lys and Arg are positively charged while Asp and Glu are negatively charged

or fragments of the polypeptide chain collapse together resulting in the formation of what are termed elements of secondary structure. A familiar example is the alpha helix, in which the peptide oxygen of the i-th residue is hydrogen bonded with the nitrogen and hydrogen of the $i+4$-th residue. Another commonly occurring secondary structure is the beta sheet. Here two extended segments of the polypeptide lie next to each other and are cross linked by hydrogen bonds between their peptides. In many cases a restricted range of ϕ and ψ angles are found for a given type of

secondary structure. There are 20 most commonly occurring amino acids. They can be distinguished in four classes: Aliphatic, aromatic, polar, non-polar and charged (see Fig. 2). In soluble globular proteins, aliphatic and aromatic residues are found in the hydrophobic interior of the protein. Charged residues are always found at the interface with the solvent (water), while polar neutral residues can be found at the surface as well as inside protein molecules. As internal residues they usually form hydrogen bonds with each other or with the polypeptide backbone.

3. Interaction Potentials

A general feature of almost all the force field developed for proteins is to divide the potential energy of the system into a few so-called bonded interaction terms and a non-bonded part. The latter includes electrostatic contributions along with dispersion and short-range repulsion terms. Of these model potentials only some are routinely used in MD simulation. The first force field developed for proteins employed variable torsions and rigid internal geometry to decrease considerably the total number of degrees of freedom.[12] This and its subsequent versions[13] are force fields developed explicitly for MC and static minimization. They cannot be efficiently used in MD simulations without performing the simulation in generalized coordinates which is currently unfeasible for complex systems. Other force fields have been developed expressly for the calculation of internal geometries and conformational energies and contain no or drastically damped electrostatic interactions (see for instance Ref. [14]). These are inappropriate in the study of condensed phase properties.

Here, I will discuss in detail only the class of potential functions developed expressly for simulations of condensed phase. In general, in addition to a non-bonded part composed of Coulombic term and of a Lennard-Jones function to represent respectively electrostatics and dispersion-repulsion terms, they contain a diagonal intramolecular part. A typical expression for the total potential energy of an isolated protein is the following:

$$
\begin{aligned}
E_{Pot} = & \frac{1}{2} \sum_{Bonds} K_r (r - r_0)^2 + \frac{1}{2} \sum_{Angles} K_\theta (\theta - \theta_0)^2 + \\
& + \frac{1}{2} \sum_{Dihed} V_\phi \left[1 + \cos \left(n\phi - \gamma \right) \right] + \\
& + \sum_{i<j}^{N} \left\{ 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{D r_{ij}} \right\}
\end{aligned}
\tag{1}
$$

The energy as given by the above equation is a function of the Cartesian coordinates, $\{ \mathbf{R}^N \}$, specifying the positions of all atoms involved. The actual calculation of this energy is performed by evaluating the internal coordinates for bonds (r), bond angles (θ), dihedral angles (ϕ), and inter-particle distance (r_{ij}) for any given geometry and using the respective energy parameters. K_r, K_θ and K_ϕ are the bonded force constants associated with bond stretching, bond and dihedral angles, while r_0, θ_0 and ϕ_0 are their respective geometrical reference values.

The non-bonded interaction parameters (the last term in Eq. 1) include the atomic diameters σ_{ij}, the minimum well depth ϵ_{ij}, the charges q_i and the dielectric constant D. In some case, to save computational time, not all the atoms which make up the protein are included in the calculation of the energy. In the earlier potential

fields[15, 16] all the hydrogens were incorporated with the bound heavy atom. Models that explicitly include only polar hydrogens are still used today.

The most popular force fields which use potential functions similar to Eq. 1 are the AMBER[17, 18], GROMOS[19], CHARMm[20], and OPLS/AMBER[21] force fields. Below, I shall examine in some details the various contributions to the total potential energy in Eq. 1, how they are typically determined and which technical problems might be involved in their evaluation during an MD simulation.

3.1. Bonded Interactions

Bonded interactions are considered as all interactions between two atoms separated by one, two or three consecutive bonds which correspond respectively in the potential function to the bond stretching, angle bending and dihedral angles terms. Fig. 3 gives a pictorial definition of the three types of bonded interactions. While only the two bound atoms contribute to the bond stretching energy, respectively three and four atoms are necessary to calculate bending and dihedral energies. In Eq. 1 the functional form of the bonded interactions is in its simplest form. It is harmonic and diagonal for bond stretchings and angle bendings, while it includes only the first two terms of a Fourier expansion for the periodic dihedral term. In more recent force fields cross terms and cubic terms are added to fit more adequately the vibrational modes of the system.[22]

Bond Stretching and Angle Bendings In the earlier force field as in more recent ones r_0, θ_0 are determined to reproduce bond length and bond angles of the X-ray structural data on the 20 most common amino acids.[15, 17, 18] On the other hand K_r and K_θ are picked to reproduce well known stretching and bending frequencies derived from vibrational analysis of smaller fragments. Sometimes *ab-initio* force constants for

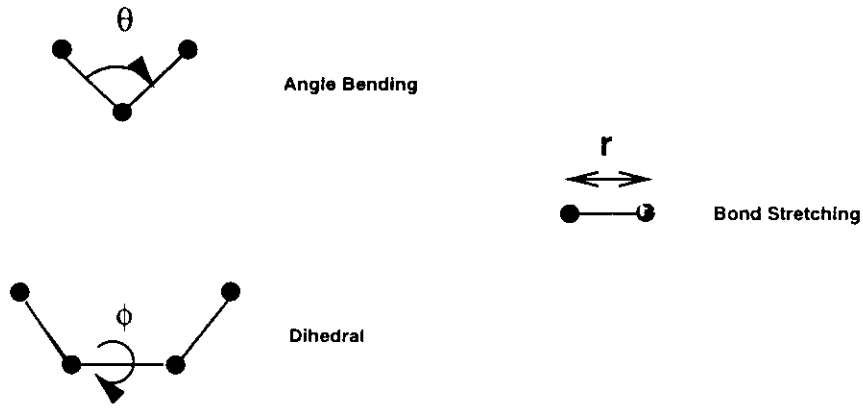


Figure 3. Bonded Interactions

bendings are utilized.[18] I must stress, that this part of the potential field is generally considered relatively less important than the non-bonded and the dihedral angles contributions. Firstly, there are no compelling spectroscopic reasons to be accurate for these high frequency vibrations: standard IR and Raman spectra of proteins are extremely messy and provide little information, although some information on the secondary structure can be obtained. Secondly, the time scale of the stretching and

bending modes is short and hopefully does not mix with more important motions with involve more than 4 atoms. In Fig. 4 I show a calculation of the stretching and bending frequencies corresponding to the bond stretching and angle bending CHARMm parameters[23]. Although in the picture the stretching and bending frequencies seem not to overlap, they do interact and mix in the actual dynamics.

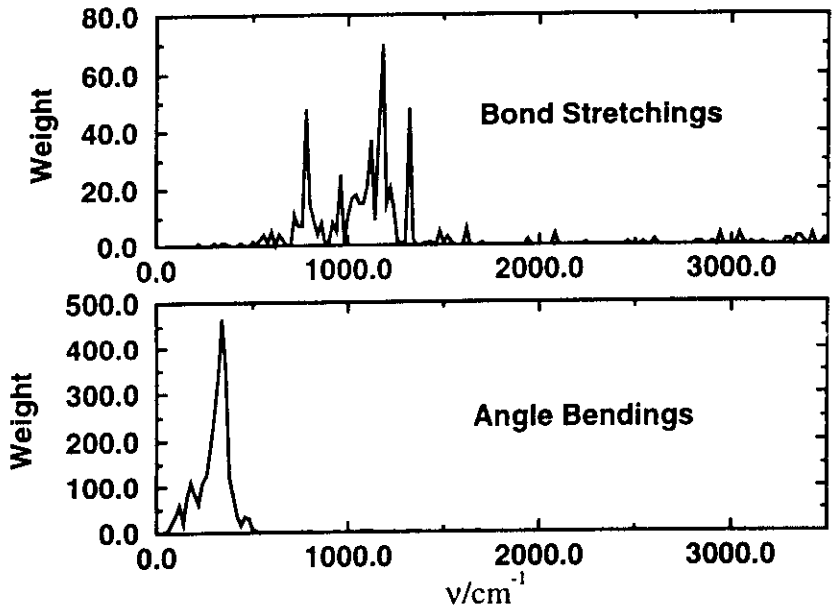
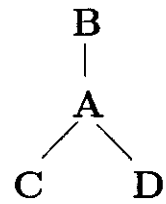


Figure 4. Distribution of bond stretching and angle bending frequencies

Dihedral Angles For any given set of four atoms connected by covalent bonds one can define dihedral angles ϕ between planes spanned by any triplets of atoms. Two types of dihedral angles are used by protein force fields. One is the proper torsion, or simply torsion, the other is the so-called improper torsion. Torsions involve four atoms connected by adjacent bonds, i.e. A1-A2-A3-A4. Torsional angles of segments of the protein backbone have specific and tabulated values if the segment belongs to a particular secondary structure. For instance, α-helices have $\phi \sim -60$ and $\psi \sim -50$.

Improper torsions have been introduced in the potential energy function to maintain the geometry of a certain type of atoms. In particular, this interaction is used to keep the amide and the carbonyl groups planar or, in united atoms force fields, to maintain $sp3$ carbons bound to one hydrogen atom tetrahedral. In the example below a possible improper torsion angle is defined by the plane A-B-C and C-D-A.



It is clear from the previous discussion that the role of dihedral angles is more important for the overall structure properties of the protein than those of stretchings and bendings. In reality, the forces involved in hindered rotations around a bond (torsions) are of both non-bonded and bonded origin. Indeed to the explicit bonded term (the third in Eq. 1) in many potential fields a contribution from dispersion and electrostatic interaction between atom 1 and 4 (see the numbering scheme at the beginning of the paragraph) is added. While the periodicity of the torsion (i.e. n in Eq. 1) is simply determined by consideration about the hybridization of the atoms involved, the K_ϕ parameters are chosen by fitting either to structural data or, in more recent developments, to *ab initio* calculations of rotational barriers.[18] Usually, the fit is performed on the simplest molecule possible which is a fragment of larger and more interesting molecules. It must be pointed out that because of the nature of this interaction a fit of the torsional parameters is not usually done without a combined fit of the non-bonded parameters.

Improper torsion potential functional forms may differ from the one in Eq. 1. A function used by the CHARMm force field is the following:

$$V(\phi) = \frac{1}{2} K_\phi (\phi - \phi_0)^2 \tag{2}$$

While the ϕ_0 or n parameters are selected from geometrical considerations, K_ϕ can be obtained from fitting to their corresponding frequencies derived from normal mode calculations (see for instance Ref. [17]).

Especially in the more recent potential fields, all the possible proper torsions occurring in a molecule are included in the calculation of the potential energy.[23, 18] On the contrary, only selected (and useful) improper torsions are included.

3.2. Non-bonded Interactions

The last term in Eq. 1 is the common functional expression for the non-bonded energy. Non-bonded interactions include all contacts involving two atoms separated by more than two bonds. Because of their long range nature they dominate the structural and aggregation properties of proteins. The *a priori* determination of the non-bonded parameters is the most difficult one. The earlier parameterizations of the non-bonded term used in MD simulations were inspired by the seminal work of Scheraga and Lifson on crystal packing of peptides.[24, 25] While the Lennard-Jones σ parameters were determined by fitting to *standard* van der Waals contact radii from crystal packing data, the ϵ and the atomic charges were empirically constructed to obtain lattice energies and to reproduce crystal structures of amides.[26, 15] In a more recent

approach the potential parameters were fitted to heat of vaporization and molecular volumes of organic liquids and aqueous solution of organic ions representative of fragments in the backbone and the side-chains of amino acids. In this fashion Jorgensen has derived a set of non-bonded parameters[21] for proteins which are capable of reproducing structures and volumes of small polypeptide crystals. Other approaches experimented in the literature derive atomic point charges directly by fitting to *ab-initio* electrostatic potential.[17, 18]

The dielectric constant contained in the Coulombic term in Eq. 1 has sometimes been used to take into account solvent effects implicitly, by increasing its value up to 80 (the dielectric constant of water). In the early simulations, but not exclusively, D was taken as equal or proportional to r, the interatomic distance. This has only physical justifications when the solvent is not explicitely included.

In all force fields used in MD simulation, each set of Lennard-Jones potential parameter (σ and ϵ) is assigned to a specific atom *type*. Atom *types* are chosen according to the atomic hybridization and the local chemical environment surrounding the atom. Consequently, the bonded parameters are assigned according to the *type* of the atoms involved in the interaction. I point out that atom of the same *type* can have different atomic charges because they can form bonds with atoms of different electronegativity in different parts of the protein. Usually, atomic point charges are provided for each individual residue.

In a following section I shall discuss problems connected to a correct handling of long range electrostatic interactions. Here, I will mention that in the great majority of protein simulation study the non-bonded interactions are spherically cut at a certain distance, generally in the order of 9-10 Å. In order to improve the convergence of the Coulombic term, which goes with $1/r$, it is common to impose a cutoff not over single couples of atoms, but between atomic groups. Such groups are defined to have a total charge of approximately zero. Thus, the Coulombic term acquires an r dependence of $1/r^3$ due to the interactions between the dipoles of two separate groups. The total non-bonded potential can then be written:

$$V = \sum_{\alpha<\beta} \sum_{ij} V\left(r_{\alpha i,\beta j}\right) + \sum_{\alpha} \sum_{ij} V\left(r_{\alpha i,\alpha j}\right) \tag{3}$$

Where α and β run on the groups of the protein and i and j labels the atoms belonging to each group.

Since an abrupt cut-off in the potential corresponds to a discontinuity in the potential function itself and its derivatives, an MD simulation carried out in these conditions will not conserve energy. Thus, in many cases a so-called switching function is used to smoothly switch the potential to zero. The most commonly used form is a third order spline function:

$$S\left(r, r_{on}, r_{off}\right) = \begin{cases} 1 & r \leq r_{on} \\ \frac{(r_{off}-r)^2(r_{off}+2r-3r_{on})}{(r_{off}-r_{on})^3} & r_{on} \leq r \leq r_{off} \\ 0 & r \geq r_{off} \end{cases} \tag{4}$$

The function is one at r_{on} and becomes zero at r_{off}. Thus, in case of cutoff on groups of atoms Eq. 3 becomes:

$$V = \sum_{\alpha<\beta} \left[\sum_{ij} V\left(r_{\alpha i,\beta j}\right) \right] S\left(r_{\alpha,\beta}, r_{on}, r_{off}\right) + \sum_{\alpha} \sum_{ij} V\left(r_{\alpha i,\alpha j}\right) \tag{5}$$

As a final remark, I will notice that atoms separated by three bonds (1-4 interactions) are commonly treated in a different way from the remaining interactions. Because of the partial bonded character of this interaction in many force fields only a fraction of the corresponding potential energy is included in the calculation. In practice both the Lennard-Jones and the Coulombic terms of 1-4 contact are multiplied by a scaling factor near to 0.5.

4. Covalent Topology and Potential Parameters

Compared to molecular liquids, simulating proteins, or any complex macromolecule, poses additional problems due to the complex covalent structure of the systems and to the related complexity of the potential force fields. Thus, it is worthwhile to show how this problem can be tackled in practice. I shall provide only my own solution and don't claim that this is the best and most efficient way of doing it.

Since we do not want to write a new program (or at least part of it) each time we decide to study another protein, It is essential that we build the covalent topology needed to evaluate the potential energy from the structure of its constituents, i.e. the amino acid. Also, we wish to minimize the size and the complexity of the actual input needed to construct this topology.

The minimal information we have to provide to describe the residue topology is the constituent atoms, the covalent bonds and, finally, the terminal atoms used to connect the unit to the rest of the chain. In addition, in order to assign the correct potential parameters to the bonds, bendings and torsions of the residue, we have to specify the type of each atom. I remind, also, that to each atom type corresponds a set of non-bonded parameters. In Fig. 5, I show a typical input I use for alanine residues.

When the bonding topology of the different residues contained in a protein is known, the units are linked together according to their occurrence in the sequence. In this fashion the total bonding topology is obtained. From this information, all possible bond angles are collected by searching for all possible couples of bonds which share one atom. Similarly, by selecting all couples of bonds linked among each other by a distinct bond, torsions can be obtained. To show how this is done in practice, I give in the Appendix a Fortran routine which for a given list of connections calculates all possible angle bending.

Once the intramolecular structure of the entire molecule is obtained, it becomes easy to assign to each non-bonded interaction, bond, angle bending and torsion its own potential parameters according to the chosen force field.

5. What to Simulate and How

Initial Conditions At the end of the operations described in the previous section, we have obtained a series of lists associated with the protein we wish to study. Thus, we dispose of the list of the protein constituent atoms and their relative types and charges. We have also calculated lists of the protein bonds, angle bendings, proper and improper torsions. Finally, we have assigned each item of these lists to its respective parameters.

What is now left is to assign initial positions and initial velocities to all particles in the system. If the latter task is easy and can be carried out in a standard fashion, the former is far from being so. Indeed, we need to know explicitly the protein X-ray

```
RESIDUE  ala ( Total Charge =   0.0 )
atoms
group
n       nh1    -0.47
hn      h       0.31
ca      ct1     0.07
ha      hb      0.09
group
cb      ct3    -0.27
hb1     ha      0.09
hb2     ha      0.09
hb3     ha      0.09
group
c       c       0.51
o       o      -0.51
end

bonds
cb   ca     n    hn     n    ca     o    c
c    ca     ca   ha     cb   hb1    cb   hb2    cb   hb3
end

imphd
n    -c   ca   hn     c    ca   +n   o
end

termatom n c
RESIDUE_END
```

Figure 5. Input topology for the alanine residue. Notice that each atom has assigned a label, a type and a charge. For instance the α carbon has label *ca*, type *ct1* and charge 0.07.

structure (the tertiary structure) which is not always available. Indeed, we have run into one major constraint faced when performing MD simulations on proteins. Because of the complexity of configurational space available to a polypeptide chain and the very long time-scales involved in conformational relaxation, starting a simulation from a random unfolded structure would not be a wise thing to do. Fortunately, in the past twenty years, there has been an increasing number of protein structures solved by X-ray and neutron scattering. More recently structures of proteins in solution have been determined by combination of 2-D NMR and molecular modeling.[8, 9] All the published solved structures are deposited to the Protein Data Bank of the Brookhaven National Laboratory and publicly available, after some delay from the date of publication.

Let us suppose that indeed the structure of our protein has been resolved. Because of the empirical nature of the whole available force fields, this starting configuration will not be at a minimum of the potential energy function. Before we can start a simulation we have to find a more suitable initial configuration. In order to do that a structural minimization is firstly carried out. Many methods can be used. If the protein is large the methods of choice would be those which use only first derivatives. The simplest to implement are the steepest descent method and the more efficient conjugate gradient method. Details of the methods can be found in Ref. [27].

MD Simulation When we have a structure at a local minimum or at least close to it, we can start the dynamic simulation which is carried out using the same techniques as for simulation of liquids. The standard algorithm to integrate the Newtonian equation of motions relies on the Verlet algorithm. Being one of the simplest time reversible integration algorithms available, it guarantees to the MD simulation stability and accuracy.

As we saw in Sec. 3.1 bonds involving hydrogens exhibit a very a high frequency stretching motion. In order to correctly integrate these high frequency stretchings a time-step of about 0.2 fs would need to be used. To avoid such a small integration step, rigid constraints are applied routinely to bonds involving hydrogens. Holonomic constraints can be imposed on the Newtonian dynamics by the method of the undetermined multipliers. A popular implicit solution to the undetermined multipliers used in conjunction with the Verlet integrator is known under the name of SHAKE. Details of this old algorithm are found in literature.[28, 29]

A few years ago, a multiple time step scheme which allows large savings of computer time combined with excellent stability and accuracy of the trajectory was introduced by Berne and coworkers.[30] Recently, this algorithm has been applied to simulation of proteins.[31, 32] This multiple timestep algorithm is based on the Liouville formulation of classical statistical mechanics. Given a observable A which is a function of a phase point $\Gamma(t)$ at time t, its time evolution will be given formally by:

$$A(\Gamma, t) = e^{iLt} A(\Gamma, 0) \tag{6}$$

Here L is the well known Liouville operator. If appropriately partitioned and discretized the time dependent operator $\exp(iLt)$ can be used to generate the integration algorithm to be used in MD simulation. This is done by applying the operator to the phase point $\Gamma(0)$. If one decomposes L in $L = L_a + L_b$, discretizes the time interval t in $\delta t = t/P$ and applies the Trotter theorem, the following expression is obtained:

$$e^{i(L_a+L_b)t} = \left[e^{i(L_a+L_b)\frac{t}{P}}\right]^P = \left[e^{iL_a\frac{\delta t}{2}} e^{iL_b\delta t} e^{iL_a\frac{\delta t}{2}}\right]^P + O\left(\delta t^3\right) \tag{7}$$

Various partitioning of the Liouville operator can be used. The simplest partitioning L in its coordinate and momenta component produces an integration algorithm identical to the velocity Verlet scheme. In this case

$$iL_a = \sum_i \mathbf{f}_i \frac{\partial}{\partial \mathbf{p}_i} \tag{8}$$

$$iL_b = \sum_i \dot{\mathbf{r}}_i \frac{\partial}{\partial \mathbf{r}_i} \tag{9}$$

Where $\mathbf{f}_i$ is the force acting on the particle i.

Both Berne[31] and Karplus,[32] who applied such a scheme to proteins, have partitioned L into a fast component containing both the intramolecular forces and the velocity dependent part (Eq. 9), and a slow component containing the intermolecular forces. Thus, the propagator for the time-step δt_1 which is compatible with the timescale of the intermolecular forces, can be written as

$$e^{iL\delta t_1} = e^{iL_a\frac{\delta t_1}{2}} e^{iL_b\delta t_1} e^{iL_a\frac{\delta t_1}{2}} \tag{10}$$

Explicitly L_a and L_b are defined as

$$iL_a = \sum_i \mathbf{f}_i^{inter} \frac{\partial}{\partial \mathbf{p}_i}$$
$$iL_b = \sum_i \dot{\mathbf{r}}_i \frac{\partial}{\partial \mathbf{r}_i} + \sum_i \mathbf{f}_i^{intra} \frac{\partial}{\partial \mathbf{p}_i} \tag{11}$$

With this partitioning one obtains a final expression for the propagator

$$e^{iL\delta t_1} = e^{iL_a \frac{\delta t_1}{2}} \left[e^{iL_b \delta t_2} \right]^n e^{iL_a \frac{\delta t_1}{2}} \tag{12}$$

Where $\delta t_2 = \frac{\delta t_1}{n}$. The practical advantage of this scheme of integration is that the inner part of the operator on the right hand side of Eq, 12 is propagated *without* recalculating the intermolecular forces in $\exp\left(iL_a \frac{\delta t_1}{2}\right)$ of the outer contribution. Although many different partitionings were attempted by Berne the one that includes bond stretching, angle bendings and torsions in L_b and all the remaining terms on L_a was found the most efficient.

6. Handling Long Range Forces

Long range electrostatic forces are involved at a fundamental level in many biological phenomena. For hydrated biomolecules the dielectric properties of the solvent are crucial to many processes such as folding of proteins, electron/proton transfer reactions and phase transitions. Indeed, while changes in the polarity of the solvent directly affects the folded state of a protein,[5] variation of the solvent ionic strength can cause phase transitions in DNA which change its biological functionality. Because of the charged nature of electrons and protons, electrostatic forces are also directly involved in the mechanism of their transfer reactions.[33, 34, 35, 36]

Thus, a correct representation of long-range interactions is essential for any theoretical study of such phenomena. Unfortunately, the most time-consuming part of an MD simulation is indeed the calculation of the long-range pairwise Coulombic forces. The computational cost for this interaction scales as $O(N^2)$ where N is the number of particles. Because of the slow $O(1/r)$ decay of the Coulombic interaction with respect to the interatomic distance, it is not appropriate to use short range cutoffs. As a consequence, other approaches need to be used.

The same type of computational problem is encountered in simulations of complex molecular fluids. The Ewald summation technique is probably the most suitable method for MD simulations and presents many advantages with respect to the other popular alternative, the reaction field method.[37, 38] Ewald summation has the attractive feature of being an exact technique for infinite periodic systems. This means that the interactions from all the periodic images of the MD box are rigorously summed over. In addition, with an appropriate choice of parameters the Ewald electrostatic potential acting on each particles of the system is a continuous function of the coordinates of the other particles. This last feature is crucial for the conservation of the total energy in MD simulations.

Detailed reviews of the Ewald method are available in many textbooks and articles.[39] In practice the Ewald summation technique by using an integral representation of the Coulombic potential splits total electrostatic energy of the system into a sum over the direct and the reciprocal space of the periodic system. Thus, for a

general molecular system composed of N atoms one obtains the following expression for the total electrostatic energy

$$V = \frac{1}{2}\sum_{ij}\sum_{n}\frac{q_i q_j}{|\mathbf{r}_i - \mathbf{r}_j + \mathbf{r}_n|} \tag{13}$$

$$= V_{qd} + V_{qr} - V_{qintra} \tag{14}$$

Where q_i and $\mathbf{r}_i$ are respectively the charge and position of the i-th particle and $\mathbf{r}_n$ is a lattice translation. The sum over n extends to all periodic images of the MD box, V_{qd} and V_{qr} are respectively the contributions from the direct and the reciprocal summations, while V_{qintra} is a correction due to the excluded electrostatic interactions discussed in Sec. 8. V_{qd} and V_{qr} are defined respectively as

$$V_{qd} = \frac{1}{2}\sum_{ij} q_i q_j \sum_{n}\frac{1}{|\mathbf{r}_{ij} + \mathbf{r}_n|} erfc(\alpha\,|\mathbf{r}_{ij} + \mathbf{r}_n|) \tag{15}$$

$$V_{qr} = \frac{1}{\pi V}\sum_{\mathbf{k}\neq 0}^{\infty}\frac{\exp\left(-\pi^2\,|\mathbf{k}|^2/\alpha^2\right)}{|\mathbf{k}|^2} S(\mathbf{k})\,S(-\mathbf{k}) - \frac{2\alpha}{\pi^{1/2}}\sum_{i} q_i^2 \tag{16}$$

Here $erfc$ is the complementary error function, V is the unit cell volume, $\mathbf{k}$ is a reciprocal lattice vector and α is an arbitrary parameter. In Eq. 16 $S(\mathbf{k})$ is the charge weighted structure factor defined as:

$$S(\mathbf{k}) = \sum_{i} q_i \exp(i\mathbf{k} \bullet \mathbf{r}_i) \tag{17}$$

Finally, the term V_{qintra} is given by:

$$V_{qintra} = \sum_{ij}\frac{erf(\alpha r_{ij})}{r_{ij}} \tag{18}$$

In the last equation the sum over i and j includes all the excluded bonded contacts.

The equations above are general and can be applied to any molecular systems whatever the size of the molecules and the number of constituent atoms. With the proper choice of the parameter α both real space and reciprocal space sums can be truncated. Given the numerical error of the standard MD algorithms, the complementary error function can be considered zero when $\alpha r \sim 3$. The choice of $\alpha = 0.3$ Å allows to truncate the real space sum to interactions within a range of 10 Å. The same degree of convergence is obtained in the reciprocal space sum choosing the maximum $\mathbf{k} \sim 1.5\text{Å}^{-1}$.

When the dimensions of the MD box increase a greater number of $\mathbf{k}$ vectors must be included in the reciprocal space sum in order to reach a reasonable degree of convergence with the same α. Although the evaluation of each $S(\mathbf{k})$ of the sum over $\mathbf{k}$ in Eq. 16 is easily vectorizable, on workstations this part of the calculation places a heavy computational burden. This problem becomes visible in simulation of large systems such as solvated proteins. As an example in Table 1 I compare timing for simulation of an hydrated reaction center protein carried out on HP and IBM workstations and on CRAY type computers. Unfortunately, satisfying solutions to the calculation of the electrostatics forces for large size systems do not exist as yet. An alternative to Ewald summations which presents great promises is the so-called fast multipole methods (FMM) introduced by Greengard and Rokhlin[40]. Briefly,

the FMM calculates the potential and the force on a particular charge as the sum of two parts: A direct part due to the particle-particle short range interactions and a multipolar part due to particle-multipole interactions. The elegant trick of

Table 1. The test was carried out on a solvated protein. The system was composed of 20627 atoms in an orthogonal box of 75 X 58 X 65 Å. A cutoff of 10 Å and 1.30 $Å^{-1}$ were used respectively for the direct and reciprocal sum with $\alpha = 0.32$ $Å^{-1}$. *CPU time* is the time for an MD step of 1.5 fs. *Slow down C94* and *Slow down J916* are the ratio between the CPU time used by the given machine and, respectively, by the CRAY C94 and by the CRAY J916.

Computer	CPU time (s)	Slow down C94	Slow down J916
HP 735	271.5	36.2	8.9
IBM SP2	109.2	14.6	3.6
IBM 580H	67.8	9.0	2.2
J916	30.6	4.1	1.0
C94	7.5	1.0	0.24

the method is to organize the multipole representation of charge distribution in hierarchically structured boxes and to calculate the multipoles of highest level boxes and the corresponding local field expansion with a recursive algorithm. The FMM was originally developed for finite systems, but more recently it has been adapted to periodic systems.[41] The CPU time required by the FMM algorithm increases logarithmically $[O(NlogN)]$ rather than with power of 3/2 of the number of particle $[O(N^{3/2})]$ found empirically for Ewald summation techniques. Compared to Ewald summation, the FMM is much more complicated to implement especially for periodic systems. Additionally, it provides energy and forces with a non negligible error for lower (i.e. computationally faster) expansions.

Since 1992 FMM have been applied to simulation of isolated proteins.[42, 43] For the much more interesting periodic systems there exists, to my knowledge, only one static implementation which shows that even for systems as large as 23000 particles Ewald summation performs better that the FMM.[44]

Some improvements of the Ewald sum performance can be obtained by using the multiple time step approach exploiting the fact that short ranged forces in the direct lattice evolve more rapidly than long range forces. To subdivide the direct space potential into terms characterized by different time scale, it is therefore sufficient to define spherical shells of increasing radius. The total direct space potential energy acting on the i-th particle can then be split up into:

$$V_{qd}(r_i) = \sum_l V_{qd}^l(r_i) \tag{19}$$

Each of $V_{qd}^l(r_i)$ contains only contacts of the l-th shell. Similarly, a subdivision of the reciprocal space term in k-vector dependent shells can be carried out. From a multiple time scale point of view, an appropriate partitioning of the Liouvillian would be done in parts containing contributions from direct and reciprocal space shells of the same time scale. For m shells one obtains:

$$L = L_0 + L_1 + ... + L_m \tag{20}$$

Where L_0 is an appropriate bottom down reference system. After the Trotter

expansion the propagator associated with the Liouvillian L can be written as:

$$e^{iL\Delta t} = [e^{iL_m \frac{\Delta t_m}{2}}[e^{iL_{m-1}\frac{\Delta t_{m-1}}{2}}...[e^{iL_1\frac{\Delta t_1}{2}}[e^{iL_0\Delta t_0}]^{P_0}e^{iL_1\frac{\Delta t_1}{2}}]^{P_1}... \\ e^{iL_{m-1}\frac{\Delta t_{m-1}}{2}}]^{P_{m-1}}e^{iL_m\frac{\Delta t_m}{2}}]^{P_m} + \sum_{k=1}^{m} O\left([\Delta t_k]^3\right) \quad (21)$$

The integration algorithm for such a subdivision uses $m+1$ time steps, i.e.

$$\Delta t_0, \Delta t_1 = P_0\Delta t_0, ..., \Delta t_m = P_{m-1}\Delta t_{m-1} \quad (22)$$

The advantage of using the propagator in Eq. 21 is that contributions to the Coulombic energy and forces due to the outer shells are calculated less frequently than inner shells. This can produce a significant saving of computational resources.

A recent implementation[45] has shown that an appropriate choice of three distance shells makes it possible to obtain speed-ups of up to 4 times with respect to standard single time step algorithm including bond constraints.

7. Simulations in Other Ensembles

In the last two decades constant pressure and constant temperature methods to simulate complex molecular condensed phases have been developed. More recently the very same methods have been applied to simulating proteins or, more in general, biomolecules.[46, 47, 48, 49] Indeed, the study of pressure dependent properties of system relevant to biology (proteins, membranes etc.) has been growing in importance in recent times.[50, 51] The pressure denaturation of monomeric proteins and the dissociation of oligomers have provided many insights on the microscopic mechanism of protein folding and on the role of solvent on this process. Here, I give a brief review of the most common MD methods to simulate systems at constant pressure and temperature stressing their application to simulation of proteins.

Andersen was the first to modify the equations of motion to sample ensemble other than the microcanonical.[52] In his approach the volume is a dynamical variable while the generalized force acting on this variable is proportional to the difference between the internal and an externally fixed pressure. It can be proved that the solution of the new set of equations of motion approximately samples the isobaric-isoenthalpic (NPH) ensemble. The fundamental idea of his method was to represent the effect of a suitable external reservoir by adding new degrees of freedom to the system and solving the equation of motion for the new extended Lagrangian (extended Lagrangian methods or ELM). Subsequently, Parrinello and Rahman[53, 54] showed that it is possible to write new equations of motion to simulate non-isotropic volume changes. Andersen's scheme was modified to allow for changes in the shape of the MD cell by introducing a new set of dynamical variables related to the cell metric. The algorithms for constant pressure were followed by an extended Lagrangian method for constant temperature by Nosé.[55]

A different approach is worth mentioning here. Indeed only a few simulation of biomolecules have been performed with ELM. Very popular among people simulating proteins[46, 47, 48] is the *weak coupling to an external bath* method (WC) of Berendsen *et al.*[56] which is simpler to implement than any of the ELM. The technique is based on a modified Langevin equation of motion in which the stochastic force is eliminated and the constant friction term is replaced with a variable friction proportional to the constraint. The method has been used to control temperature and

pressure in simulated system. In so far there is no proof that such equations sample a statistical mechanical ensemble. It must be pointed out that in contrast with the oscillatory behavior of the extended variables due to the second order character of the EL equations, the WC method does not produce a periodic behavior in the system variables coupled to the pressure and/or to the temperature baths.

A detailed derivation of the two classes of constant pressure methods can be found in literature[57]. Here I will focus on the choice of coupling scheme for the barostat. When dealing with molecules, one has the choice of coupling the barostat to each atomic degree of freedom or to each molecular centers of mass.[58] In the former case the effect of the barostat corresponds to a space scaling that affects the position of each atom and then the distances between atoms of the same molecule. In the latter case the scaling occurs only for the position of the molecular centers of mass.

Its is clear that the definition of the pressure tensor will depend on the chosen coupling scheme. Thus, according to the virial theorem one can write the atomic pressure:

$$P_{atomic} = \langle \Pi_{atomic} \rangle = \left\langle \frac{1}{3V} \sum_{\alpha} \sum_{i} \left(\frac{\mathbf{P}_{\alpha i}^2}{m_{\alpha i}} + \mathbf{r}_{\alpha i} \cdot \mathbf{f}_{\alpha i} \right) \right\rangle \tag{23}$$

and the molecular pressure:

$$P_{molecular} = \langle \Pi_{molecular} \rangle = \left\langle \frac{1}{3V} \sum_{\alpha} \left(\frac{\mathbf{P}_{\alpha}^2}{M_{\alpha}} + \mathbf{R}_{\alpha} \cdot \mathbf{F}_{\alpha} \right) \right\rangle \tag{24}$$

Here upper and lower case letters are used respectively for molecular and atomic properties. r is the coordinate, f the force and V the volume. Indices α and i apply respectively to molecules and atoms.

If the system is ergodic and the molecules do not dissociate (which is true for non dissociative potentials), the pressure equations 23 and 24 have been demonstrated equivalent for flexible molecules[29]. In this fashion, the choice of how to couple the barostat will not affect the results of the constant pressure simulation, but only how efficiently the system will relax and equilibrates at a certain pressure. For rigid or for small flexible molecules coupling to the center of mass has been preferred.[59] In simulations of large flexible molecules at constant pressure carried out with the weak coupling method of Berendsen only atomic coupling has been employed.[46]

However, if the ELM is used to simulate solvated proteins the choice of molecular coupling might be more advantageous. The detailed dynamics of the extended variables depends upon the value of the corresponding masses chosen. Within certain approximations the associated fundamental frequencies can be readily estimated by linearizing the equations of motion.[59] Thus, in the limit of large Q, the barostat mass, it can be shown that in case of isotropic volume variations (Andersen case) the characteristic frequency of the volume oscillations is given by:

$$\omega_Q = \left(\frac{3LB}{Q} \right)^{1/2}, \tag{25}$$

Where L is the length of the simulation box and B is the bulk modulus.

Experimentally, the internal compressibility of a protein is one order of magnitude smaller than that of water.[51] The coupling of the barostat to the atoms would involve effects due to the two distinct compressibilities. One related to the intermolecular

interactions between water molecules and between water molecules and the protein. The other associated with the intramolecular (bonded and non-bonded) interactions of the protein. This will produce oscillations of the extended variables with at least two distinct and uncoupled timescales: A slower frequency due to the water and a faster one due the protein. It is likely that in case of molecular coupling, where only one timescale related to intermolecular interactions is involved, the system will equilibrate easily.

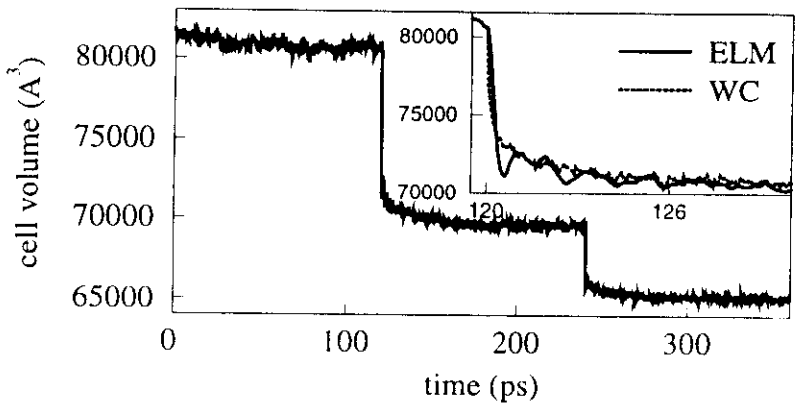


Figure 6. Relaxation of volume of the whole simulation cell of a system composed by one SOD solvated in 1457 SPC water molecules. The system, initially at atmospheric pressure, has been brought at 1000 MPa after 120 ps, and at 2000 MPa after other 120 ps. In the inset, 'the relaxation of the volume obtained by both ELM (solid line) and WC (dotted line).

In Fig. 6 I show the volume relaxation of a small dimeric protein superoxide dismutase, or SOD, solvated by 1451 water molecules when the pressure is brought from 0 to 10 Kbars and subsequently to 20 Kbars. The simulations were done in the NPT ensemble using the ELM with a molecular coupling of the stress tensor and at constant temperature and pressure with WC method. The total volume has relaxed within 30 ps from the change in pressure.

8. Molecular Dynamics Simulation of the Type I Crystal of BPTI

In this last section I will present an application of MD to a crystal of bovine pancreatic trypsin inhibitor (BPTI) which is a globular protein soluble in water. The high resolution structure of this protein has been one of the first to be resolved by X-ray crystallography. This aspect and its small size (832 atoms including hydrogens) can explain why in the past BPTI has been employed as a testing ground for force fields or new simulation methods. Indeed, the first MD simulations of a protein in vacuum[3] and of a solvated protein[16] were carried out on BPTI. From the biological point of view its importance is more limited and BPTI is used as a model compound to test theories, for instance on the detailed mechanism of protein folding.

The BPTI molecule is composed of 58 residues. The sequence is shown below:

```
ARG PRO ASP PHE CYS LEU GLU PRO PRO TYR THR GLY PRO
```

```
CYS LYS ALA ARG ILE ILE ARG TYR PHE TYR ASN ALA LYS
ALA GLY LEU CYS GLN THR PHE VAL TYR GLY GLY CYS ARG
ALA LYS ARG ASN ASN PHE LYS SER ALA GLU ASP CYS MET
ARG THR CYS GLY GLY ALA
```

It contains also 6 cysteine residues which form three sulphur bridges connecting different regions of the protein. Depending on the type of salt contained in the solution, BPTI can crystallize in at least two types of crystals: Type I and type II from respectively NaCl and K_2HPO_4 solutions. According to its most recent high resolution structure[60] the type I crystal is orthorhombic of space group P $2_12_12_1$ and cell constant a = 43.1 Å, b = 22.9 Åc = 48.6 Å. The unit cell contains 4 BPTI molecules. For each protein molecule the crystallographers have detected 60 crystallization water molecules. The X-ray coordinates of the asymmetric unit are available through the Protein Data Bank.

To limit the size of the system, I chose to include only one crystallographic unit cell in the MD simulation box. Thus, by applying the symmetry operations of the P $2_12_12_1$ space group to the coordinates of the BPTI protein and its crystallization water molecules contained in the asymmetric unit, I obtained an orthogonal MD box composed of 4288 atoms (4 BPTI and 240 water molecules). To this system extra molecules need to be added. Firstly, crystallographers can only observe water molecules with little or no mobility. Hydration water molecules with short residence times can not be detected by X-ray, but must be added to the system. Secondly, each BPTI molecule contains at its surface a few charged residues whose charge adds up to a total charge of $+6e$. These positive charges were neutralized in the simulation by 6 negative chloride ions.

In order to insert the extra molecules, a box with the same dimensions of the MD box containing a lattice of randomly oriented waters at the standard water density at 25 °C was generated. This box was then overlapped with the original MD box. All the water molecules of the lattice within less than 90 % of the van der Waals radius from any of the atoms of the proteins or of the crystallization waters were eliminated. In this fashion, a new simulation box containing 110 additional water molecules was obtained. Finally, 24 of the extra waters were chosen at random and replaced by the same number of chloride ions. The simulation box now contained 336 water molecules, 24 chloride ions and 4 molecules of BPTI, i.e. a total of 4600 atoms.

The whole atom potential parameters of the most recent CHARMm force field[23] were chosen for the protein molecules. Rigid bond constraints were applied only to bonds involving hydrogen atoms. To handle long range forces Ewald summation was used with α = 0.32 and a cutoff on the k-vectors of 1.3. A total of 842 reciprocal space vectors were included in the calculation of the energy.

Before the MD simulation was started, 3000 cycles of steepest descent minimization of both atomic coordinates and cell parameters were carried out. The aim here was not to find a good minimum, but to release some of the stress due to imperfections in the potential parameters and to the bad contacts added during the MD box set up. During the minimization, the total energy of the system decreased from about +75000 to −66000 KJmol^{-1}.

Subsequently, an MD simulation run in the NPT ensemble at 1 bar and 275 K was started. The masses of the barostat and thermostat were chosen to have the same oscillatory frequency of 60 cm^{-1}. The integration timestep was 2.0 fs. It took 1.5 s of CRAY C94 CPU time to advance the simulation of one step. After 1000 steps

Figure 7. The simulation box after a 200 ps run. The picture includes the 4 BPTI molecules, the 336 water molecules, the 24 chloride ions. Of all possible images only those molecule with their center of mass in the first cell are shown. The backbone of the BPTI molecule is shown as a ribbon of 1.5 Å .

during which the temperature of the barostat and thermostat were rescaled if they raised over 2000 K, the simulation was carried out for 200 ps. A pictorial view of the simulation cell is given in Fig. 7.

As shown in Fig. 8 during the first 100 ps of simulation the total potential energy of the system relaxed to an equilibrium value. This slow relaxation is typical of MD simulations of proteins. When simulating solvated proteins not in their crystalline environment this relaxation time tends to increase.

Despite the slow relaxation of the potential energy, the cell parameters after an initial decrease during the minimization phase did not change appreciably during the MD run. This is shown in Fig. 9 where the cell parameters are plotted. The average volume after 200 ps of simulation was $45,980$ Å^3 (error is estimated at ≤ 0.2 %) which corresponds to a 4 % decrease from the experimental volume of $47,967.7$ Å^3.

The root mean square difference between the experimental and the calculated coordinates of the BPTI molecules was computed. To this purpose, for each

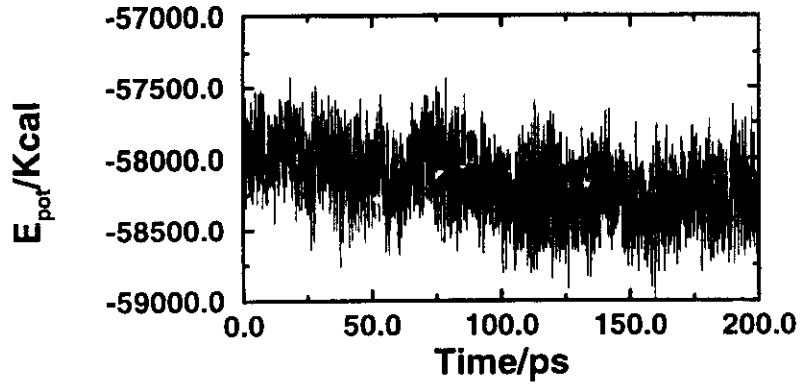


Figure 8. The total potential energy of the system as a function of the simulation time. This energy includes contributions from non-bonded and bonded terms.

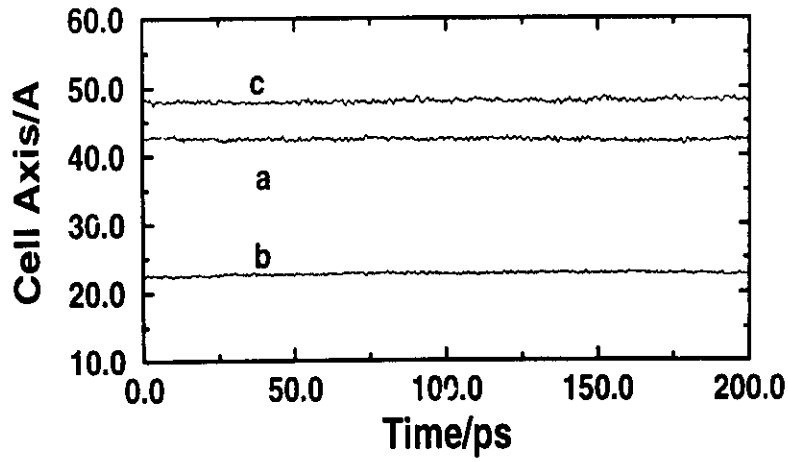


Figure 9. The cell axis a b and c during the simulation. The angles α, β and γ remained within 2 degrees from orthogonality.

configuration obtained in the simulation the root mean square difference between the coordinates of each one of 4 molecule of BPTI and their X-ray counterparts is minimized. This quantity (X-rms) was defined as

$$X - \mathrm{rms}_\alpha = \min \left(\sum_{i \in \alpha} (\mathbf{r}_i - \mathbf{r}_i^X)^2 \right)^{1/2} \tag{26}$$

Where α is the index of protein, $\mathbf{r}_i$ and $\mathbf{r}_i^X$ are respectively the atomic coordinates from the simulation and the X-ray.

In such minimizations only selected atoms are considered. In Fig. 10 I show the

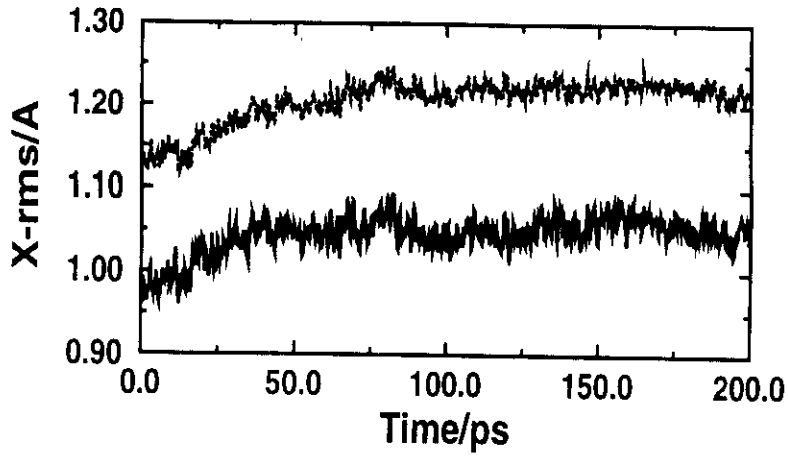


Figure 10. The X-rms deviation from the crystallographic structure versus simulation time. The dotted line is the X-rms calculated over all the heavy atoms, while the solid line refers to the C_α's.

cumulative X-rms calculated during the run for the C_α's and all the heavy atoms averaged over the 4 BPTI molecules. The two curves show the same behavior. After 75 ps of initial increase they reach a plateau which, in agreement with the behavior of the total potential energy in Fig. 8, signals the equilibration of the sample. For the 4 BPTI molecules contained in the simulation box the average C_α X-rms were calculated at 1.15, 1.07, 0.99 and 0.98 Å.

If the minimization of the difference between the coordinates X-ray and simulated coordinates is carried out for the whole 4 BPTI molecules at once, larger deviations, of the order of 1.2 Å are obtained.

The structural results of the MD simulation presented in this last section agrees well with experimental data. The discrepancy with the X-ray coordinates should not be misleading. Although experimental structures are nowadays more accurate than in the past, it has been shown[61] that errors between independent X-ray determinations are unlikely to be less than 0.5 to 0.8 Å even for small size proteins. Simulations of solvated proteins outside their crystalline environment show larger X-rms, of the order of 2.0 Åfor the heavy atoms.

In light of the results obtained in this MD simulation and in many others in literature, it is clear that the force fields and the simulation techniques available today are sufficiently sophisticated that qualitative and, hopefully, quantitative information regarding phenomena occuring in the timescale of picoseconds or at most nanoseconds can reliably be obtained.

Appendix: Routine to Compute Angle Bendings

```
      SUBROUTINE sbend(connct,n1,natom,ibendl,nbend,n2)

*************************************************************************
*   Time-stamp: <95/07/19 21:55:32 marchi>                              *
*                                                                       *
*     Find all bendings from the connection table. A bending is         *
*     defined as a three-atom {a,b,c}, in which a is connected          *
*     to b and c connected to b. Since each bending occurs in           *
*     the equivalent forms a-b-c and c-b-a, only the the first          *
*     is kept. The bending angle is the angle between the               *
*     connections a-b and b-c.                                          *
*                                                                       *
*     CONNCT  : Connection Table.                                 (I)   *
*               CONNCT(I,1)=Coord. number of atom I,                    *
*                          I=1,..,NATOM;                                *
*               CONNCT(I,J)=Neighbors of atom I,                        *
*                          J=2,..,1+CONNCT(I,1).                        *
*     N1      : Physical row dimension of CONNCT.                 (I)   *
*     NATOM   : Number of atoms.                                  (I)   *
*     IBENDL  : List with all bendings.                           (O)   *
*               IBENDL(1..3,I) contains the numbers of atom 1-2-3       *
*               in bending no. I, I=1,..,NBEND.                         *
*     NBEND   : Number of bends.                                  (O)   *
*     N2      : Physical row dimension of IBENDL.                 (I)   *
*                                                                       *
*=======================================================================*
*              Author:  Massimo Marchi                                  *
*              CEA/Centre d'Etudes Saclay, FRANCE                       *
*                                                                       *
*              - Wed Jul 19 1995 -                                      *
*************************************************************************

*---- This subroutine is part of the program ORAC ----*


*======================== DECLARATIONS ================================*

      IMPLICIT none

*----------------------------- ARGUMENTS -------------------------------*

      INTEGER n1,n2,natom,nbend,connct(n1,*),ibendl(3,*)

*------------------------- LOCAL VARIABLES -----------------------------*

      INTEGER ibend,i1,i2,i3,i4,a,b,c,d,coorda,coordb,coordc
      LOGICAL alldif
      CHARACTER*80 errmsg

*---------------------- EXECUTABLE STATEMENTS --------------------------*

      DO ibend=1,n2
         ibendl(1,ibend)=0
         ibendl(2,ibend)=0
         ibendl(3,ibend)=0
      END DO


      nbend=0
```

```
*=====================================================================
*--- Start loop on all atoms ---------------------------------------
*=====================================================================

      DO i1=1,natom
         a     =i1
         coorda=connct(a,1)

*=====================================================================
*------- Now loop on neighbors of i1 ---------------------------------
*=====================================================================

         DO i2=1,coorda
            b     =connct(a,1+i2)
            coordb=connct(b,1)

*=====================================================================
*---------- Third neighbor -------------------------------------------
*=====================================================================

            DO i3=1,coordb
               c     =connct(b,1+i3)
               IF(a .ne. b .AND. a .ne. c  .AND. b .ne. c) THEN
                  coordc=connct(c,1)
                  DO i4=1,coordc
                     d     =connct(c,1+i4)
                     IF(d .EQ. a) GOTO 100
                  END DO
                  alldif=.true.
               ELSE
                  alldif=.false.
               END IF
               IF(alldif) THEN
                  DO ibend=1,nbend
                     if(a .EQ. ibendl(3,ibend) .AND.
     .                    b .EQ. ibendl(2,ibend) .AND.
     .                    c .EQ. ibendl(1,ibend)      ) THEN
                        GOTO 100
                     END IF
                  END DO

*=====================================================================
*---------------- If alldiff and not already there add bend to the list-
*=====================================================================

                  nbend=nbend + 1
                  IF(nbend .GT. n2) THEN
                     errmsg='In SBEND: physical dimension of'//
     x                    ' ibendl are insufficient. ABORT!'
                     CALL xerror(errmsg,80,1,2)
                  END IF
                  ibendl(1,nbend)=a
                  ibendl(2,nbend)=b
                  ibendl(3,nbend)=c
               END IF
100            CONTINUE
            END DO
         END DO
      END DO
```

```
*----------------- END OF EXECUTABLE STATEMENTS -----------------------*

      RETURN
      END
```

References

[1] Rahman, A.;. *Phys. Rev.*, 136:405, 1964.
[2] Verlet, L.;. *pr*, 159:98, 1967.
[3] McCammon, J. A.; Wolynes, P. G.;. *J. Chem. Phys.*, 66:1452, 1977.
[4] Deisenhofer, J. O.; Steigemann, W. R.;. *Acta Crystallogr.*, B 31:238, 1975.
[5] Creighton, T. E.;. *Biochem. J.*, 270:1, 1990.
[6] Bruñger, A. T.;Kuriyan, J.; Karplus, M.;. *Science*, 235:459, 1987.
[7] Jack, A.; Levitt, M.;. *Acta Crystallogr.*, A 34:931, 1978.
[8] Kaptein, R.; Zuiderweg, E. R. P.; Scheek, R. M.; Boelens, R.; van Gunsteren, W. F.;. *J. Mol. Biol.*, 182:179, 1985.
[9] Clore, G. M.; Gronenborn, A. M.; Bruñger, A. T.; Karplus, M.;. *J. Mol. Biol.*, 186:435, 1985.
[10] Monge, A.; Friesner, R. A.; Honig, B.;. *Proc. Nat. Acad. Sci. U.S.A.*, 91:5027, 1994.
[11] Monge, A.; Lathrop, J. P.; Gunn, J. R.; Shenkin, P. S.; Friesner, R. A.;. *J. Mol. Biol.*, 247:995, 1995.
[12] Warme, P. K.; Scheraga, H. A.;. *Biochemistry*, 13:757, 1974.
[13] Roterman, I. K.; Lambert, M. H.; Gibson, K. D.; Scheraga, H. A.;. *J. Biomol. Struct. Dyn.*, 7:421, 1989.
[14] Clark, M.; Cramer, R. D.; van Oppenbosch, N.;. *J. Comput. Chem.*, 10:982, 1989.
[15] Gelin, B. R.; Karplus, M.;. *Biochemistry*, 18:1256, 1979.
[16] van Gunsteren, V. F.; Karplus, M.;. *Nature*, 293:677, 1981.
[17] Weiner, S. J.; Kollman, P.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S., Jr. ; Weiner, P.;. *J. Am. Chem. Soc.*, 106:765, 1984.
[18] Cornell, W. D.; Cieplak, P.; Bayly, C. I.; Gould, I. R.; Merz, K. M., Jr.; Ferguson, D. M.; Spellmeyer, D. C.; Fox, T.; Caldwell, J. W.; Kollman, P.;. *J. Am. Chem. Soc.*, 117:5179, 1995.
[19] van Gunsteren, W. F.; Berendsen, H. J. C.;. *Groningen Molecular Simulation (GROMOS) Library Manual.* Biomos, Groningen, 1987.
[20] Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; Slater, D. J.; Swaminathan, S.; Karplus, M.;. *J. Comput. Chem.*, 4:187, 1983.
[21] Jorgensen, W. L.;Tirado-Rives, J.;. *J. Am. Chem. Soc.*, 110:1657, 1988.
[22] Maple, J. R.; Hwang, M. J.; Stockfish, T. P.; Dinur, U.; Waldman, M.; Ewig, C. S.; Hagler, A. T.;. *J. Comput. Chem.*, 15:162, 1994. See also references therein.
[23] Polygen Corp. Parameter and topology files for charmm, version 22, Copyright 1986, Release May 1993. The.
[24] Momany, F.; McGuire, R.;Burgess, A. ; Scheraga, H.;. *J. Phys. Chem*, 79:2371, 1975.
[25] Warshel, A.; Levitt, M.; Lifson, S.;. *J. Mol. Spectrosc.*, 33:84, 1970.
[26] Hagler, A.; Euler, E.; Lifson, S.;. *J. Am. Chem. Soc.*, 96:5319, 1974.
[27] Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P;. *Numerical Recepies in Fortran. The Art of Scientific Computing.* Cambridge University Press, second edition edition, 1992.
[28] Ryckaert, J. P.; Ciccotti, G.; Berendsen, H. J. C.;. *J. Comput. Phys.*, 23:327, 1977.
[29] Ciccotti, G; Ryckaert, J. P.;. *Computer Phys. Rep.*, 4:345, 1986.
[30] Tuckerman, M.; Martyna, G. J.; Berne, B. J.;. *J. Chem. Phys.*, 97:1990, 1992.
[31] Humphreys, D. D.; Friesner, R. A.; Berne, B. J.;. *J. Phys. Chem*, 98:6885, 1994.
[32] Watanabe, M; Karplus, M.;. *J. Phys. Chem*, 99:5680, 1995.
[33] Marcus, R. A.; Sutin, N.;. *Biochim. Biophys. Acta*, 811:265, 1985.
[34] Moser, C. C,; Keske, J. M.; Warncke, K.;Farid, R. S.; Dutton, P. L.;. *Nature*, 355:796, 1992.
[35] Marchi, M.; Gehlen, J. N.; Chandler, D.; Newton, M.;. *J. Am. Chem. Soc.*, 115:4178, 1993.
[36] Warshel, A.; Chu, Z.-T.; Parson, W. W;. *J. Photochem. Photobiol. A - Chem.*, 82:123, 1994.
[37] Barker, J. A.; Watts, R. O.;. *Mol. Phys.*, 26:789, 1973.
[38] Barker, J. A.;. The problem of long-range forces in the computer simulation of condensed matter. volume 9, page 45. NRCC Workshop Proceedings, 1980.
[39] de Leeuw, S. W.; Perram, J. W.; Smith, E. R.;. *Proc. Royal Soc.*, A 373:27, 1980.
[40] Greengard, L.; Rokhlin, V.;. *J. Comput. Phys.*, 73:325, 1987.
[41] Schmidt, K. E.; Lee, M. A.;. *J. Stat. Phys.*, 63:1223, 1991.

[42] Hong-Qiang Ding; Karasawa, N.; Goddard, W.A., III.;. *Chem. Phys. Lett.*, 196:6, 1992.
[43] Board, J.A., Jr.; Causey, J.W.; Leathrum, J.F., Jr.; Windemuth, A.; Schulten, K.;. *Chem. Phys. Lett.*, 198:89, 1992.
[44] Shimada, J.; Kaneko, H.; Takada, T.;. *J. Comput. Chem.*, 15:28, 1994.
[45] Procacci, P.; Marchi, M.;. Taming the Ewald Sum in Molecular Dynamics Simulations of Solvated Proteins via a Multiple Time Scale Algorithm. *J. Chem. Phys.*, August 1995. Submitted.
[46] Kitchen, D. B.; Reed, L. H.; Levy R. M;. *Biochemistry*, 31:10083, 1992.
[47] Bassolino-Klimas D.; Alper H. E.; Stouch, T. R.;. *Biochemistry*, 32:12624, 1993.
[48] Marrink, S. J.; Berendsen, H. J. C.;. *J. Chem. Phys.*, 98:4155, 1994.
[49] Paci, E.; Marchi, M.;. Constant Pressure Molecular Dynamics Techniques applied to Complex Molecular Systems and Solvated Proteins. *J. Chem. Phys.*, August 1995. Submitted.
[50] Frauenfelder, H.; Alberding, N.A.; Ansari, A.; Braunstein, D.; Cowen, B.R.; Hong, M.K.; Iben, I.E.T.; Johnson, J.B.; Luck, S.; Marden, M.C.; Mourant, J.R.; Ormos, P.; Reinisch, L.; Scholl, R.; Shulte, A.; Shyamsunder, E.; Sorensen, L.B.; Steinbach, P.J.; Xie, A.; Young,R.D.; Yue K.T.;. *J. Phys. Chem*, 94:1024, 1990.
[51] Silva, J. L.; Weber, G.;. *Annu. Rev. Phys. Chem.*, 44:89, 1993.
[52] Andersen, H. C.;. *J. Chem. Phys.*, 72:2384, 1980.
[53] Parrinello, M.; Rahman, A. *Phys. Rev. Lett.*, 45:1196, 1980.
[54] Parrinello, M.; Rahman, A.;. *J. appl. Phys.*, 52:7182, 1981.
[55] Nosé, S.;. *Mol. Phys.*, 52:255, 1984.
[56] Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R.;. *J. Chem. Phys.*, 81:3684, 1984.
[57] Allen, M. P.; Tildesley, D. J.;. Clarendon Press, Oxford, 1987.
[58] Ferrario, M.; Ryckaert, J. P.;. *Mol. Phys.*, 54:587, 1985.
[59] Nosé, S. ; Klein, M. L.;. *Mol. Phys.*, 50:1055, 1983.
[60] Wlodawer, A.; Deisenhofer, J. O.; Huber, R. *J. Mol. Biol.*, 193:145, 1987.
[61] Smith, L. J.; et al.;. *Nature: Structural Biology*, 1:301, 1994.