UNIVERSITY OF TRENTO

Faculty of Mathematical, Physical and Natural Sciences

Physics Department

Ph.D. Thesis in Physics

Computer simulation of Biological Systems

Supervisors:
Dr. Maurizio Dapor
Dr. Giovanni Garberoglio

Ph.D. candidate:
Anna Battisti

Doctoral School in Physics, XXIV cycle
Trento, 24th Febrary 2012

AUTHOR'S EMAIL ADDRESS:
battisti@fbk.eu
battistia@libero.it

Introduction

This thesis investigates two biological systems using *atomistic modelling* and *molecular dynamics simulation*. The work is focused on: (a) the study of the interaction between a segment of a DNA molecule and a functionalized surface; (b) the dynamical modelling of protein tau, an intrinsically disordered protein. We briefly describe here the two problems; for their detailed introduction we refer respectively to chapter 4 and chapter 5.

The interest in the study of the *adsorption of DNA on functionalized surfaces* is related to the considerable effort that in recent years has been devoted in developing technologies for faster and cheaper genome sequencing. In order to sequence a DNA molecule, it has to be extracted from the cell where it is stored (e.g. the blood cells). As a consequence any genomic analysis requires a purification process in order to remove from the DNA molecule proteins, lipids and any other contaminants. The extraction and purification of DNA from biological samples is hence the first step towards an efficient and cheap genome sequencing. Using the chemical and physical properties of DNA it is possible to generate an attractive interaction between this macromolecule and a properly treated surface. Once positioned on the surface, the DNA can be more easily purified. In this work we set up a detailed molecular model of DNA interacting with a surface functionalized with amino silanes. The intent is to investigate the free energy of adsorption of small DNA oligomers as a function of the pH and ionic strength of the solution.

The tau protein belongs to the category of Intrinsically Disordered Proteins (IDP), which in their native state do not have an average stable structure and fluctuate between many conformations. In its physiological state, tau protein helps nucleating and stabilizing the microtubules in the axons of the neurons. On the other hand, the same tau - in a pathological aggregation - is involved in the development of the Alzheimer disease. IDPs do not have a definite 3D structure, therefore their dynamical simulation cannot start from a known list of atomistic positions, like a protein data bank file. We first introduce a procedure to find an initial dynamical state for a generic IDP, and we apply it to the tau protein. We then analyze the dynamical properties of tau, like the propensity of residues to form temporary

secondary structures like $\beta-$sheets or $\alpha-$helices.

Contents

Part I

Theory

Chapter 1

Overview of Molecular Modelling

In this work we use Molecular Modelling to study biological systems on the atomic scale, representing and manipulating the structure of molecules to study properties that are dependent on these three-dimensional structures. Molecular Modelling encompasses quantum mechanics and classical mechanics, and uses minimization procedures, dynamical simulations, conformational analysis and other computer based methods to understand and predict the behaviour of molecular systems. A first step is the selection of a model to describe the intra- and inter-molecular interactions in the system; these models enable the computation of the energy of any arrangement of the atoms and molecules in the system, and allow to determine how the energy of the system varies with the positions of the atoms. Secondly, one can make a calculation such as an energy minimisation, a Molecular Dynamics (MD) or Monte Carlo (MC) simulation, or a conformational search. Finally, after a check that the calculation has been performed properly (usually a comparison with experimental data) the results can be analysed to calculate specific properties of the system.

The 1950s were very important for the development of this discipline. In this year Watson and Crick discovered the structure of DNA [1]. The determination of its relationship to the biological function of this molecule had a tremendous impact, and was the cornerstone of the paradigm of modern biochemistry and molecular biology. The so-called *lock-and-key* paradigm established the primary importance of molecular structure for the function of

biological molecules, and the need to investigate this relationship to advance our understanding of the processes of life [2].

The subsequent development of Molecular Modelling was very fast, and gained effectiveness and popularity when computers were introduced in this field, with an improvement in hardware and software tools that is still progressing. In 1963 Ramachandran made one of first experiments that showed the potential of a computational approach in understanding the atomic details of biomolecules: he made a prediction of the allowed conformations of amino acids, the basic building blocks of proteins, using a simple hard-sphere model [3]. With the ever increasing computing power and with faster and efficient numerical algorithms, computational chemistry can be now used very efficiently to solve complex chemical and biological problems. In a computational approach the principal required inputs are: molecular energies and structures, atomic charges, and surface properties. So there is a close connection between theory and experiment: computational models evolve as more experimental data become available; biological theories are developed and new experiments are performed, as a result of computational results.

An effective implementation of Molecular Modelling techniques implies a through understanding of the method and of the nature of the data used in the parameterization of the models. With this knowledge one can define the limits of applicability and one can develop new tools. This entire field is based on approximate solutions, it is therefore important to understand how each of these approximations determines the level of accuracy that can be expected. The models used to represent a system are mathematical descriptions (sometime visual descriptions) and give us a way to describe and predict microscopic properties without performing the complex mathematical calculation dictated by an *ab-initio* theory; the mathematical complexity might be so great that the exact solution of a problem is just not feasible. Different types of approximation are possible: one can use a coarse-grained rather than an atomistic description; one can treat the most relevant part of the system with its complexity and the remaining part using a mean field approximation. One can also use simplified semi-empirical functions that are able to reproduce experimental results with a resonably small set of fitted parameters.

1.1 Ab-initio Calculations

The term *ab-initio* refers to computations which are derived from the basic principles of quantum mechanics, such as the Schrödinger equation, to describe the motion of electrons and nuclei, with no inclusion of parameters derived from experimental data. Since a complete solution of Schrödinger equation cannot be obtained, *ab-initio* methods are usually mathematical approximations of the full theory, at various levels of accuracy.

The most common type of *ab-initio* method is the Hartree-Fock (HF) method: by using the variational principle, it can calculate the ground state wave function and ground state energy of a quantum many-body system. The starting point are the spin-orbitals, a set of one-electron wave functions. On these acts the Fock operator, an effective one-electron Hamiltonian operator, that takes into account kinetic energy, internuclear repulsion energy, nuclear-electronic Coulomb attraction, and Coulomb repulsion between electrons. Inclusion of the latter term, that is considered in a mean-field theory context, is the critical point of the method: neglecting electron correlation can lead to large deviations from experimental results.

Most computations begin with a HF calculation, followed by further corrections for the explicit electron-electron repulsion. Among these procedures are the Møller-Plesset perturbation theory (MPn, where n is the order of correction) [4], the Generalized Valence Bond (GVB) [5], the Configuration Interaction (CI) [6], and the Coupled Cluster theory (CC) [7]; these methods are called correlated calculations.

A different method is Quantum Monte Carlo (QMC) [8]. There are several flavors of QMC, namely variational, diffusion, and Green's function. These methods work with an explicitly correlated wave-function and evaluate integrals numerically using a Monte Carlo integration. These calculations can be very time consuming, but they are very accurate.

An alternative *ab-initio* method is the Density Functional Theory (DFT) [9], in which the total energy is expressed in terms of the total electron density rather than through the wave-function.

A positive aspect of *ab-initio* methods is that they converge to the exact solution by making the approximations progressively more accurate, and the error sufficiently small in magnitude. The unfavorable aspect is that these methods often take enormous amounts of computer CPU time, memory and disk space, and are therefore very expensive. In practice they can be applied

only to very small systems.

1.2 Semi-empirical Calculations

Semi-empirical calculations have the same general structure as HF, but certain pieces of information, such as two electron integrals, are approximated or completely omitted, to reduce the complexity of the system. In order to correct for the errors introduced by omitting part of the calculation, and to give the best possible agreement with experimental data, the functionals are parameterized. The parameters are determined by fitting the data of a suitable database, entailing either experimental properties or properties derived by *ab-initio* calculations. This method is much faster than *ab-initio* calculations. On the other hand, if the studied molecule is similar to one found in the database used by the method, the results may be very good; but if the molecule is significantly different from anything in the known set, the results may be very poor. As a matter of fact, semi-empirical calculations have been very successful in the description of organic molecules, where there are only a few different atoms even in large molecules.

1.3 Molecular Mechanics

If a molecule is too complex to use a semi-empirical calculation, it is still possible to model its behaviour totally avoiding an explicit use of a quantum mechanics formalism. The Molecular Mechanics method uses an algebraic expression for the total energy, that consists of simple classical terms. A harmonic oscillator potential may be used to describe the energy associated with bond stretching, valence angle bending, and dihedral rotation; and a classical potential (e.g. Lennard-Jones) may be used to describe intramolecular or intermolecular forces, such as Van der Waals interactions. Hydrogen bonds can be defined based on a simple geometric criterium, specified by the maximum hydrogen-donor-acceptor angle and donor-acceptor distance. All parameters in these functions are obtained from experimental data or *ab-initio* calculations.

In Molecular Mechanics, the database of properties used to parameterize a system is crucial for the success of the procedure. The parameterization allows the modelling of very large molecules, such as proteins and segments

of DNA, making it the primary tool of computational biochemistry. The main limit of this method is that there are many chemical process that are not defined within the method (e.g. we cannot consider classical reactions or electronic excitation processes).

1.4 Molecular Simulation

Molecular Simulation is a computational experiment performed on a molecular model. The molecular simulation software may be based on different approaches: Monte Carlo (MC) simulation, Molecular Dynamics (MD) simulation, or Car-Parrinello Molecular Dynamics (CPMD). In this work we used GROMACS, a MD software package [10, 11]. A very important part in the simulation of a biological molecule is the Molecular Graphics software, which allows a first insight into a generated trajectory through the visualization of the molecule conformations. Molecular graphics systems have greatly evolved in recent years, and a number of computer programs are now available to visualize complex systems in the 3D space, like VMD [12].

Chapter 2

Atomistic Models and Force Field

The results that one can obtain in a computational experiment on chemical or physical systems are strictly dependent on the mathematical model used to represent the energy terms that define the interatomic interactions. The quantum mechanics (QM) *ab-initio* approach is inapplicable to biochemical systems involving macromolecules that contain several thousand atoms, plus their environment due to its very large computational cost. To investigate the systems that are the object of the present study a different approach is necessary, and we have used Molecular Dynamics, implementing force fields defined by the Molecular Mechanics approach.

2.1 Empirical energy functions

Empirical energy functions can fulfill the accuracy and feasibility requirements of computational studies of biochemical and biophysical systems. The equations of empirical energy functions include relatively simple terms to describe the physical interactions; the atomistic model is used, so that atoms are the smallest particles in the system, rather than electrons and nuclei as in QM. These two simplifying assumptions allow studying structural and dynamical properties of biological molecules, and a very good accuracy can be achieved by optimizing the parameters of the model.

Model	Degrees of freedom		Some computable properties
	Considered	Removed	
QM	nucleus, electrons	nucleons	chemical reactions
Atomistic Force Field	atoms, dipoles	electrons	interactions, structural properties
Implicit solvent	solute atoms	solvent atoms	folding of macromolecules

2.2 Parameterization of empirical force fields

The empirical energy represents the potential energy $V(\mathbf{r})$ as a function of all the relevant degrees of freedom of a given system. A typical function used in a classical MD code is:

$$
\begin{aligned}
V(\mathbf{r}) = & \sum_{bonds} \frac{1}{2} k_{b_i} (b_i - b_{0i})^2 + \sum_{angles} \frac{1}{2} k_{\theta_i} (\theta_i - \theta_{0i})^2 \\
& + \sum_{improper\, dihedral} \frac{1}{2} k_{\gamma_i}^{(idh)} (\gamma_i - \gamma_{0i})^2 + \sum_{dihedrals} k_{\gamma_i}^{(dh)} [1 + \cos(n\gamma_i - \delta_i)] \\
& + \sum_{atom\, pairs} \Big[\Big(\frac{C_{ij}^{12}}{r_{ij}^{12}} - \frac{C_{ij}^{6}}{r_{ij}^{6}} \Big) + \frac{1}{4\pi\epsilon} \frac{q_i q_j}{r_{ij}} \Big]
\end{aligned}
\tag{2.1}
$$

(the symbols introduced here are explained in the next sections).
The first four terms represent the bonded interactions, the last sum is the non-bonded term; they entail the parameters describing the equilibrium 3D structure (e.g. b_{0i}, θ_{0i}...) and the force parameters (e.g. k_{b_i}, k_{θ_i}...) . The former parameters are obtained experimentally by X-ray crystallography, NMR spectroscopy, electron microscopy (EM); or by homology modelling, Molecular Dynamics (MD), or Monte Carlo (MC) simulations. On the other hand, the force parameters are associated with the particular type of interacting atoms, and they are fitted to reproduce experimental data and the quantum mechanical calculations, where available. The combination of potential energy function and parameters is called a `Force Field`.

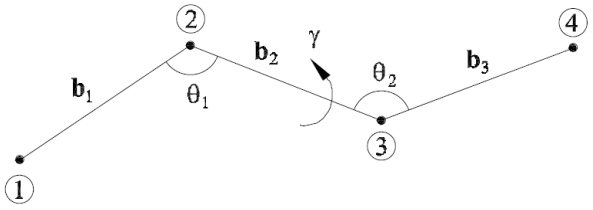


Figure 2.1: Schematic representation of the relevant degrees of freedom.

2.3 Bonded interaction

In reference to Fig. 2.1, which illustrates the most common potential terms between atoms connected by chemical bonds, denoting by $\mathbf{r}_i = \{r_{i\alpha}\}$ ($\alpha = x, y, z$) the Cartesian coordinates of the $i-th$ atom, we introduce the following parameters:

- the bond distance:

$$\mathbf{b}_i = \mathbf{r}_{i+1} - \mathbf{r}_i \;\; (i = 1, 2, \ldots N) \; ; \tag{2.2}$$

- the bond angle θ_i between atoms $i, i+1$ e $i+2$ ($i = 1, 2, \ldots N$), that is between two bonds sharing a common atom:

$$\cos\theta_i = -\frac{\mathbf{b}_i\,\mathbf{b}_{i+1}}{b_i\,b_{i+1}} \; ; \tag{2.3}$$

- the torsional angle γ_i between the plane encompassing atoms i, $i+1$, i+2 and the plane encompassing atoms i+1, i+2 and i+3. Introducing the normal vectors $\boldsymbol{\xi}_i$, $\boldsymbol{\xi}_{i+1}$ to these two planes:

$$\boldsymbol{\xi}_i = \mathbf{b}_i \times \mathbf{b}_{i+1} \qquad \boldsymbol{\xi}_{i+1} = \mathbf{b}_{i+1} \times \mathbf{b}_{i+2}, \tag{2.4}$$

  γ_i is defined by:

$$\cos\gamma_i = -\left(\frac{\boldsymbol{\xi}_i}{|\boldsymbol{\xi}_i|}\right) \cdot \left(\frac{\boldsymbol{\xi}_{i+1}}{|\boldsymbol{\xi}_{i+1}|}\right) \tag{2.5}$$

2.3.1 Bond stretching

Many types of interaction potentials can be used to model a covalent bond in a molecular structure, such as the Morse potential or the finitely

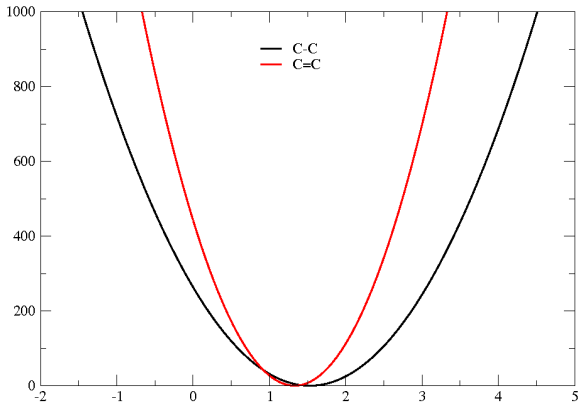
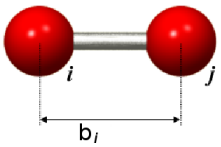


Figure 2.2: Behaviours of the two bond stretching interaction C−C [black line] and C=C [red line].

extendible nonlinear elastic (FENE) potential. However, the most common potential used is the harmonic bond potential:



$$V_s = \sum_i \frac{1}{2} k_{b_i} (b_i - b_{0i})^2 \tag{2.6}$$

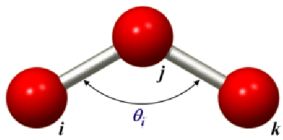
b_i and b_{0i} are respectively the distance and the equilibrium distance between two atoms. The parameters b_{0i} are associated with the 3D structure, while the set k_{b_i} is characteristic of atoms i and $i+1$, and of the type of bond between them. In Fig. 2.2 the parameterization for C−C and C=C bonds is shown. This example highlights how the same function, with different parameters, can represent the interaction of a pair of atoms with a different type of bonds.

When the potential is known, one can calculate the force acting on the $i-th$ atom:

$$\mathbf{f}_i = -\frac{\partial V_s}{\partial \mathbf{r}_i} = -\frac{\partial V_s}{\partial r_{ij}} \frac{r_{ij}}{\mathbf{r}_i} = k_{b_i}(b_i - b_{0i}) \frac{\mathbf{b}_i}{b_i} \tag{2.7}$$

2.3.2 Harmonic angle potential

The deviation of angles from their equilibrium value is frequently described by a harmonic term called bending potential:



$$V_b = \frac{1}{2} k_{\theta_i} (\theta_i - \theta_{0i})^2 \tag{2.8}$$

The forces exchanged between atoms are:

$$\mathbf{f}_i = -\frac{\partial V_b}{\partial \mathbf{r}_i} \qquad \mathbf{f}_k = -\frac{\partial V_b}{\partial \mathbf{r}_k} \qquad \mathbf{f}_j = -\mathbf{f}_i - \mathbf{f}_k \tag{2.9}$$

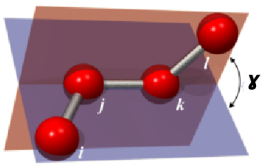
The bond-stretching and angle-bending terms are often regarded as "hard" degrees of freedom because a large energy is required to have a significant deviation from their equilibrium value. Most of the variation in structure of a molecule is thus due to the complex interplay between the torsional and non-bonded contributions, and the knowledge of barriers to rotation around chemical bonds is fundamental to understand the structural properties of the molecule.

2.3.3 Torsion potential

Usually two types of torsion potentials are considered: dihedral angle and improper torsion potentials. The proper dihedral angle depends on the position of four consecutive bonded atoms, whereas the improper dihedral angle depends on the position of four bonded but nonconsecutive atoms.

Proper dihedral

The dihedral angle potential defines the rotation around a bond. This term is oscillatory in nature and requires the use of a sinusoidal function. The parameters included in this term are the force constant k_γ, the number of minima n, and the phase δ.



$$V^{dh} = k_{\gamma_i}^{(dh)} [1 + \cos(n\gamma_i - \delta_i)] \tag{2.10}$$

In Fig. 2.3 the proper dihedral potential for the butane molecule is shown. We observe a global minimum in $\gamma = 0$, two local minima and two local maxima respectively in $\gamma = \pm 2/3\pi$ and $\gamma = \pm 1/3\pi$

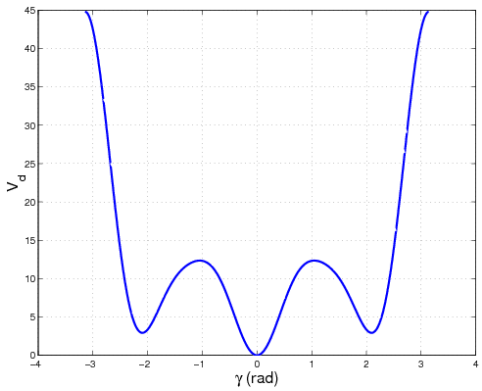
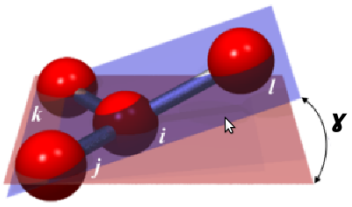


Figure 2.3: A typical behaviour of a dihedral potential. The parametrization is refered to a butane molecule where the dihedral potential is a sum of five terms like V^{dh}.

Improper dihedral

This term is included in order to keep a group in a planar or tetrahedral structure; to represent it a harmonic potential may be used, with only one minimum.

$$V^{idh} = k_{\gamma_i}^{(i-dh)}(\gamma_i - \gamma_{0i})^2 \qquad (2.11)$$



γ_i and γ_{0i} are respectively the dihedral and the equilibrium dihedral angle. The parameters γ_{0i} are associated with the 3D structure, while the set $k_{\gamma_i}^{(i-dh)}$ is characteristic of atoms involved.

Fig. 2.4 shows the structure of cyclobutane. It is known from experimental data that the correct geometry is obtained when atoms 1, 2, 3, and 5 are on one plane as on the right; but without the the improper dihedral term in the force field the system would shift to configuration on the left.

The mathematical derivation of the force for these terms is quite complicated (see the DL-Poly manual) [13].

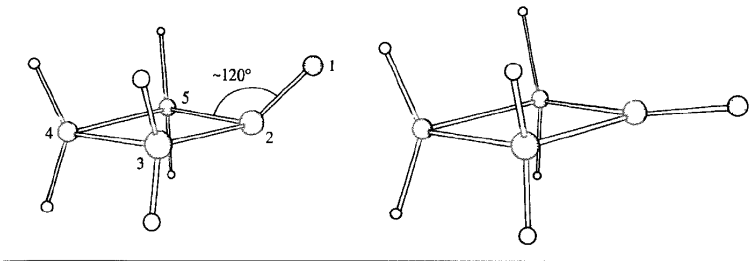
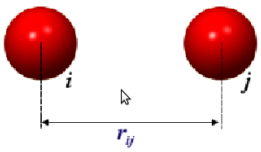


Figure 2.4: Conformation of cyclobutane corresponding to diffent values of the improper dihedral angle.

2.4 Non-bonded interactions



The non-bonded part of the potential describes interactions between atom pairs that are not covalently bonded to another; solute-solvent, solvent-solvent, solute-ions and solvent-ions interactions are included; furthermore the non-bonded term describes the interactions between distant parts of the solute. This term is very important for computational studies of biological systems because the property of macromolecules are strongly influenced by the environment and so the proper treatment of non-bonded interactions is essential for successful biomolecular computations. The mathematical expression of the function referring to these terms can be relatively simple.

2.4.1 Van der Waals interactions

The Van der Waals interactions are generated by correlations in the fluctuating polarizations of atoms; taking into account the Pauli exclusion principle, the polarization effects and the multipoles interactions, it is the macroscopic result of the quantum effects. These forces are relatively weak (see Tab. 2.1) when compared with bonded forces, but play a fundamental role in structural biology as well as in other physical and chemical fields.

The mathematical function widely used to represent these forces is the Lennard-Jones potential that fulfills the request that $v(r) \to \infty$ at small distances (Pauli exclusion principle), and $v(r) \to 0$ as $1/r^6$ for large distances (Van der Waals dispersion).

$$V^{LJ} = 4\,\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] \tag{2.12}$$

Fig. 2.5 shows the tipical behaviour of the Van der Waals potential, computed for the SPC model of water (described below). The parameters

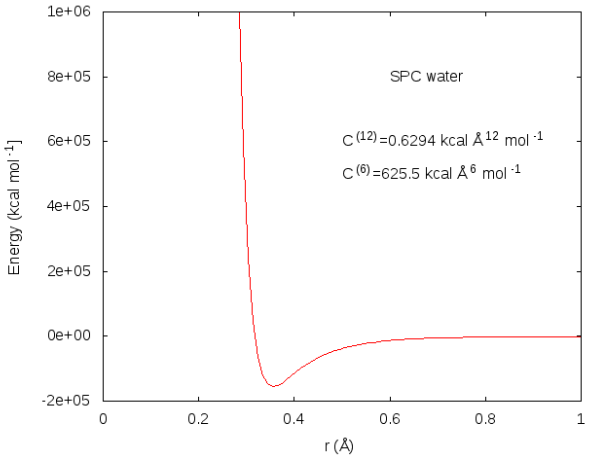


Figure 2.5: Lennard-Jones potential for SPC model of water.

needed to define this interaction are the well depth ϵ_{ij} of the potential and the minimum interaction radius σ_{ij}, defined for every pair of atoms in the system. GROMACS uses an equivalent expressions for the Lennard-Jones function, namely:

$$V^{LJ} = \left(\frac{C_{ij}^{(12)}}{r_{ij}^{12}}\right) - \left(\frac{C_{ij}^{(6)}}{r_{ij}^{6}}\right) \tag{2.13}$$

The forces acting on the particle are defined by:

$$\mathbf{f}_i = -\frac{\partial V^{LJ}}{\partial \mathbf{r}_i} = \left(12\frac{C_{ij}^{(12)}}{r_{ij}^{12}} - 6\frac{C_{ij}^{(6)}}{r_{ij}^{6}}\right)\mathbf{f}_j = -\mathbf{f}_i \tag{2.14}$$

2.4.2 Electrostatic interactions

In the study of biomolecules electrostatic interactions play a fundamental role; they are crucial in a proper modelling of conformational stability, fold-

ing, enzyme activity, binding energies, and of the interactions between a macromolecule and a surface.

The electrostatic interactions are defined by the Coulomb law:

$$V = \frac{1}{4\pi\epsilon_0}\frac{q_i q_j}{\epsilon_r r_{ij}} \tag{2.15}$$

where ϵ_0 is the electric constant and ϵ_r the relative dielectric constant. The only parameters necessary to define this function are the charges, but in spite of its simplicity the numerical treatement of this term is quite demanding, and it is necessary to use specialized algorithms. We well discuss below of treatment of this term.

As mentioned before, the equation (2.1) is adequate to treat the physical interactions that occur in biological systems, its accuracy depending on the set of the parameters used.

	kJ mol^{-1}
Covalent	200-4000
Electrostatic	10-30
Hydrogen-bonds	5-20
Van der Waals	5-10

Table 2.1: Order of magnitude of the principal interactions involved in the molecular structures.

2.5 Explicit solvent model: the water parametrization

An understanding of a wide variety of phenomena concerning biomolecules requires considering solvation effects. In this section we present and discuss the parameterization of water molecules. There are several water models for (bio)molecular simulation. They can be classified in the following way:

- rigid models with a fixed geometry;
- flexible models, including vibrational degrees of freedoms;

- polarizable models that explicitly account for polarization;

- implicit models.

The most used models for biomolecular simulation treat water as a rigid structure: various versions of TIP (Transferable Interaction Potential) and SPC (Simple Point Charge) are available. In Fig. 2.6 we show the general shape of 3-, 4-, 5-, and 6-site water models. The geometric parameters (OH distance and HOH angle) vary depending on the model.

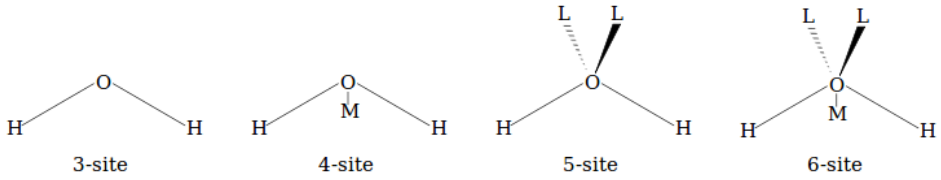


Figure 2.6: Different models of water molecule that correspond to different parametrizations.

2.5.1 TIP model

The TIP series has been developed by William Jorgensen and coworkers [14].

The TIP3P model has 3 interaction sites centered on the atomic nuclei. Positive partial charges are on the hydrogen atoms and a negative one is on the oxygen. The Lennard-Jones parameters are non-zero only for the oxygen, as in most water models. In the TIP4P model there are 4 interaction sites. The negative charge is shifted 0.015 nm off the oxygen along the bisector of the HOH angle. This model is more accurate than TIP3P, but computationally more expensive. The TIP5P model has 5 interaction sites. The negatively charged sites are located symmetrically along the lone-pair directions. This model is the best in reproducing bulk water properties, but is computationally expensive. This and more complicate models are impracticable for biomolecular simulation. There is a wide literature on these models and on their successive development.

2.5.2 SPC model

The SPC series has been developed by Herman Berendsen, Wilfred van Gunsteren and coworkers [15]. As for the TIP3P model, all SPC models have 3 interaction sites centered on the atomic nuclei, positive partial charges on the hydrogens and a negative partial charge on the oxygen; the Lennard-Jones parameters are non-zero only for the oxygen.

SPC/E: Extended SPC

This model adds an average polarization correction to the potential energy function:

$$E_{pol} = \frac{1}{2}\sum_i \frac{(\mu - \mu_0)^2}{\alpha_i}$$

where μ is the dipole of the effectively polarized water molecule (2.53 D), μ_0 is the dipole moment of an isolated water molecule (1.85 D from experiments) and α_i is an isotropic polarizability constant whose value is $1.68 \cdot 10^{-40} F\,m$. Since the charges in the model are constant this correction just results in adding 1.25 Kcal/mol (or 5.22 kJ/mol) to the total energy. The SPC/E model is better than the SPC model in reproducing density and diffusion properties [16].

	SPC	SPC/E	TIP3	TIP 4
r(OM) [Å]	-	-	-	0.15
r(OH) [Å]	1.0	1.0	0.9572	0.9572
HOH [deg]	109.47	109.47	104.52	104.52
$C^{(12)}10^{-3}$ [kcal Å^{12} mol^{-1}]	629.4	629.4	582.0	600
$C^{(6)}$ [kcal Å^{6} mol^{-1}]	625.5	625.5	595.0	610
q(O)	-.82	-.8476	-.834	-
q(H)	+.41	+.4238	+.417	+.52
q(M)	-	-	-	-1.04

Table 2.2: Parameter for the water-topology.

All MD simulation packages (e.g. GROMACS, CHARMM, AMBER, LAMMPS, GROMOS...) offer the possibility to choose the water model independently of the biomolecular force field. Explicit inclusion of water molecules provides a good representation of the kinetic and thermodynamic properties of the solute molecules. Most biomolecular simulations are made in an all-atoms approximation, using periodic boundary conditions. This yields a large number of water molecules and a considerable increas in the number of degrees of freedom of the system. For this reason and because of the smallness of the time step necessary to integrate the equations of motions (of the order of a femtosecond), the best algorithms can simulate events in the range of 10^{-9} s to 10^{-8} s for typical proteins, and 10^{-6} s for very small proteins. Many biological processes occur on a larger time scale, therefore much work is aimed at developing alternative models and computationally less expensive approaches.

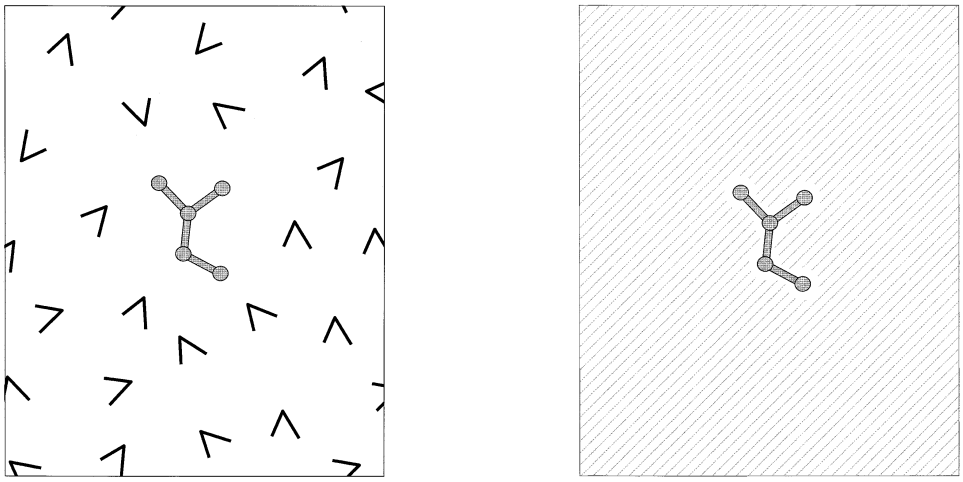
2.6 Implicit solvent model

In an all-atoms representation of solvated biomoleculs, a large fraction of the overall computational time is used to calculate the detailed trajectories of the solvent molecules. An alternative approach consists in incorporating implicitly the effect of the solvent. The use of a continuum model of the solvent greatly decreases the number of degrees of freedom in the system, and consequently the computing time. Fig. 2.7(a) shows schematically a molecule surrounded by explicit water molecules; Fig. 2.7(b) represents the same biological system but in a medium field that implicitly incorporates the influence of the solvent. This approximation can provide useful quantitative estimates of solvation free energies.

2.6.1 Potential of mean force

In this section we see the statistical approach to the mathematical formulation of the implicit solvent model, and we introduce the classical electrostatic equation (CE).

Let us consider a molecule α immersed in a bulk solution β. $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2 ... \boldsymbol{x}_N)$ is the molecule's configuration vector and in $\boldsymbol{Y}$ there are all the other degrees of freedom (solvent and ions, if present in solution). The system fluctuates over a large number of configurations,



(a) (b)

Figure 2.7: Schematical rapresentation of a biomolecule surrounded by explicit water molecules [panel (a)], and in a medium field that implicitly incorporates the influence of the solvent [panel (b)].

and so one uses a statistical approach to define the probability of a given configuration:

$$P(\boldsymbol{X},\boldsymbol{Y}) = \frac{\exp\big[-\beta U(\boldsymbol{X},\boldsymbol{Y})\big]}{\int d\boldsymbol{X}d\boldsymbol{Y}\exp\big[-\beta U(\boldsymbol{X},\boldsymbol{Y})\big]} \tag{2.16}$$

with $\beta = 1/\mathrm{k}_B\mathrm{T}$, k_B Boltzmann's constant and T the equilibrium temperature of the system.

The total potential energy of the system $U(\boldsymbol{X},\boldsymbol{Y})$ can be separated into the following terms:

$$U(\boldsymbol{X},\boldsymbol{Y}) = U_\alpha(\boldsymbol{X}) + U_\beta(\boldsymbol{Y}) + U_{\alpha\beta}(\boldsymbol{X},\boldsymbol{Y}) \tag{2.17}$$

where:

- $U_\alpha(\boldsymbol{X})$ is the intramolecular potential of the solute;
- $U_\beta(\boldsymbol{Y})$ is the potential representing the solvent-solvent interaction;
- $U_{\alpha\beta}(\boldsymbol{X},\boldsymbol{Y})$ is the solute-solvent interaction.

Using $P(\boldsymbol{X},\boldsymbol{Y})$ one can calculate every observable of the system as an expected value:

$$\langle\Theta\rangle = \int d\boldsymbol{X}d\boldsymbol{Y}\Theta(\boldsymbol{X},\boldsymbol{Y})P(\boldsymbol{X},\boldsymbol{Y}) \tag{2.18}$$

If one integrates the function $P(\boldsymbol{X}, \boldsymbol{Y})$ over the $\boldsymbol{Y}$ coordinate, one can define the reduced probability distribution, depending only on the solute configurations:

$$\overline{P}(\boldsymbol{X}) = \int d\boldsymbol{Y} P(\boldsymbol{X}, \boldsymbol{Y}); \tag{2.19}$$

by introducing (2.16) and (2.17) into (2.19) we have:

$$\overline{P}(\boldsymbol{X}) = \frac{\int d\boldsymbol{Y} \exp\left\{-\beta\Big[U_\alpha(\boldsymbol{X}) + U_\beta(\boldsymbol{Y}) + U_{\alpha\beta}(\boldsymbol{X}, \boldsymbol{Y})\Big]\right\}}{\int d\boldsymbol{X} d\boldsymbol{Y} \exp\left\{-\beta\Big[U_\alpha(\boldsymbol{X}) + U_\beta(\boldsymbol{Y}) + U_{\alpha\beta}(\boldsymbol{X}, \boldsymbol{Y})\Big]\right\}} \tag{2.20}$$

If one defines a function $W(\boldsymbol{X})$ such that:

$$\exp\{-\beta W(\boldsymbol{X})\} = \int d\boldsymbol{Y} \exp -\beta U(\boldsymbol{X}, \boldsymbol{Y}) \tag{2.21}$$

one can write (2.20) in the canonical form for a system in equilibrium at temperature T:

$$\overline{P}(\boldsymbol{X}) = \frac{\exp\big[-\beta W(\boldsymbol{X})\big]}{\int d\boldsymbol{X} \exp\big[-\beta W(\boldsymbol{X})\big]}. \tag{2.22}$$

The function $W(\boldsymbol{X})$ is an effective configuration-dependent free energy, and is called the potential of mean force (PMF); it is a function only of solute configurations, and its gradient is related to the average force:

$$\frac{\partial W(\boldsymbol{X})}{\partial \boldsymbol{x}_i} = \langle \frac{\partial U}{\partial \boldsymbol{x}_i} \rangle_{\boldsymbol{X}} = -\langle \boldsymbol{F}_i \rangle_{\boldsymbol{X}} \tag{2.23}$$

where $\langle \cdots \rangle_{\boldsymbol{X}}$ is the average over all solvent coordinates, with the solute in a fixed configuration specified by $\boldsymbol{X}$. All solvent effects are included in $W(\boldsymbol{X})$ and consequently in the reduced probability distribution $\overline{P}(\boldsymbol{X})$.

It is instructive to calculate from the equation (2.23) the variation of $W(\boldsymbol{X})$ between two different solute configurations:

$$\begin{aligned} W(\boldsymbol{X}_2) &= W(\boldsymbol{X}_1) + \int_{\boldsymbol{X}_1}^{\boldsymbol{X}_2} \sum_i d\boldsymbol{x}_i \cdot \frac{\partial W(\boldsymbol{X})}{\partial \boldsymbol{x_i}} \\ &= W(\boldsymbol{X}_1) - \int_{\boldsymbol{X}_1}^{\boldsymbol{X}_2} \sum_i d\boldsymbol{x}_i \cdot \langle \boldsymbol{F}_i \rangle_{\boldsymbol{X}} \end{aligned} \tag{2.24}$$

This relationship clearly shows that the PMF in not simply an average potential energy but represents the reversible work performed by the solute molecule against the average solvent force, and is therefore defined up to an additive constant. Usually, it is rescaled by the solvent-solvent interaction, which still satisfies the normalization condition (2.22) :

$$\exp\{-\beta W(\boldsymbol{X})\} = \frac{\int d\boldsymbol{Y} \exp -\beta U(\boldsymbol{X}, \boldsymbol{Y})}{\int d\boldsymbol{Y} \exp -\beta U_\beta(\boldsymbol{Y})} \tag{2.25}$$

In $W(\boldsymbol{X})$ we have the intramolecular solute potential contribution $U_\alpha(\boldsymbol{X})$, the solute-solute and solute-solvent interactions. The contribution of the last two terms is due to: (a) a short-range repulsive interaction arising from Pauli's exclusion principle; (b) the Van der Waals attractive force arising from quantum dispersion; (c) long-range electrostatic interactions arising from a non-uniform charge distribution; (a) and (b) are usually called non-polar interactions. One can consider the following splitting:

$$U_{\alpha\beta}(\boldsymbol{X}, \boldsymbol{Y}) = U^{(np)}_{\alpha\beta}(\boldsymbol{X}, \boldsymbol{Y}) + U^{(elec)}_{\alpha\beta}(\boldsymbol{X}, \boldsymbol{Y}) \tag{2.26}$$

that produces a similar separation in the free energy, and is commonly used in biomolecular force fields (e.g. AMBER [17], CHARMM [18] OPLS [19]). This separation leads usally to the following formulation:

$$W(\boldsymbol{X}) = U_\alpha(\boldsymbol{X}) + \Delta W^{(np)}(\boldsymbol{X}) + \Delta W^{(elec)}(\boldsymbol{X}) \tag{2.27}$$

Many methods developed for the simulation of biomolecules use the Solvent Accessible Surface Area (SASA) [20] to compute the non-polar free energy contribution $\Delta W^{(np)}(\boldsymbol{X})$. The Poisson-Boltzmann (PB) equation is usually used to compute the electrostatic contribution $\Delta W^{(elec)}(\boldsymbol{X})$. The combination of SASA and PB defines the implicit solvent approach.

2.6.2 Non-polar free-energy contribution

The term $\Delta W^{(np)}$ is often defined *the free energy of cavity formation* because the principal contribution is the work necessary to rearrange the solvent molecules around the solute to minimize the hydrofobic force. A second, small contribution is due to the Van der Waals interactions. Computer simulation studies [21] attribute the hydrofobic effects primarily to a decrease in the number of hydrogen bonds among water molecules near a nonpolar

surface. Therefore, in a first approximation, the non-polar free energy can be considered proportional to the number of solvent molecules in the first solvation shell. One assumes that the non-polar free energy contribution is directly related to the SASA

$$\Delta W^{(np)}(\boldsymbol{X}) = \gamma \mathcal{A}_{tot}(\boldsymbol{X}); \tag{2.28}$$

γ has the dimension of a surface tension and its value is assigned by matching experimental data. $\mathcal{A}_{tot}(\boldsymbol{X})$ is the configuration dependent SASA. The limitations of this model are well studied in the literature [22].

2.6.3 Electrostatic free-energy contribution

In order to describe the electrostatic free energy contribution, it is useful to introduce a parameter λ so that if $\lambda = 0$ the interactions solute-solvent are absent, and if $\lambda = 1$ the full set of interactions is taken into account. The free energy function has a particularly simple form assuming λ as a scaling factor of the solute charge (i.e. $U^{(elec)}_{\alpha\beta}(\boldsymbol{X}, \boldsymbol{Y}; \lambda) = \lambda U^{(elec)}_{\alpha\beta}(\boldsymbol{X}, \boldsymbol{Y})$). One has:

$$\Delta W^{(elec)}(\boldsymbol{X}) = \int_0^1 d\lambda \langle U^{(elec)}_{\alpha\beta} \rangle_\lambda. \tag{2.29}$$

$\langle U^{(elec)}_{\alpha\beta} \rangle_\lambda$ is proportional to λ because the interaction energy of the solvent is proportional to the charge of solute:

$$\Delta W^{(elec)}(\boldsymbol{X}) = \int_0^1 d\lambda \sum_i q_i \phi_{rf}(\boldsymbol{x}_i; \lambda) \approx \frac{1}{2} \sum_i q_i \phi_{rf}(\boldsymbol{x}_i, \lambda = 1) \tag{2.30}$$

$\phi_{rf}(\boldsymbol{x}_i, \lambda = 1)$ is the field acting on the *ith* solute atomic charge at the position $\boldsymbol{x}_i$: it is called *reaction field* and represents the electrostatic potential exerted on the solute by the polarized solvent [23].

2.7 Application of the electrostatic continuum to biological systems

In order to calculate the reaction field and from this the electrostatic contribution to the solvation free energy, one can use a classical electrostatic continuum representation. The first model where the polar solvent was represented by a continuum dielectric medium, was devised by Born in 1920 to

calculate the hydration free energy of spherical ions [24]. Later Kirkwood [25] and Onsager [26] extended Born's model to study an arbitrary charge distribution inside a spherical cavity.

In this approximation, to study a macromolecule in a solvent one has to solve the Poisson equation to find the reaction field:

$$\nabla \cdot [\epsilon(\boldsymbol{x}) \nabla \phi(\boldsymbol{x})] = -4\pi \rho_\alpha(\boldsymbol{x}), \tag{2.31}$$

where $\phi(\boldsymbol{x}), \rho_\alpha(\boldsymbol{x})$ and $\epsilon(\boldsymbol{x})$ are respectively the electrostatic potential, the charge density of the solute, and the position dependent dielectric constant. $\epsilon(\boldsymbol{x})$ reflects the reorientation of permanent and induced dipoles under the local electric field. Permanent dipoles occur when the distribution of charge over neighboring atoms is not symmetric: typical examples are peptide bonds and water molecules. In liquid water, the relative freedom of the molecules allows a high dipolar rotation and consequently one finds a high dielectric constant (78.5 at 298 K). In contrast, permanent dipoles in a macromolecule can be considered fixed, and the dielectric constant is much smaller. Experimental and theoretical studies suggest that the average dielectric response of such a solute macromolecule can be approximated with a dipole in the range 2-4; furthermore it has been shown that the use of a single dielectric constant appears to be a reasonable approximation to account for the electronic polarization response of the entire macromolecule. On the other hand, induced dipoles arising from electronic polarization, i.e., from the distortion of electronic clouds immersed in an electric field, give a small contribution of electronic polarization (~ 4).

In Fig. 2.8 a scheme of a protein immersed in a continuum medium with a specific dielectric constant is reported. The Poisson equation (2.31) can be numerically solved by mapping the system on a discrete grid. Several programs are available to compute the electrostatic potential using this approach, for example the PBEQ program [27] which is part of the simulation program CHARMM. The results depend sensitively on the atomic radii and on the atomic partial charges assigned to the atoms at the solvent exclusion-boundary. The boundary can be constructed using the SASA model. The optimal radius of an atom is not a property of the isolated atom, but is an effective empirical parameter depending on its charge, on its neighbors and also on the nature of the solvent molecules: some algorithms have been developed and are available in the simulation packages. The partial background charges of the solute are specified in the topology of the system, and

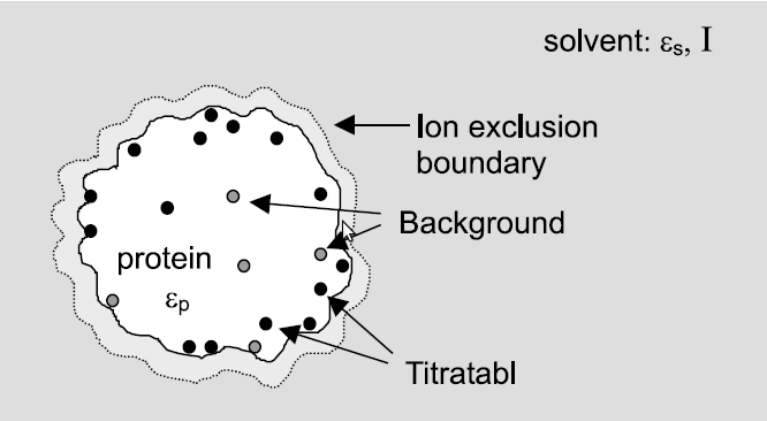


Figure 2.8: Schematic representation of the charge distribution in a protein immersed in a continuum medium; the two groups, the background charges and the titratable charges, are represented.

implemented in the definition of the force field of a specific system.

When in solution there is a ionic concentration (counterions), one can use the equation (2.31) modified according to the Debye-Hückel theory of electrolytes [28]:

$$\nabla \cdot [\epsilon(\boldsymbol{x})\nabla\phi(\boldsymbol{x})] - \kappa^2(\boldsymbol{x})\epsilon(\boldsymbol{x})\phi(\boldsymbol{x}) + 4\pi\rho_\alpha(\boldsymbol{x}) = 0 \tag{2.32}$$

where the charge density $\rho_\alpha(\boldsymbol{x})$ refers only to the protein charges and the ionic effect is totally contained in the second term of the equation. $\kappa(\boldsymbol{x})$, called reciprocal Debye length, is a two-valued function; it assumes the value

$$\kappa = \left(\frac{8\pi e^2 N_A I}{\epsilon\ \mathtt{k}_B\mathtt{T}}\right)^{1/2} \tag{2.33}$$

if a point in space is ion-accessible, and zero otherwise. N_A is the Avogadro number, e is the proton charge, k_B is the Boltzmann constant and ϵ the solvent dielectric constant. The ionic strength I of the solution is defined as:

$$I = \frac{1}{2}\sum_i c_i z_i^2 \tag{2.34}$$

with the sum running over all ionic species in solution, each with its charge z_i; c_i is ionic bulk concentration.

2.7.1 pH and solute charge

With the simple model in Fig. 2.8 we see that the charges of the protein can be divided into two groups, the background charges and the titratable charges. The former are independent of the protonation state of the molecule; the latter represent the charge of the ionic (titratable) residues (Asp, Glu, Lys, Arg, His, Tyr, free Cys, N- and C-terminal), and are generally pH dependent as a consequence of the protonation/deprotonation reactions.

The pH of the solution is an important parameter, strictly correlated to the electrostatic field. At first glance it would seem a trivial matter to define the charge state of a given group by considering the Henderson-Hasselbach equation of the acid-base equilibrium [29]:

$$pK_a = pH - \log \frac{f}{1-f}; \tag{2.35}$$

f is the degree of protonation, i.e. the fraction of molecules that is protonated; pK_a is the pH value at which half of the molecules is protonated: its values for the free titratable residues are well known. However, the situation is more complicated because of the other charged sites: the local environment in the protein may shift by several pH units the pK_a of a given site from its value typical of the free state. If one defines:

- pK^{free}: pK_a of a titratable residue in its free state;
- pK^{int}: pK_a of a titratable residue in the protein taking the other groups as neutral;
- pK^{eff}: pK_a of a titratable residue in the protein taking into account all the charges in the protein;

one finds the following relations [30]:

$$pK^{int} = pK^{free} + \frac{1}{2.3K_bT}\Delta G^{env} \tag{2.36}$$

where ΔG^{env} is the free energy change when moving the residue from water into the neutral protein;

$$pK^{eff} = pK^{int} + \frac{1}{2.3K_bT}\Delta G^{int} \tag{2.37}$$

ΔG^{int} is the free energy due to the electrostatic contribution of the other charged residues.

The determination of the charge of the titratable residues as a function of the solution pH is a complicated calculation, and some software has been developed for this purpose, e.g. DelPhy (see [31]).

2.8 A simple example: the Born model

For its historical importance and for its simplicity in the following we show the simple Born model, that estimates the free energy of a ion in water. The ion is modelled with a sphere of radius a centered in the origin characterized by a dielectric constant $\epsilon = 1$, whereas the water is modelled by a continuum with a large dielectric constant ϵ.

At any point $\boldsymbol{r}$, with $r > a$, for the electric field and for the electrostatic potential we have respectively:

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{q}{\epsilon r^2}\frac{\boldsymbol{r}}{r} \qquad \phi(\boldsymbol{r}) = \frac{q}{\epsilon r} \tag{2.38}$$

Inside the sphere, for $r \neq 0$, the electric field is given by

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{q}{r^2}\frac{\boldsymbol{r}}{r}\ . \tag{2.39}$$

The electrostatic potential at $\boldsymbol{r}$ is given by its value on the surface of the ionic sphere plus the work to move, against the field, the charge inside:

$$\phi(\boldsymbol{r}) = \frac{q}{\epsilon a} + \frac{q}{r} - \frac{q}{a}\ . \tag{2.40}$$

If we remove the work required to move the ion from infinity to $\boldsymbol{r}$ in vacuum electrostatic field, we have the work performed by the dielectric medium in moving the ion:

$$\phi(\boldsymbol{r}) = -\left(1 - \frac{1}{\epsilon}\right)\frac{q}{a} \tag{2.41}$$

replacing in equation (2.30) $\Delta W^{(elec)}(\boldsymbol{X}) \approx \frac{1}{2}\sum_i q_i\phi_{rf}(\boldsymbol{x}_i, \lambda = 1)$ we arrive to the Born solvation free energy:

$$\Delta W^{(elec)} = -\frac{1}{2}\left(1 - \frac{1}{\epsilon}\right)\frac{q^2}{a} \tag{2.42}$$

2.9 Generalized Born Model

In order to take into account the conformational flexibility in the energy minimization of various initial conformations, and in the dynamics, we have to introduce the analytical first derivative of the solvation free energy with respect to the atomic coordinates of the solute. But this calculation is computationally too expensive for large systems. For this reason, it is usual to introduce an approximation to the exact continuum electrostatics, using a semianalytical function expressed as superposition of pairwise additive terms. One of the most popular approximations is the Generalized Born (GB) model [32].

The GB function for the solvation free energy is:

$$\Delta G^{pol} = -\frac{1}{2}\left(1 - \frac{e^{-\kappa f_{GB}}}{\epsilon}\right)\sum_{ij}\frac{q_i q_j}{f_{GB}} \tag{2.43}$$

q_i and q_j are atomic partial charges and the double sum runs over all pairs of atoms, ϵ is the solvent dielectric constant, and κ is the Debye-Hückel screening parameter taking the salt concentration into account. f_{GB} is a function that interpolates between an effective Born radius α_i, when the distance r_{ij} between atoms is short, and r_{ij} itself at large distances:

$$f_{GB} = [r_{ij}^2 + \alpha_i \alpha_j \exp\left(-r_{ij}^2/4\alpha_i\alpha_j\right)] \tag{2.44}$$

As we have seen, the effective Born radius α_i describes how deeply a charge of a biological system is placed in the low-dielectric medium. It depends not only on the intrinsic radius ρ_i of atom i, but also on the relative positions and intrinsic radii of all other atoms. Several algorithms are available in the biomolecular software to calculate Born radii, e.g. STILL [33], HTC [34], OBC [35].

2.10 Biomolecular Force Fields

Central to the success of any computational approach is the quality of the model used; therefore much work has focused on the improvement in the available force fields and on the development of new ones. In this section

we present a list of the most popular force fields used in biomolecular simulations. Each force field has strengths and weaknesses; there isn't a best one, but one can choose the better one for a particular case. Many studies compare the properties of different force fields [36–38].

AMBER (Assisted Model Building with Energy Refinement)

http://amber.scripps.edu
has been developed since the early 80's under the leadership of Peter Kollman at University of California at San Francisco. In Table 2.3 we report the major versions.

CHARMM (Chemistry at Harvard Molecular Mechanics)

http://charmm.org
has been developed since the early 80's under the leadership of Martin Karplus at Harvard University. In Table 2.4 we report the major versions.

GROMOS (Groningen Molecular Simulation)

http://www.igc.ethz.ch/gromos
has been developed since the early 80's under the leadership of Wilfred van Gunsteren, ETH Zurich. In Table 2.5 we can see the major versions.

OPLS (Optimized Potentials for Liquid Simulation)

http://zarbi.chem.yale.edu
has been developed since the early 80's under the leadership of William Jorgensen, Yale University. In Table 2.6 the major versions are listed..

Version	Principal characteristic
ff86	United-atom and all atom variants; fixed partial charges centered on atoms.
ff94	Reparametrization, all-atoms force field and fixed partial charges centered on atoms. Charges based on multipole-conformation calculations and a Restrained Electrostatic Potential fit.
ff96	All-atoms force field. Fixed partial charges. Modification of backbone ϕ, ψ torsional parameters based on *ab-initio* calculations for alanine tetrapeptide.
ff99	All-atoms force field. Fixed partial charges. Minor changes on protein parameters.
ff02	Polarizable variant of ff99. Polarizable dipoles at the atoms, which can be calculated iteratively at each step or propagated with the atomic positions as additional dynamical variables. Two variant, with centered point charges and with additional point charges.
ff03	All-atoms force field. Fixed partial charges centered on atoms. Derived from ff99, with charges obtained from *ab-initio* calculations with a continuum dielectric to mimic solvent polarization and new backbone ϕ, ψ torsional parameters.

Table 2.3: Major versions of AMBER force field.

Version	Principal characteristic
param19	United atom force field. Fixed partial charges centered on atoms.
param22	Reparametrization of param19. Charges based on *ab-initio* dimer energies and geometries. Newer CHARMM parameter sets do not include changes in protein parameters.
CHARMM fluctuating-charge	Polarizable force field, based on fluctuating-charge model i.e. the partial atomic charges in a molecule are allowed to redistribute to yield equivalent electronegativity on each atom.
CHARMM Drude	Polarizable force field based on the Shell or Drude model, (i.e. introduction of "massless" virtual sites /Drude particles) carrying partial electronic charge and attached to individual atoms via a harmonic spring.

Table 2.4: Major versions of CHARMM force field.

Version	Principal characteristic
37C4	United-atoms force field. Fixed partial charges centered on atoms.
43A1	Reparametrization of 37C4 using the liquid properties and hydration free energies.
53A6	United-atoms force field (implicit aliphatic H). Fixed partial charges centered on atoms. Reppresentation of polar groups based on liquid properties and of side chains based on solvation free energies.

Table 2.5: Major versions of GROMOS force field.

Version	Principal characteristic
OPLS	United-atoms force field. Fixed partial charges centered on atoms.
OPLS-AA1	Reparametrization of OPLS using liquid properties and hydration free energies.
OPLS-AA1	Major reparametrization with special emphasis on torsional parameters.

Table 2.6: Major versions of OPLS force field.

Chapter 3

Simulation methods

Interest in the dynamics of biomulecular systems derives from its relevance in issues such as folding and unfolding of proteins, the role of dynamics in biological functions, the interaction between proteins or nucleic acids and other systems. Molecular Dynamics (MD) is a useful tool to study many such problems. Using a selected force field, Newton's equations have to be solved to obtain coordinates and momenta of all atoms of the system along the time trajectory:

$$m_i \ddot{\boldsymbol{r}}_i = \boldsymbol{f}_i \quad i = 1, 2 \ldots N \tag{3.1}$$

$$\ddot{\boldsymbol{r}}_i = \frac{d^2 \boldsymbol{r}_i}{dt^2} \quad \boldsymbol{f}_i = -\frac{\partial V(\boldsymbol{r})}{\partial \boldsymbol{r}_i} \tag{3.2}$$

$\boldsymbol{r}_i = (x_i, y_i, z_i)$, $\ddot{\boldsymbol{r}}_i$ and $\boldsymbol{f}_i$ are respectively the Cartesian coordinates and the corresponding acceleration of the $i - th$ atom, and the force acting on it. Newton's law of motion is a second order differential equation that requires two initial values for each degree of freedom to be numerically integrated. To start a simulation one need also of a molecular description (or topology) of the system to be simulated containing information on the system, e.g. which atoms are covalently bonded and other physical information. In Fig. 3.1 there is a simple scheme of a MD computer simulation.

3.1 Initial coordinates

The 3D structures are usually obtained from spectroscopic experiments (e.g, X-ray crystallography, Nuclear magnetic resonance NMR, Electron microscopy EM) [39]. Alternatively, one can use the homology protein

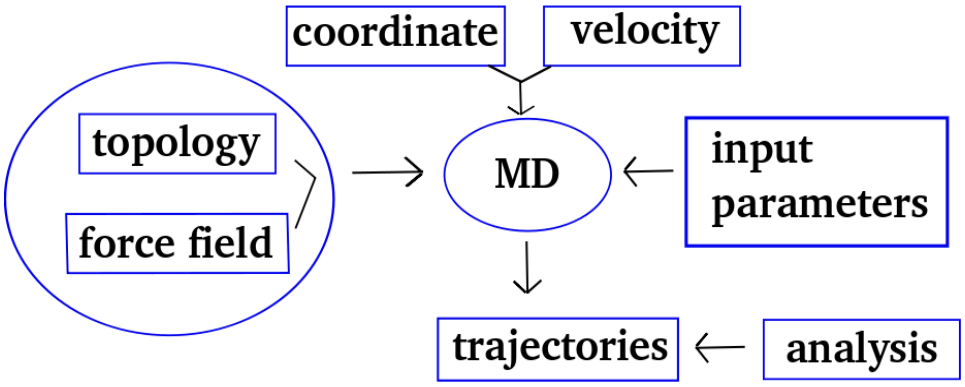


Figure 3.1: The global MD procedure.

structure modelling [41]. Homology modelling, also known as comparative modelling of protein, refers to constructing an atomic-resolution model of the *target* protein from its amino acid sequence and an experimental three-dimensional structure of a related homologous proteins (the *templates*). Homology modelling relies on the identification of one or more known protein structures likely to resemble the structure of the target sequence, and on the production of an alignment that maps residues in the target sequence to residues in the template sequence. It has been shown that protein structures are more conserved than protein sequences amongst homologues, but sequences falling below a 20% sequence identity can have very different structure. We can describe the homology technique with four simple steps, repeated if needed (Fig. 3.2). For the unknown system (target sequence) we have to:

1. identify related structures with a 3D known structure (template structures);

2. align the target to the template sequence;

3. using information from the template structures to build a model for the target sequence;

4. evaluate the model comparing the information obtained from this model with the experimental results.

The spectroscopic experimental techniques are used to produce the 3D structures, but also to investigate the results of a computer simulation in order to validate the model.

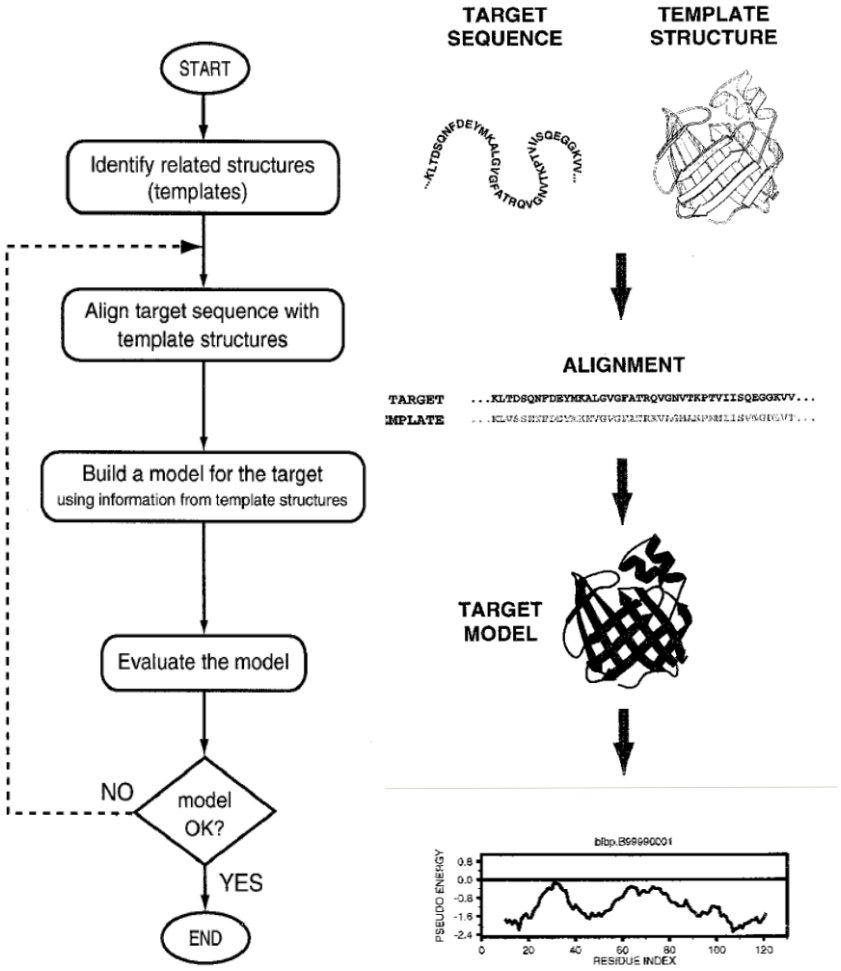


Figure 3.2: The flow for comparative protein structure modelling. The figure has been taken from Ref. [40].

3.2 Initial velocities

The velocities are usually taken at random from a standard Maxwellian velocity distribution. For a system in equilibrium at temperature T, one has for a given component of the velocity:

$$P(v)dv = \left(\frac{m}{2\pi \mathrm{k}_B \mathrm{T}}\right)^{1/2} \exp\left(-\frac{mv^2}{2\mathrm{k}_B \mathrm{T}}\right) dv. \tag{3.3}$$

In order to avoid a thermal schock of the system, one usually starts with a low temperature and increases it gradually by scaling the velocities, allowing the system to relax. This slow heating continues until the simulation reaches the desired temperature.

According to the equipartition theorem, the temperature $\mathrm{T}(t)$ is defined by:

$$\mathrm{T}(t) = \frac{1}{\mathrm{k}_B N_{dof}} \sum_{i=1}^{N_{dof}} m_i |v_i|^2 \tag{3.4}$$

where N_{dof} is the number of unconstrained degree of freedom ($N_{dof} = 3N - n$, N is the number of atoms and n is the number of constraints).

3.3 Time step

The most common integration algorithms used in the MD simulation package (e.g. AMBER, CHARMM, GROMACS...) are Verlet and Leap-frog [42].

A very important parameter is the integration time step Δt: a smaller Δt produces greater accuracy but on the other hand is computationally more expensive; a good choice of the time step yields a good balance between economy and accuracy. The choice of Δt is correlated to the problem under examination: different types of analysis require different values of the time step. In Tab. 3.1 we give the reference value for some properties.

When the oscillatory frequency of a bond is greater than $\mathrm{k}_B\mathrm{T}/\hbar$ ($\hbar = h/2\pi$ Planck's constant)[1] it is necessary to treat the motion with quantum mechanics and this is true for the vibration of covalent bonds at room temperature (see Tab. 3.1). If one is not interested in observables related to the stretching of the chemical bonds, it is customary to fix them to their equilibrium values, using geometrical constraints to keep their length constant. This procedure has two related advantages: it removes the frequencies that

[1] $\mathrm{k}_B\mathrm{T}/\hbar \sim 4 \times 10^{13} Hz$ which correspond to a relaxation time of 25 fs.

Force	Relaxation time (fs)	Suitable time step (fs)
Hight frequency: vibration of covalent bonds	10	0.5
Medium frequency: angle deformation, dihedral angle torsion, Van der Waals and Coulomb short range interactions	40	2
Low frequency: Coulomb long range interaction	1000	20

Table 3.1: Reference values of time steps in different range of frequency.

can not be processed classically, and allows the use of a larger time step with a typical speed up of a factor of ~ 4.
A number of algorithms have been developed to implement geometrical constraints, e.g. SHAKE [43], or LINCS [44].

3.4 Thermodynamic Boundary Conditions

A direct solution of Newton's equations yields trajectories typical of the microcanonical ensemble NVE (where number of atoms, volume and energy of the system are costants), but other algorithms have been developed to generate trajectories in NVT or NPT ensemble (where number of atoms, volume and temperature, or number of atoms, pressure and temperature of the system are kept constant). The latter algorithms are used in particular in biology, where the systems are naturally coupled with a thermal bath.

The principal algorithms available to produce a constant temperature are the Berendsen one [45] that produces a weak coupling, and the Nosé-Hoover one [46, 47], that defines an extended Hamiltonian system; Berendsen's barostat [45] and Parrinello-Rahman's barostat [48] can be used to generate NPT trajectories.

3.5 Long-range and short-range interactions

As discussed in the previous chapter, a typical potential energy function has the form:

$$E = \underbrace{E_b + E_\theta + E_\gamma}_{\text{bonded}} + \underbrace{E_{VdW} + E_{el}}_{\text{nonbonded}} \tag{3.5}$$

where E is the total molecular energy, E_b and E_θ are harmonic terms describing bond and angle vibrations, and E_γ describes the torsion energy (we include in the E_γ the proper and the improper dihedral terms); E_{VdW} and E_{el} are nonbonded terms that describe interactions between atom pairs that are not part of a common bond, valence or torsion angle. E_{VdW} takes into account dispersion and repulsion terms, whereas E_{el} is the Coulomb interaction.

The computer time required to calculate the potential energy of a particular conformation in a large system is dominated by the calculations of the nonbonded interactions. This is due to the number of nonbonded pairs, which is much larger than the number of terms involved in the bond, angle and torsion interactions. In a system of N atoms (10^4 is a typical number of atoms in a biomolecule) there are about N bond terms and roughly the same number of angle and torsion terms; by contrast, there are $\frac{N(N-1)}{2}$ nonbonded pairs: a straightforward calculation is too expensive.

The functional form (see sec. 2.4.1) of E_{VdW} shows that this term describes *short-range* interactions. Short-range means that the total potential energy of a given particle i is dominated by interactions with neighboring particles that are closer than some cutoff distance r_c, and that the error that results when we ignore interactions with particles at a larger distance can be made arbitrarily small by choosing r_c sufficiently large. If the interactions decay rapidly enough, one can evaluate the error by the following expression [42]:

$$U^{tot} = \sum_{i<j} u_c(r_{ij}) + \frac{N\rho}{2} \int_{r_c}^{\infty} dr\ u(r)\, 4\pi r^2 \tag{3.6}$$

where u_c is the truncated potential energy function, ρ is the average number density, and we have assumed that the radial distribution function $g(r) = 1$ for $r > r_c$.

The introduction of a cutoff produces a negligible error if the potential goes to zero at least as $1/r^3$. This is the case for the E_{VdW} term, but not for the electrostatic interaction where a cutoff approximation is quite

inaccurate. Auffinger and Beveridge discussed the drawback of using simple cutoff in electrostatics [49].

The electrostatic interaction is a very important issue when studying biomolecular systems: (i) from the physical point of view, because there is an increasing evidence that the electrostatic interaction plays a relevant role in folding, conformational stability, enzyme activity, and binding energies as well as in protein-protein interactions; (ii) from the computational point of view, because the evaluation of its contribution to the total energy of the system is so expensive that it is necessary to develop specific approximate algorithms.

In most force fields the atoms of the system are parametrized using partial charges on the atomic sites. When the charges q_1, $q_2 \dots q_N$ are at positions $\boldsymbol{r}_1, \boldsymbol{r}_2 \dots \boldsymbol{r}_N$, the electrostatic energy due to the whole system of charges is given by the Coulomb's equation:

$$U = \frac{1}{2} \sum_{i=1}^{N} q_i \phi(\boldsymbol{r}_i) \tag{3.7}$$

where:

$$\phi(\boldsymbol{r}_i) = \sum_{j \neq i} \frac{q_j}{r_{ij}} \qquad\qquad r_{ij} = |\boldsymbol{r}_i - \boldsymbol{r}_j| \tag{3.8}$$

If instead of discrete charges, the charge is described by a smooth charge density $\rho(\boldsymbol{x})$, the the main equation used to model electrostatic interactions is the Poisson's equation, given by:

$$\nabla \cdot [\epsilon(\boldsymbol{x}) \nabla \phi(\boldsymbol{x})] = -4\pi\rho(\boldsymbol{x}) \tag{3.9}$$

In a computer experiment the electostatic interactions are treated solving (3.7) or (3.9) depending on the required accuracy and on the boundary condition.

3.5.1 Continuum electrostatics

In continuum boundary conditions the macromolecule is considered as a low-dielectric region carrying a fixed charge distribution and surrounded by a continuum high-dielectric medium representing the solvent. The molecule's internal forces are described by a standard force field including the Coulombic interactions, while the solution of Poisson's equation determines the solvation free energy.

3.5.2 Discrete and continuum electrostatics

Approximations based on continuum electrostatics, in which the solvent is represented as a featureless dielectric material, are remarkably successful in representing the electrostatic contribution to the solvation free energy. Nevertheless sometimes, depending of the problem, a description in which the structural details of the solvent molecules are ignored may not be appropriate; e.g. the structure of water is important in the folding process and the creation of secondary structures.

In order to obtain by computer simulation of a finite cluster a statistics similar to that of an infinite system, Belou and Roux have developed an intermediate approach: here one takes into account the solute and a small number of explicit solvent molecules in the vicinity of the solute (a layer of solvent), and represents the influence of the bulk with an effective solvent potential. This approximation follows from a formal separation of the multidimensional configurational integral in the solvent molecules nearest to the solute and the remaining ones.

Even in this approach, in which the number of explicit molecules is very reduced, the evaluation of the electrostatic interaction by means of the Coulomb potential is computationally too expensive, and an approximation is necessary. A first step to solve this problem was the use of the multiple expansion series. To evaluate the Coulomb energy at $\boldsymbol{r}$, let us consider N charges $q^1, q^2 \ldots q^N$ at the positions $\mathbf{r}^{(k)} = \{r_\alpha^{(k)}\}$ $(\alpha = x, y, z)$ with $k = 1, 2 \ldots N$, close to the point $\boldsymbol{b} = \{b_\alpha\}$, so that their distances $|\boldsymbol{r}^{(k)} - \boldsymbol{b}|$ are small compared to $|\boldsymbol{b} - \boldsymbol{r}|$ (see Fig. 3.3).

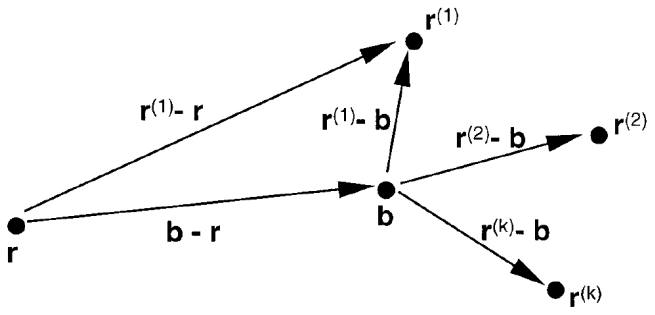


Figure 3.3: Multipole expansion of the potential at $\boldsymbol{r}$ due to charges near $\boldsymbol{b}$. Assume that $|\boldsymbol{r}^{(k)} - \boldsymbol{b}| < |\boldsymbol{b} - \boldsymbol{r}|$ for all k.

In the second-order approximation (quadrupole approximation) we have:

$$\sum_{i=1}^{N} \frac{q^{(i)}}{|\boldsymbol{r}^{(i)} - \boldsymbol{r}|} \simeq \frac{Q}{|\boldsymbol{b} - \boldsymbol{r}|} - \sum_{\alpha=1}^{3} \frac{d_\alpha (b_\alpha - r_\alpha)}{|\boldsymbol{b} - \boldsymbol{r}|^3} + \frac{3}{2} \sum_{\alpha=1}^{3} \sum_{\beta=1}^{3} \frac{\Theta_{\alpha\beta} (b_\alpha - r_\alpha)(b_\beta - r_\beta)}{|\boldsymbol{b} - \boldsymbol{r}|^5} - \frac{1}{2} \sum_{\alpha=1}^{3} \frac{\Theta_{\alpha\alpha}}{|\boldsymbol{b} - \boldsymbol{r}|^3} \tag{3.10}$$

where $Q = \sum_{i=1}^{N} q^{(i)}$ is the total charge of the system, $\boldsymbol{d}$ its dipole moment, with $d_\alpha = \sum_{i=1}^{N} q^{(i)} (r_\alpha^{(i)} - b_\alpha)$, and $\boldsymbol{\Theta}$ its quadrupole moment, with $\Theta_{\alpha\beta} = \sum_{i=1}^{N} q^{(i)} (r_\alpha^{(i)} - b_\alpha)(r_\beta^{(i)} - b_\beta)$ (α, β indicate the Cartesian coordinates).

The Fast Multipole Algorithm (FMA) [50], that reduces the cost of the electrostatic calculation to order N for a system of N particles, is based on this approximation. The simulation of a cell containing the solute and the solvent is divided into subcells. At the beginning of the electrostatic calculation, total charge, dipole and quadrupole moments of each subcell are calculated; the potential energy at $\boldsymbol{r}$ is calcutated exactly for the charges in the same and in the adjacent subcells, and using the approximation (3.10) for the nonadjacent subcells. The key of FMA algorithm is to consider the more distant charges as grouped into large subcells.

3.5.3 Discrete electrostatics

In the all-atoms approximation all solvent molecules are treated explicitly; the periodic boundary conditions (PBC) are introduced to minimize surface effects and to reproduce the bulk phases with a minimal number of atoms. The simulation cell (for convenience here we consider a cubic cell of side L), containing the solute and the solvent, is considered at the center of an infinite system obtained by replicating the original cell in all directions, in order to mimic the presence of an infinite bulk surrounding the N-particle system (Fig. 3.4). For every particle i at position $\boldsymbol{r}_i = (x_i, y_i, z_i)$, there are infinite images at positions $\boldsymbol{r}_i + \boldsymbol{n}L = (x_i + n_1 L, y_i + n_2 L, z_i + n_3 L)$, where $\boldsymbol{n}$ is an integer vector. The total potential energy is:

$$U_{tot} = \frac{1}{2} {\sum_{i,j,\boldsymbol{n}}}' u(|\boldsymbol{r}_{ij} + \boldsymbol{n}L|) \tag{3.11}$$

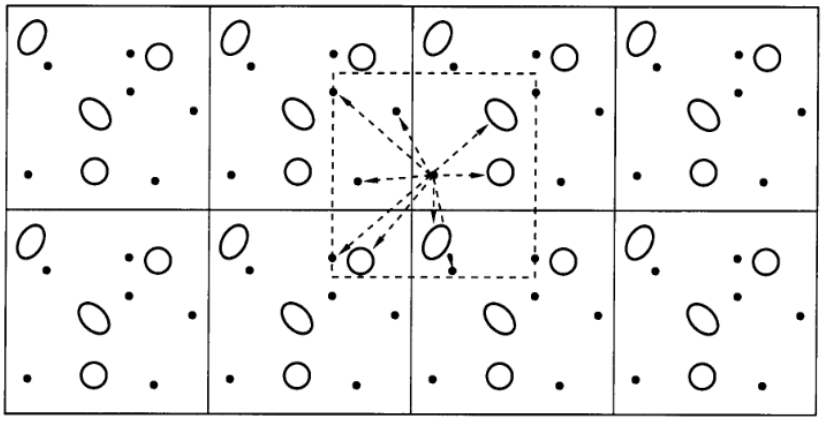


Figure 3.4: Schematic representation of periodic boundary conditions.

the prime over the sum indicates that the terms with $i = j$ have to be excluded when $\boldsymbol{n} = 0$.

In order to calculate the short-range interactions (e.g. Van der Waals interactions), all intermolecular interactions are usually truncated beyond a certain cutoff distance r_c. As seen before, a suitable choice of r_c can yield the desired degree of accuracy.

Ewald sums

The treatment of electrostatic interactions is more complicated; in this case the introduction of a cutoff produces inaccurate results. The Ewald algorithm [42, 51] has been developed to treat the electrostatic interactions in an appropriate way. Applied successfully for many years to the simulation of liquids, it is the reference algorithm for macromolecular simulations.

The basic idea is the modification of charge density. In equation 3.7 the charge density is a sum of δ-functions and its contribution to the energy decays as $1/r$. In the Ewald algorithm every charge q_i is surrounded by a charge density distribution such that the total charge of this cloud cancels exactly q_i (Fig. 3.5). In this new situation, the electrostatic potential at large distance is due only to the fraction of q_i that is not screened by the cloud, and this fraction goes to 0 rapidly. With this expedient the electrostatic energy is split into two parts: (a) a short-range part that one computes

introducing a cutoff; (b) a long-range part defined in the Fourier space, that can be computed using the Poisson equation.

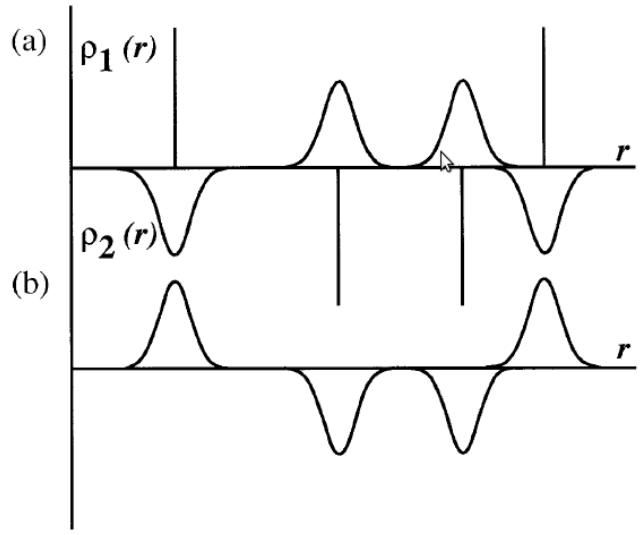


Figure 3.5: (a) Density $\rho(1)$ is the sum of the point charge and of the Gaussian densities. This generates the short-range term. (b) Density $\rho(2)$ equals the Gaussian density with opposite sign, and produces the long-range reciprocal sum potential.

3.6 Free energy calculation

Free energy, usually expressed as the Helmholtz function F or the Gibbs function G, is perhaps the most important thermodynamic quantity and a central concept in modern studies on biochemical systems: in fact many physical properties relevant from the chemical or biochemical point of view depend directly or indirectly on the free energy of the system. For example, binding constants, association and dissociation constants, and conformational preferences are all related to the difference in free energy between states.

The Helmholtz function is appropriate for a system with constant number of particles, temperature and volume and the corresponding ensemble is referred to as the canonical (NVT) ensemble; on the other hand, the Gibbs free energy is appropriate for constant number of particles, pressure and temperature (NPT ensemble). In the following we will refer to the Helmholtz

free energy, but one can extend all considerations to the Gibbs function.

The Helmoltz free energy is defined in thermodynamics as:

$$\mathrm{F{=}U - TS} \tag{3.12}$$

where U is the total energy, T the temperature and S the entropy of a system. From the microscopic point of view, it is a statistical property, which measures the probability of finding a system in a given state. Furthermore, it is a global property that depends on the extent of the phase (or configurational) space accessible to the molecular system.

The statistical physics definition of this quantity is the logarithm of the partition function Z:

$$F = -\frac{1}{\beta}\ln Z \tag{3.13}$$

$$= -\frac{1}{\beta}\ln \int e^{-\beta H(q,p)} dpdq \tag{3.14}$$

where $\beta = 1/(\mathrm{k}_B\mathrm{T})$ (k_B denotes the Boltzmann constant), and q and p represent respectively positions and momenta.

To obtain a good estimate of the absolute free energy, in theory one should sample the whole phase space, which is computationally not possible. But in many applications the important quantities are actually the free energy differences between various macroscopic states of the system, rather than the absolute free energy. Free energy differences allow to quantify the relative likelihood of different macroscopic states; each of these states is the collection of all possible microscopic configurations corresponding to the macroscopic parameters, distributed according to the canonical measure μ; the latter is defined as:

$$\mu(dq\,dp) = Z^{-1}\exp\left[-\beta H(q,p)\right] dq\,dp. \tag{3.15}$$

Many efforts have been made to devise algorithms to overcome sampling barriers and to compute free energy differences. Indeed, in many cases the time trajectory generated in the phase space by the numerical integration is trapped for a long time in some region of the phase space, and hops only occasionally to another region, where it again remains trapped for a long time. This occurs when several regions of low free energy exist in the phase space, separated by regions of high free energy (that is, of very low probability). In [52] one can find a good overview of different methods

of atomistic simulation that can be applied to force a complex system to overcome sampling barriers, classified according to their scope and range of applicabilty; in our work we have used *umbrella sampling* and *metadynamics* as discussed in the following.

3.6.1 Metadynamics

The metadynamics algorithm works in the following way. One considers a dynamical system, in equilibrium at a temperature T, described by a set of coordinates x and by a potential $V(x)$. The algorithm is based on a dimensional reduction: one is usually interested in exploring the properties of the system as a function of a finite number of collective variables CVs such as some angles, some distances, a coordination number, the potential energy or any explicit function of x, assuming that they provide a good coarse-grained description of the system. The algorithm calculates the probability distribution of the system as a function of one or few of these predefined collective variables. For example, in a chemical reaction one would choose the distance between two atoms that have to form a bond. The dynamics in the space of the chosen CVs is enhanced by a history-dependent potential constructed as a sum of Gaussians centered along the trajectory in the CVs space.

The metadynamics method [52] provides in many cases an efficient framework used both for accelerating rare events and for computing the free energy. The method is schematically represented in Fig. 3.6, which shows how it makes the system escape local free energy minima through the lowest free energy saddle point. The same figure illustrates how the method can be used for estimating the free energy.

In the following the capital S is used for denoting CVs as a function of the microscopic coordinates $S(x)$, while lower case s is used for denoting the value of the CVs.

The equilibrium behaviour of these variables is completely defined by the probability distribution:

$$P(s) = \frac{\exp\left[-F(s)/\mathrm{T}\right]}{\int ds \exp\left[-F(s)/\mathrm{T}\right]} \tag{3.16}$$

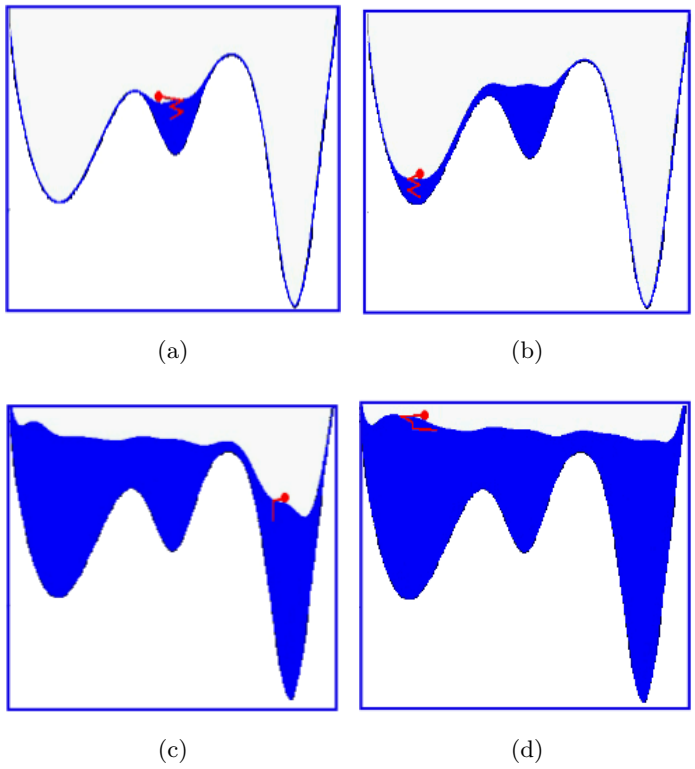


Figure 3.6: The dynamics begins from a minimum of free energy(a). This minimum is quickly filled with Gaussians, and the system evolves through the lowest saddle point in towards a near minimum (b). Afterwards, as the dynamics continues, the free energy profile is progressively filled with Gaussians (c, d). At the end, the sum of the Gaussians provides the negative image of the free energy.

where $F(s)$ denotes the free energy and is given by:

$$F(s) = -\mathrm{T} \ln \left(\int dx \exp \left[-\frac{1}{\mathrm{T}} V(x) \right] \delta \left[s - S(x) \right] \right) \tag{3.17}$$

If one generated a very long trajectory $x(t)$, $P(s)$ could be obtained by taking the histogram of the CVs, i.e., at time t one would have:

$$P(s) \sim \frac{1}{t} \int_0^t dt' \delta \left(s - S \left[x(t') \right] \right) \tag{3.18}$$

If the phase space of the system displays regions of metastability, the motion of S will be often bound in some local minimum of the free energy $F(s)$, or equivalently in a local maximum of $P(s)$, and it will escape from this

regions with very low probability. On the other hand, in metadynamics even if the system is initially at the bottom of a well, the algorithm produces a good sampling by means of the history-dependent potential, that provides an expanding exploration of a progressively larger portion of the configuration space. In the simplest molecular dynamics implementation of this algorithm one introduces a repulsive Gaussian potential every τ_G MD steps. Therefore the potential that acts on the system at time t is given by:

$$V_G\left(S(x),t\right) = w \sum_{\substack{t'=\tau_G,2\tau_G,\dots \\ t'<t}} \exp\Big(-\frac{\left[S(x)-s(t')\right]^2}{2\delta_s^2}\Big) \tag{3.19}$$

where $s(t) = S(x(t))$ are the values of the CVs at time t. The parameters that enter the definition of the V_G and influence the accuracy and efficiency of the free energy reconstruction are:

1. the Gaussian height w;
2. the Gaussian width δ_s;
3. the frequency τ_G by which the Gaussians are added.

If the Gaussians are large, the free energy surface will be explored at a fast pace, but the reconstructed profile will be affected by large errors. Instead, if the Gaussians are small or are placed with low frequency, the reconstruction will be accurate, but it will take a longer time.

The basic assumption of metadynamics is that $V_G(s,t)$ as defined in equation (3.19) provides a good estimate of the underlying free energy after a sufficiently long time:

$$\lim_{t\to\infty} V_G\big(s,t\big) \sim -F(s) \tag{3.20}$$

This relation does not derive from any standard identity for the free energy, but was postulated heuristically [53].

Equation (3.20) can be qualitatively understood in the limit of slow deposition, i.e. $w \to 0$. In this limit, $V_G(s,t)$ varies slowly and the probability of observing s is approximately proportional to

$$P(s) \propto \exp\left(-\frac{1}{\mathrm{T}}\Big(F(s)+V_G(s,t)\Big)\right)$$

If the function $F(s) + V_G(s,t)$ has some local minimum, S will be preferentially localized in the neighborhood of this minimum, and an increasing number of Gaussians will be added there until this minimum is completely filled. On the other hand, in the region where $F(s) = -V_G(s,t)$, the probability distribution will be approximately flat; in this case the corrugations in the free energy are an undesired effect of the number and of the size of the newly added Gaussians.

The efficiency of metadynamics is strongly dependent on the choice of the CVs. In order to obtain a good reconstruction of the free energy, the CVs should describe all the slow events that are relevant to the process of interest; it is also important that all degrees of freedom other than CVs are allowed to relax in the new potential between two depositions of the Gaussians. The CVs should assume clearly distinguishable values in the initial, final, and intermediate states. Finally, it is important to stress that if the number of CVs is too large, it will take a very long time to fill the free energy surface.

A more detailed discussion of the method can be found in Refs. [52, 53].

3.6.2 Umbrella Sampling

The free energy function can be considered as a potential of mean force (PMF) (see section 2.6.1). The concept of potentials of mean force is frequently used to characterize the energetics of transitions in solids, fluids, and biomolecular systems. A routinely used technique to compute the PMF along a given reaction coordinate ξ is the umbrella sampling [54]. This technique aims at overcoming the bias of a limited sampling of energetically unfavorable configurations; it generates a series of initial conditions, each corresponding to a possible state of the system, and confining the simulation around this state with an additional potential, usually harmonic. In this way one can exlplore very low probability areas of the phase space. The Fig. 3.7 illustrates the method. The steps for this procedure are as follows:

1. one generates a series of configurations along a single degree of freedom (reaction coordinate $\xi(t)$) (Fig. 3.7(a)). To produce these configurations one can generate a trajectory by applying an extra force to the system: for example, one can pull apart the system by applying a sufficently strong potential;

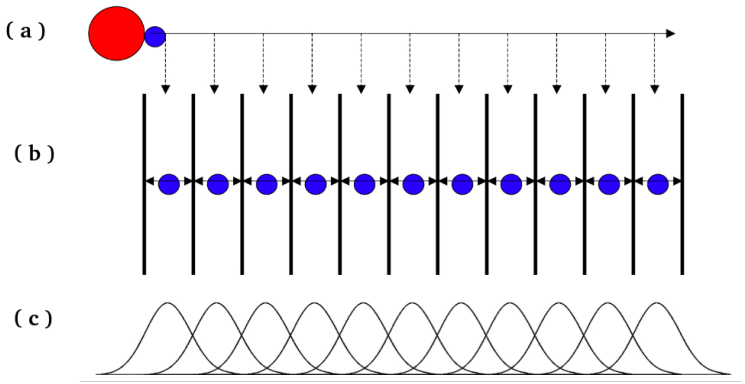


Figure 3.7: Panel (a) illustrates the pulling simulation: a part of the system (blue circle) is pulled away from the rest (red circle). This generates a series of configurations along the reaction coordinate, which in this case is the distance between the centers of mass of the two parts of the system. These configurations are extracted from the constrained trajectory after the simulation is complete, at points indicated by the dashed arrows. Panel (b) schematizes the independent simulations within each sampling window, the center of mass of the subsystem being confined in that window by the umbrella biasing potential. Panel (c) shows the ideal result expected for the histogram of configurations when one uses a harmonic potential; if neighboring windows overlap a continuous energy function can be derived from these simulations.

2. one then extracts from the trajectory produced as in step 1 various frames corresponding to a desired spacing between configurations, as indicated in Fig. 3.7(a) where the dashed arrows point at different positions of the center of mass;

3. a run of the umbrella sampling simulation is started from each configuration, in which the system is restrained within a window centered on the chosen configurations (Fig. 3.7(b));

4. the last step is the calculation of the PMF from the histograms produced in each run (Fig. 3.7(c)). The most widely used technique to compute the PMF from histograms is the weighted histogram analysis method (WHAM) [55].

In this procedure a set of N separate umbrella simulations are carried out, with the usual umbrella potential

$$w_i(\xi) = \frac{k_i}{2}\Big(\xi - \xi_i^c\Big)^2 \tag{3.21}$$

which restrains the system at positions ξ_i^c $(i = 1, ..., N)$ with a force constant k_i. From each of the N umbrella simulations an umbrella histogram is recorded, as in Fig. 3.7(c); it represents the probability distribution $P_i^b(\xi)$ along the reaction coordinate biased by the umbrella potential $w_i(\xi)$. We thus obtain a sequence of biased distribution functions $P_1^b(\xi), P_2^b(\xi), \cdots P_N^b(\xi)$, such that $P_1^b(\xi)$ overlaps with $P_2^b(\xi)$, $P_2^b(\xi)$ with $P_3^b(\xi)$, etc. The unbiased distribution function $P_i(\xi)$ on the $ith-$window can be written in terms of the $P_i^b(\xi)$:

$$P_i(\xi) = \exp\Big(\beta\big(w_i(\xi) - f_i\big)\Big)P_i^b(\xi) \tag{3.22}$$

where f_i is the free energy obtained by adding the biasing potential, and is defined by:

$$\exp\left(-\beta f_i\right) = \int d\xi \exp\big(-\beta w_i(\xi)\big)P(\xi) = \langle \exp\big(-\beta w_i(\xi)\big)\rangle. \tag{3.23}$$

The total unbiased probability distribution can be obtained as a linear combination of the unbiased probabilities:

$$P(\xi) = \sum_{i=1}^{N} c_i(\xi)P_i(\xi) = \sum_{i=1}^{N} c_i(\xi)\left(e^{\beta\Big(w_i(\xi)-f_i\Big)}P_i^b(\xi)\right). \tag{3.24}$$

The $P(\xi)$ is related to the PMF via:

$$W(\xi) = -\frac{1}{\beta}\ln\Big[P(\xi)/P(\xi_0)\Big] \tag{3.25}$$

where ξ_0 is an arbitrary reference point, with $W(\xi_0)$ set equal to zero.

The weights in the $P(\xi)$ equation are:

$$c_i(\xi) = \frac{n_i e^{-\beta\Big(w_i(\xi)-f_i\Big)}}{\sum_{j=1}^{N} n_j e^{-\beta\Big(w_j(\xi)-f_j\Big)}} \tag{3.26}$$

and satisfy the normalization condition $\sum_{i=1}^{N} c_i(\xi) = 1$.

The equations (3.24) cannot be solved directly because they contain two unknown quantities: the free energy constants f_j and the unbiased distribution $P(\xi)$. Therefore, an iterative procedure is necessary; the WHAM equation is:

$$\begin{cases} P(\xi) = \dfrac{\sum_{i=1}^{N} n_i P_i^b(\xi)}{\sum_{j=1}^{N} n_j e^{-\beta\left(w_j(\xi)-f_j\right)}} \\ \\ \exp\left(-\beta f_j\right) = \displaystyle\int d\xi \exp\big(-\beta w_j(\xi)\big) P(\xi) \end{cases} \tag{3.27}$$

It is also possible to introduce in (3.27) a parameter describing the statistical inefficiency: when the sampling of the phase space is not uniformly distributed, a lower weight is assigned to histograms corresponding to trajectories with longer autocorrelations [56]. In solving iteratively equation (3.27), the WHAM procedure estimates the statistical uncertainty of the unbiased probability distribution and subsequently computes by the relation the PMF that corresponds to the smallest uncertainty.

The WHAM implementation in the GROMACS package allows one to compute the statistical error estimates for the derived PMFs, using different bootstrap techniques. Bootstrapping is a resampling technique that can be applied to estimate the uncertainty of a quantity $A(a_1, ..., a_n)$ which is computed from a large set of n observations $a_l (l = 1, ..., n)$.

In theory one could repeat the n observations multiple times in order to calculate several independent estimates of A and to determine the uncertainty on A. But in practice this procedure would require too many observations, and is therefore not feasable for complex systems because it is computationally too expensive. The idea of bootstrapping is to estimate $P(a)$ using the n observations, and to subsequently generate new random sets of n hypothetical observations, based on the estimated distribution. Each of the sets of n hypothetical observations is then used to calculate a hypothetical value for A. The uncertainty on A is then given by the standard deviation of the hypothetical values for A. For a detailed introduction into the bootstrap technique we refer to [57].

The WHAM procedure computes the PMF based on the N trajectories, each taken from one of the umbrella windows $i = 1, ..., N$ along the reaction coordinate. All positions $\xi_i(t)$ during the N simulations may thus

be considered as the large set of observations - which we referred to before as a_l - and are accordly distributed in accordance with $P_i(\xi)$. Thus, one can generate from each window distribution a new hypothetical observation, that is a bootstrapped trajectory $\xi_{b,i}(t)$. Each bootstrapped trajectory $\xi_{b,i}(t)$ yields a new histogram. The new set of N $P_{b,i}(\xi)$ functions is subsequently used in WHAM to compute a bootstrapped PMF $W_b(\xi)$. The whole procedure is repeated N_b times yielding a large set of bootstrapped PMFs $W_{b,k}(\xi)(k = 1, ..., N_b)$. The uncertainty on the PMF is then given by the standard deviation as computed by the N_b bootstrapped PMFs:

$$\sigma_{PMF} = \sqrt{\frac{1}{N_b - 1} \sum_{k=1}^{N_b} \Big(W_{b,k}(\xi) - \langle W_b(\xi) \rangle \Big)^2} \tag{3.28}$$

where:

$$\langle W_b(\xi) \rangle = \frac{1}{N_b} \sum_{k=1}^{N_b} W_{b,k}(\xi) \tag{3.29}$$

denotes the average of the bootstrapped PMFs at position ξ.

In [56] the authors demonstrate that, given sufficient sampling, the bootstrapped new trajectories allow for an accurate error estimate; they presented the results obtained on two test systems.

3.7 Essential Dynamics

The biological functions of the proteins or nucleic acids are connected to the correlations of internal atomic motions. Such complex correlations between internal atomic motions is inherent in the structures and in the characteristic interactions of each system. It is a challenge to derive the correlated motions from the knowledge of molecular structure and about the interactions, in order to identify the functional role of biomolecules.

To investigate the correlations between atomic positional fluctuations one can use the covariance analysis, also called principal component analysis or essential dynamics [58]. The covariance analysis uses the covariance matrix C of the atomic coordinates defined as:

$$C = \langle (\boldsymbol{x} - \langle \boldsymbol{x} \rangle)(\boldsymbol{x} - \langle \boldsymbol{x} \rangle)^T \rangle \tag{3.30}$$

where $\boldsymbol{x}(t)$ is a trajectory generated with a molecular dynamics simulation, and $\langle\ \rangle$ denote an average over time. The generic element of C is:

$$C_{ij} = \langle(\boldsymbol{x}_i - \langle\boldsymbol{x}_i\rangle)(\boldsymbol{x}_j - \langle\boldsymbol{x}_j\rangle)^T\rangle. \tag{3.31}$$

For a system of N particles C is a $3N \times 3N$ symmetric matrix (if desired $\boldsymbol{x}(t)$ can represent a trajectory of a subset of atoms, usually the C_α atoms for the proteins), can be diagonalized by an orthogonal coordinate trasformation T:

$$(\boldsymbol{x} - \langle\boldsymbol{x}\rangle) = T\boldsymbol{q} \qquad \text{or} \qquad \boldsymbol{q} = T^{\mathrm{T}}\left(\boldsymbol{x} - \langle\boldsymbol{x}\rangle\right). \tag{3.32}$$

T transforms C into a diagonal matrix $\Lambda = \langle\ \boldsymbol{q}\boldsymbol{q}^T\rangle$ of eigenvalues λ_i, and the $i-th$ column of T is the eigenvector belonging to λ_i:

$$C = T\Lambda T^{\mathrm{T}} \qquad \text{or} \qquad \Lambda = T^{\mathrm{T}}CT \tag{3.33}$$

The eigenvalues λ_i are the average square displacements along the associated eigenvector directions. One found that in the proteins most of the positional fluctuations are concentrated in a subspace of only a few degrees of freedom [58]. This subspace is called the *essential subspace*. On the other hand, the all other degrees of freedom represent much less important, basically indipendent, Gaussian fluctuations, orthogonal to the essential subspace. This offers the possibility to consider the motion outside the essential subspace as essentially constrained and to represent the protein dynamics in the essential subspace only.

Using the equation $\boldsymbol{q}(t) = T^{\mathrm{T}}\left(\boldsymbol{x}(t) - \langle\boldsymbol{x}\rangle\right)$, the trajectory can be projected on the essential directions to give the principal components $q_i(t)$.

We refer to the literature for more details [58–60].

Part II

Computational Experiments

Chapter 4

Adsorption of DNA oligomers on amine-functionalized surface

4.1 DNA sequencing and diagnostic tool

In recent years, a considerable effort has been devoted to improve the speed and reliability of full-genome sequencing. A human genome is about 3 billion base pairs long, and certain variations in the sequence are well known to cause serious health problems. The sequence of the human genome can be determined with the current technology, but at a high cost and considerable time. Alternative technologies for faster and cheaper DNA sequencing are likely to emerge, and the improvements in this field have also been driven by the expectation that it could become an important diagnostic and medical tool, possibly leading to personalized therapies.

To this end, many methods have been proposed, using various physical properties of the DNA molecule to perform the analysis. For example, the DNA can be forced to pass through a nanopore, while reading the order of nucleotides from the different electrostatic properties of the base pairs. In this respect, computer modelling has been proven to be a good tool to understand the details of the interaction of DNA with the sequencing apparatuses [61, 62].

Before DNA can be sequenced, it has to be extracted from the cell where it is stored (e.g. the blood cells); therefore, any genomic analysis requires a

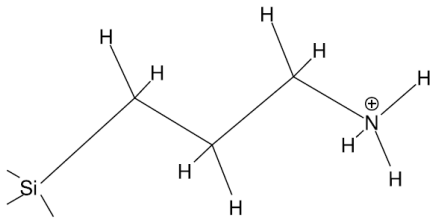


Figure 4.1: Schematic representation of the APTES moiety used to functionalize silica surfaces.

purification process in order to remove proteins, lipids and any other contaminants. One of the methods used to purify DNA from a biological sample is based on the fact that under physiological conditions the DNA molecule is negatively charged, with an average of 2 electronic charges per base pair. Under the same conditions amine groups are positively charged, and in principle the electrostatic interaction could be used to separate DNA from the other unwanted molecules, e.g. by functionalizing a surface with amine-carrying molecules so that DNA is preferentially absorbed onto it.

The strength of the electrostatic interaction between charged bodies immersed in an electrolytic solution largely depends on the kind and concentration of the ions present in solution. Under particular circumstances, it might even happen that the interaction between like-charged bodies becomes attractive [63, 64]. The origin of this peculiar effect can be ascribed to the presence of doubly-charged ions in the solution, which develop very strong spatial correlations through a divalent-ion mediated like-charge attraction [65, 66]. The possibility of modulating the electrostatic interaction using dissolved electrolytes suggests that DNA might be efficiently purified by adsorbing it first onto a surface with opposite charge and then releasing it by appropriately changing the type and concentration of dissolved ions.

This has been recently studied in various experimental setups, mostly based on the functionalization of a silica surface with 3-aminopropyltriethoxysilane (APTES) [67, 68] that contains an amino group. These functionalized surfaces can be used into lab-on-a chip devices, and have been shown to be able to extract up to 40% of the DNA present in blood samples [69, 70]. In Fig. 4.1 the relevant moiety of APTES' structure is shown, while Fig. 4.2 depicts schematically the reactions between an oxide surface and the APTES molecules in a functionalization process.

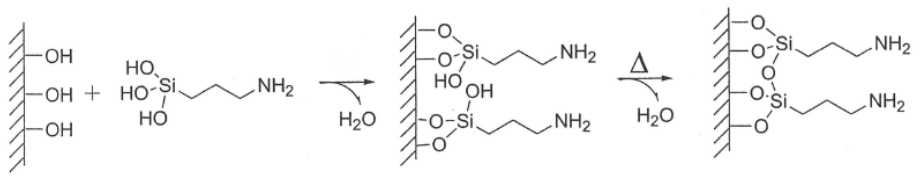


Figure 4.2: Reaction between an oxide surface and APTES molecules.

Divalent-ion mediated like-charge attraction has received considerable attention; on the other hand, the possibility of an opposite-charge repulsion has not been systematically investigated, although there are experimental indications that such a regime can actually happen [71–73]. In particular, the general features of the interaction of DNA with an amine-functionalized surface have not been yet investigated theoretically. Therefore, little is known about the details of how the interaction of DNA with a functionalized surface depends on the pH and on the ionic strength of the surrounding solution, as well as on the charge of the dissolved ions.

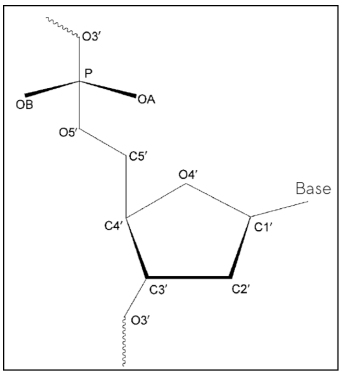
In this thesis we first present the effects of the charge density on the adsorption properties of DNA, obtained by using classical density functional methods to investigate the free energy landscape of DNA oligomers adsorbed on positively-charged surfaces. Density functional methods, that have been proven to be very effective in determining the conditions of like-charges attraction, have been used to investigate adsorption of polyelectrolites [74–76], ion interactions with polyelectrolites [77, 78], and the electric double layer in various geometries [79–81].

We then investigate in detail by computer simulation how the interaction between DNA and an amine-functionalized surface depends on the pH of the solution, and on the type and the concentration of dissolved ions. In this first exploratory work we consider DNA oligomers, with a size able to produce the desired effects within an affordable computational cost. We have performed free energy calculations using an atomistic model of DNA and of the functionalized surface, which allowed us to provide a more detailed description of the adsorption mechanism. This approach showed that the detachment of DNA oligomers from the surface can indeed take place, and is the result of an interplay between the attractive electrostatic interactions and an entropic contribution to the potential of mean force between the surface and the DNA.

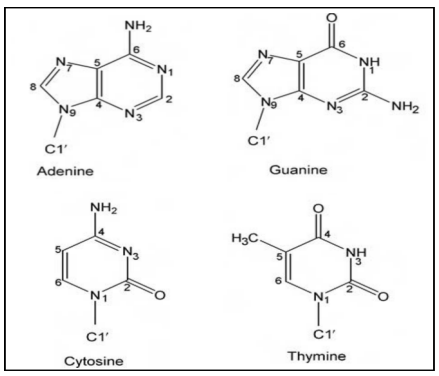
We also report experimental results obtained by the `Biofunctional Surfaces and Interface (BioSInt)` group@FBK (Fondazione Bruno Kessler) using an atomic force microscope (AFM) [82] to detect in a similar system the phenomenon described before. This technique has been utilized for the study of interactions occuring within biological systems such as antibodies-antigens [83, 84], and biological processes such as molecular recognition [85]. The experiments described in section 4.3.3 point out a significant dependence of the DNA-surface interaction on the pH of the solution, which is directly related to the average surface charge, and on the concentration of dissolved ions.

4.2 DNA structure

In this section we describe the principal properties of DNA.



(a) Nucleotide: repeating unit in a polynucleotide chain.



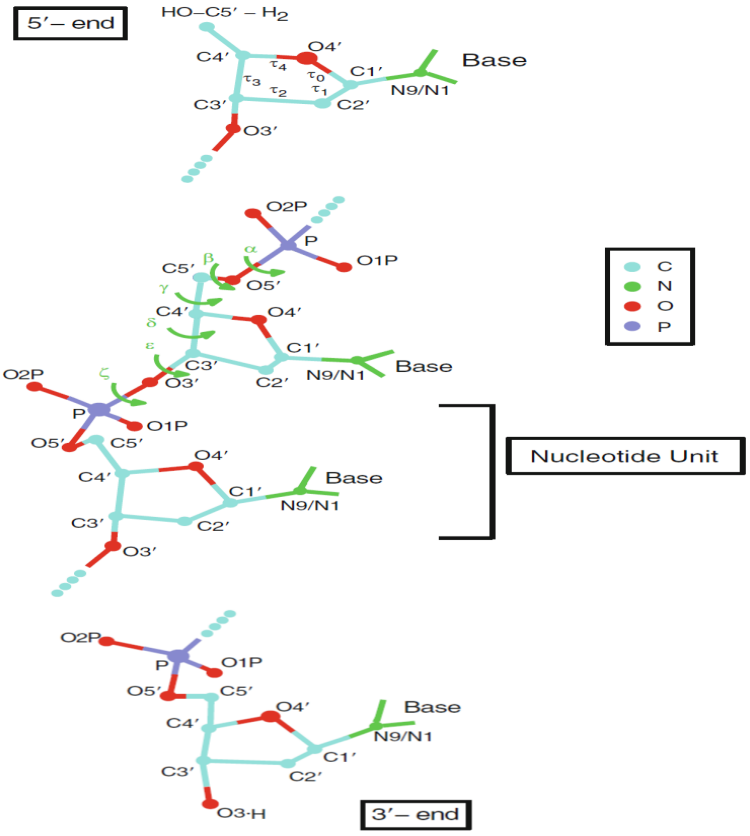
(b) The bases are planar aromatic heterocyclic molecules and are divided into two groups: the pyrimidine bases, thymine and cytosine; and the purine bases, adenine and guanine.

Figure 4.3: The standard nomenclature for the atoms is taken from the International Union of Biochemistry.

The deoxyribonucleic acid is a long polymer built from four different building blocks, the nucleotides. The sequence in which the nucleotides are arranged contains the entire information required to shape the cells and their functions. Despite its essential role in cellular functions, DNA molecules adopt surprisingly simple structures. Each nucleotide contains two parts:

a backbone consisting of sugar (the deoxyribose) and phosphate unit (Fig. 4.3(a)), and an aromatic base (Fig. 4.3(b)).

The unit containing just the sugar and the base is called the *nucleoside*. As shown in Fig. 4.4, individual nucleoside units in a nucleic acid are joined together in a linear way, through phosphate groups attached to the 3' and 5' positions of the sugars. Hence the full repeating unit in a nucleic acid is a 3', 5'-nucleotide.



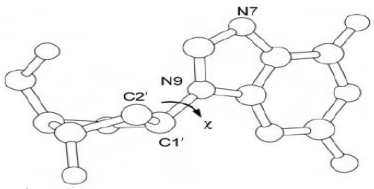


Figure 4.4: A polynucleotide chain (standard atom labeling is used) runs from the 5' end (of the sugar atom C5') to the 3' end (of the sugar atom C3'). Nucleotide linkage is via the 3' to 5' phosphodiester bonds. $\alpha, \beta, \gamma, \delta, \epsilon,$ and ζ label nucleic acid torsion angles along the polynucleotide chain, and χ the torsion angle around a glycosyl bond, connecting sugar and base; τ_0 through τ_4 represent the length of endocyclic bonds in the sugar.

The phosphate group is negatively charged and, being located on the exterior of the helix (see Fig. 4.5), is readily available for physical and chemical interactions with solvent water molecules and ions present in the cell.

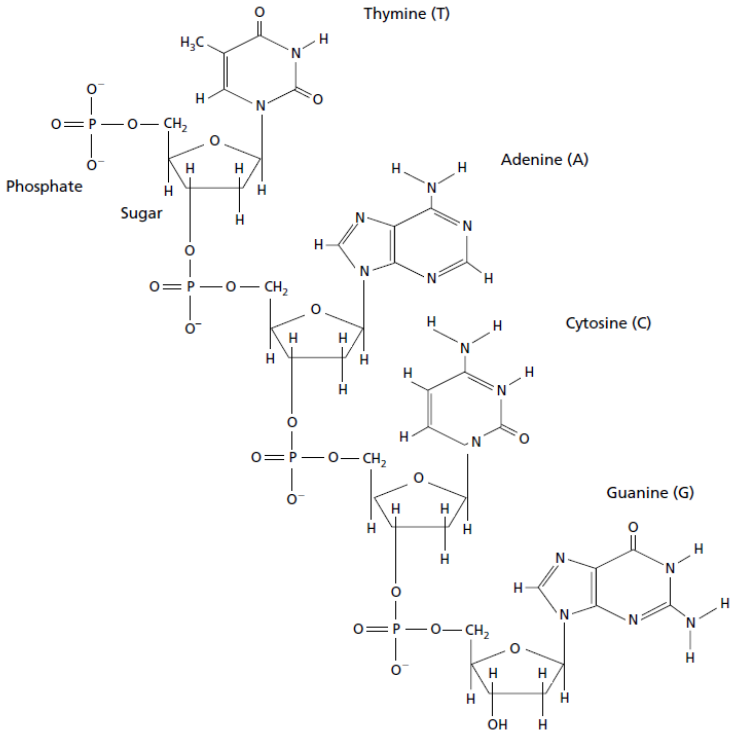


Figure 4.5: Primary structure of one strand of DNA.

Accurate bond lengths and angle geometries for all bases, nucleosides and nucleotides have been determined by X-ray crystallography. In structural surveys [86–89] the mean values for these parameters, that define their equilibrium values, have been calculated from the most reliable structures in the Nucleic Acid Databases. These parameters have been incorporated into several implementations of the AMBER force fields that we have used in our computer simulations, as well as in other force fields. Accurate crystallographic analyses, at very high resolution, can also yield quantitative information on the electron-density distribution in a molecule, and hence on individual partial atomic charges. These charges have been obtained by *ab-initio* quantum mechanical calculations, but are also available experimentally [90, 91].

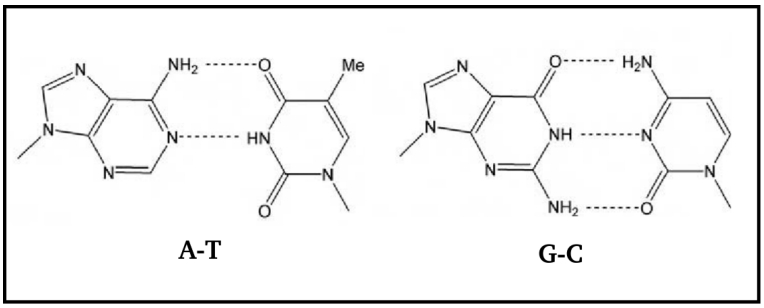


Figure 4.6: In the classic Watson-Crick base pairing scheme of DNA double helices, thymine (T) pairs with adenine (A) forming two hydrogen bonds, and cytosine (C) pairs with guanine (G) forming three hydrogen bonds.

The realization that the planar bases can bind in particular ways by means of hydrogen bonding (see Fig. 4.6), was a crucial step in the elucidation of the structure of DNA. Important early experimental data by Chargaff showed that the molar ratios of adenine-thymine and cytosine-guanine in DNA were both unity. This led to the proposal by Watson and Crick that in each of these pairs the purine and pyrimidine bases are held together by specific hydrogen bonds, to form planar base pairs. In native, double-helical DNA the two bases in a base pair necessarily arise from two separate strands of DNA (with intermolecular hydrogen bonds) and so hold together the DNA double helix (Watson and Crick, 1953) [1].

In a double helix (Fig. 4.7) only specific pairs of bases can bind together: A (purine) with T (pyrimidine), and G (purine) with C (pyrimidine). In other words if an A is a member of a pair, on either chain, then the other member must be T; similarly for G and C. The sequence of bases on a single chain does not appear to be restricted in any way. Hovewer, if only specific pairs of bases can be formed, it follows that if the sequence of bases on one chain is given, then the sequence on the other chain is automatically determined. This arrangement produces CG and TA base pairs (bps) whose size along the double-helix are nearly identical, and therefore the overall structure appears quite uniform.

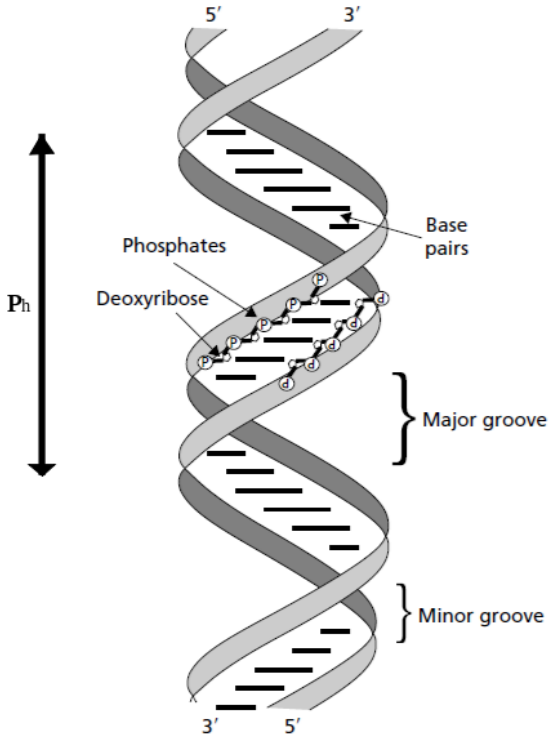


Figure 4.7: The Watson and Crick model of DNA double helix. Alternating subunits of phosphates and deoxyriboses held together by hydrogen-bonded bps, with adenine pairing with thymine and guanine pairing with cytosine. The A-T and G-C bps have the same distance between the C1' atoms of their sugars and can form a regular helix. The two chains of the DNA double helix run in opposite directions, through the center of each bp. The *pitch* of the helix P_h is the helix axis for one complete turn, and n_b is the number of bps per turn (10 - 10.5). The unit *twist* is defined as $\Omega = 360/n_b$ (about $34 - 36$), and the helical rise is $h = P_h/n_b$. The major grooves (12 Å) and the minor grooves (8.5 Å) are indicated.

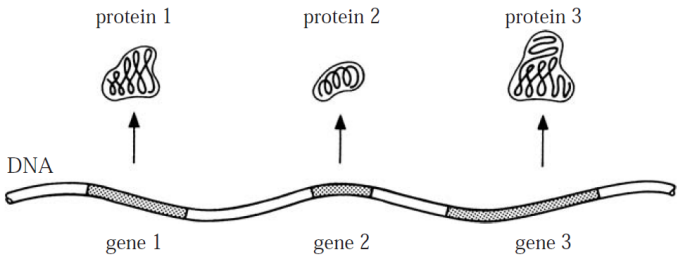


Figure 4.8: The DNA is schematically represented as a long structure. Some stretches of the sequence, the *genes*, contain all information needed to code the proteins.

The order in which the nucleotides appear in one DNA strand defines its sequence. Some stretches of the sequence, called genes, contain information that can be translated first into an RNA molecule and then into a protein, as schematically represented in Fig. 4.8. The ensemble of all genes of an organism constitutes its genetic information, and is called the genome.

4.2.1 Classical Density Functional Calculations

Electrostatic interactions are the driving force for many phenomena relevant to biophysics, and electrostatics is most likely the main driving force in the reversible adsorption of DNA on functionalized surfaces. From the point of view of modelling, a continuum description of the electrostatics of a given problem is a first approximation that is usually capable of giving a considerable insight into the relevant scales characterizing a phenomenon. To this end, we present the results of a simplified model of the DNA molecule interacting with a functionalized surface. In this model, a DNA segment is

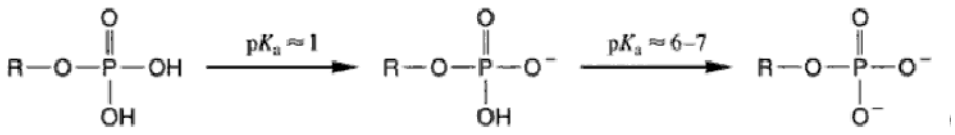


Figure 4.9: Deprotonation reaction of a phosphate group, as it occurs in the exterior backbone of the double helix of DNA when the pH value is similar to the indicated pK_a value.

pH	average charge/group
7.65	0.10
7.45	0.15
7.30	0.20
7.04	0.314
6.88	0.40

Table 4.1: Average charge in units $|e|$ per amine group on the functionalized surface as a function of the pH. The equilibrium costant of the oxidation reaction of the amine group is pK_a=6.7.

approximated as a rigid cylindrical molecule parallel to the functionalized surface, with radius $R = 1$ nm and length $l = 0.34$ nm per base pair. Under physiological conditions, each base pair carries a charge of $-2|e|$, where e is the electron charge, due to the deprotonation of the phosphate groups (as shown in Fig. 4.9). This reaction has pK_a=7 and this charge can be safely considered constant, up to very acid values of the pH. One assumes that the negative charge on the DNA is uniformly spread onto the cylinder, which, as a consequence, acquires a surface charge density of

$$\sigma_c = -\frac{2|e|}{2\pi R l} \tag{4.1}$$

$$= -0.936\,|e|\mathrm{nm}^{-2} \tag{4.2}$$

This cylinder interacts with an amine-functionalized surface, which we also assume to carry a uniform charge density, due to the protonation reaction of the amine groups. The APTES moieties are assumed to be uniformely distributed on the surface with a density of 1.00 nm^{-2}. This is the actual experimental density of amine groups on APTES functionalized silica surfaces [92].

Using the Henderson-Hasselbach equation (2.35), we calculate the charge per amine group as a function of pH (see Fig. 4.10 and Tab. 4.1 where some values extracted from the curve are reported).

Given these representations of the DNA molecule and the amino-functionalized surface, the free energy of the system depends on the concentration and distribution of the other ions present in the solution. Analogously to the Hohenberg-Kohn theorem for the electron liquid, it can be shown that the free energy is a universal functional of the ionic densities

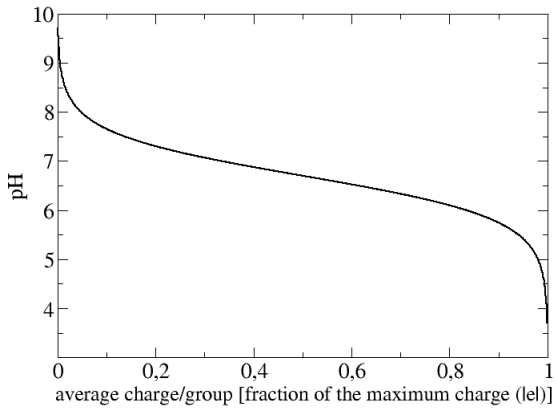


Figure 4.10: Behaviour of pH value as a function of charge per amine group. The function is calculated using the Henderson-Hasselbach equation.

[93–95], even though the proof leading to this result does not give the actual form of this functional. Nevertheless, this theorem is the foundation of the application of Density Functional Theory (DFT) to the study of the interactions between polyelectrolytes in ionic solutions.

As a first approximation, the free energy can be taken as the sum of the electrostatic energy of interaction between the ionic species and of the free energy of the ideal gas for each species, taken in the local density approximation. Denoting by $\rho_i(x)$ the density of species i, ($i = 1, 2 \ldots n$ where n is the total number of ionic species), one has:

$$\begin{aligned} \mathfrak{F}_{PB} = \; & k_B T \sum_{i=1}^{n} \int d^3x \rho_i(x) log[\rho_i(x)\Lambda_i^3 - 1] \\ + \; & \frac{|e|^2}{2} \sum_{i,j}^{1,n} \int d^3x d^3y \frac{z_i \rho_i(x) z_j \rho_j(y)}{\epsilon |x-y|} \\ + \; & \sum_{i=1}^{n} \int d^3x |e| z_i \rho_i(x) V_{ext}(x) \end{aligned} \tag{4.3}$$

where Λ_i is the de Broglie thermal wavelength of the particles of species i, z_i its charge in units of $|e|$, ϵ is the dielectric constant of the solvent (~ 80 for water), which is assumed to be uniform, and $V_{ext}(x)$ is the electrostatic potential generated by an external charge distribution, which is assumed to be fixed. The minimization of Eq. (4.3) with the relevant boundary condi-

tions produces the average density of the electrolytes that is a solution of the Poisson-Boltzmann (PB) (Eq. (2.31)). In Eq. (4.3) one completely neglects the presence of ion-ion correlations. However, these correlations have been shown to play an important role in the description of the interaction between charged bodies, especially in the presence of divalent ions, where like-charge attraction phenomena can be observed [66].

To overcome the limitations of Eq. (4.3), various functionals have been proposed and tested in the past years. In general, there are two physical phenomena producing ion-ion correlations. First, one has to consider excluded volume effects, deriving from the fact that the ions have a finite diameter ($\sigma \sim 0.3$ nm); therefore, the first term on the right-hand side of Eq. (4.3), which is the free energy of an ideal gas of point particles, is better approximated with the free energy of a hard-sphere fluid. The results presented here are obtained with this approximation, exploiting the results obtained by Rosenfeld and collaborators using Fundamental Measure Theory (FMT) [96, 97]; in particular the White Bear version of this functional has been used, also known as FMT3 [98]. The set of this approximations, that is hard-sphere ions immersed in a uniform solvent, is known as the Restricted Primitive Model (RPM) of the electrolytes.

Secondly, ion-ion correlations are also induced by the long-range Coulomb interaction: the second and third terms on the right-hand side of Eq. (4.3) can be improved to take into account at least pair correlations. The calculation uses the free-energy density functional term calculated by Tang and collaborators using the Mean Spherical Approximation solution of the Ornstein-Zernicke equation for interacting charges [65]. It has been shown that RPM plus ion-ion electrostatic correlations is a very good model, able to describe efficiently the main effects of electrostatics in a wide range of systems of biological interest.

Classical DFT simulations have been performed by Giovanni Garberoglio [99] using the freely available TRAMONTO software [100]. The approximations discussed above reduce the system under study to a two-dimensional system, whose dimension is taken as 20×10 nm^2, where the first length is along the direction perpendicular to the functionalized surface. Two different charge densities for the surface, namely $\Sigma_1 = 0.15\,|e|$ nm^{-2} and $\Sigma_2 = 0.314\,|e|$ nm^{-2}, are considered. Assuming a density of amine groups ~ 1.0 nm^{-2}, the two charged states correspond to values of pH close to 7.45

and 7.04, respectively (see Tab. 4.1). The concentration of ions has also been varied, considering both monovalent and divalent ones. In the DFT simulations, the ionic concentration was set as the boundary condition far from the charged surface.

We report in Fig. 4.11 and Fig. 4.12 the results obtained for the DNA free-energy as a function of the distance from the surface, for the two charge densities. The distance has been defined as the smallest one between any point on the surface and any point of the DNA molecule.

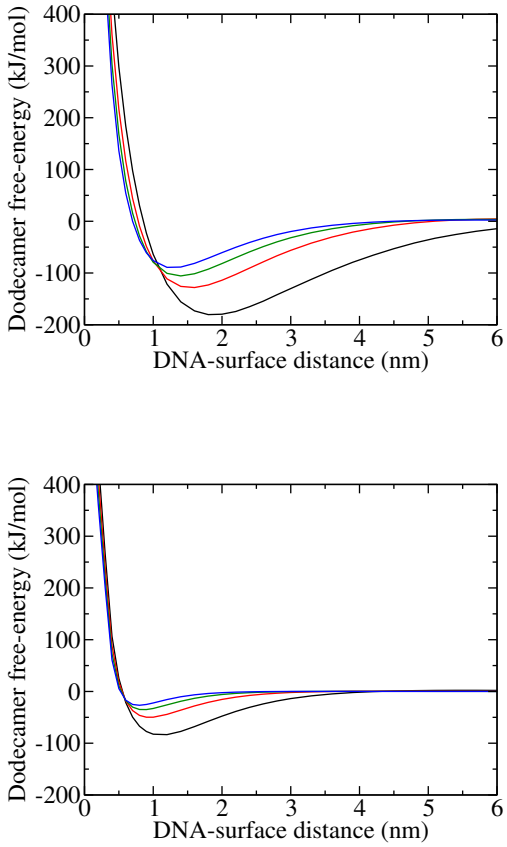


Figure 4.11: Free energy of a DNA dodecamer as a function of the distance from an amine functionalized surface with fixed charge density $\Sigma = 0.15\,|e|$ nm^{-2}, calculated using DFT. The two panels refer to monovalent (top) and divalent (bottom) ions. The four curves in each panel refer to different concentrations in the solution: 0.1 M (black), 0.2 M (red), 0.3 M (green), 0.4 M (blue).

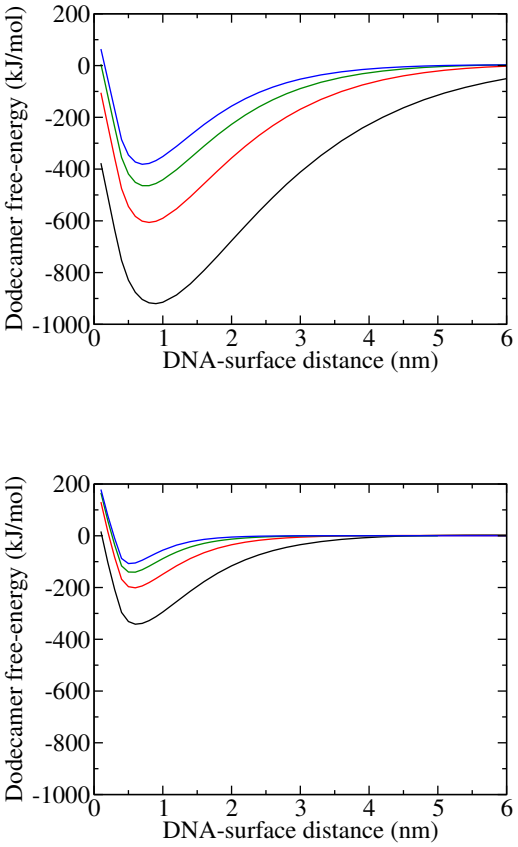


Figure 4.12: Free energy of a DNA dodecamer as a function of the distance from an amine functionalized surface with fixed charge density $\Sigma = 0.314\,|e|$ nm^{-2}, calculated using DFT. The two panels refer to monovalent (top) and divalent (bottom) ions. The four curves in each panel refer to different concentrations in the solution: 0.1 M (black), 0.2 M (red), 0.3 M (green), 0.4 M (blue).

At the lower pH, that is when the charge density on the functionalized surface is higher (see Fig. 4.10), the free-energy profiles have a well defined and deep minimum, meaning that the DNA is very likely to be adsorbed under these conditions. We recall that at room temperature $k_BT = 2.5$ kJ mol^{-1}, and therefore negligible when compared to the depth of the well. When passing from a 0.1 M solution of monovalent ions to a 0.4 M solution of divalent ions the free energy of adsorption can be reduced 9-fold. When the pH increases, or equivalently the charge density on the functionalized

surface decreases, the free-energy profiles still show a well defined minimum close to the surface, but the values of the free energy at this minimum is higher than in the previous case; that is, the depth of the well is smaller.

The reduction of the well's depth when the ionic concentration is increased indicates that the dissolved ions are able to screen the electrostatic interaction between the DNA and the surface. Smaller charge densities on the surface have not been studied because the pH needed to produce these states would be so high (see Fig. 4.10) as to result in the denaturation of DNA.

These calculations indicate that divalent ions can be very effective in screening the electrostatic interaction. For the same charge density, namely the molar concentration times the ionic-charge, the adsorption free-energy in the presence of divalent ions is significantly smaller than when monovalent ions are present. This can be seen, in both Fig. 4.11 and Fig. 4.12, by comparing the curves corresponding to 0.1 M and 0.2 M of divalent ions (bottom panels) with, respectively, the curves corresponding to 0.2 M and 0.4 M for monovalent ions (top panels).

To investigate more in detail the influence of charge screening on the adsorption free-energy, the ionic distributions have been computed as a function of the concentration for a fixed value of the charge on the adsorbing surface (Fig. 4.13 and Fig. 4.14). One sees that the negative-ion density is larger close to the positively charged surface, as expected, and its peak value increases with increasing ionic concentration. As a consequence, the DNA-surface electrostatic interaction is progressively screened, resulting in a smaller adsorption free-energy. A complementary picture is obtained by analyzing the distribution of positive ions, as shown in Fig. 4.14. The positive ions are most likely found around the negatively charged DNA molecule, with a density increasing with the salt concentration, thus contributing to the screening of the DNA-surface interactions.

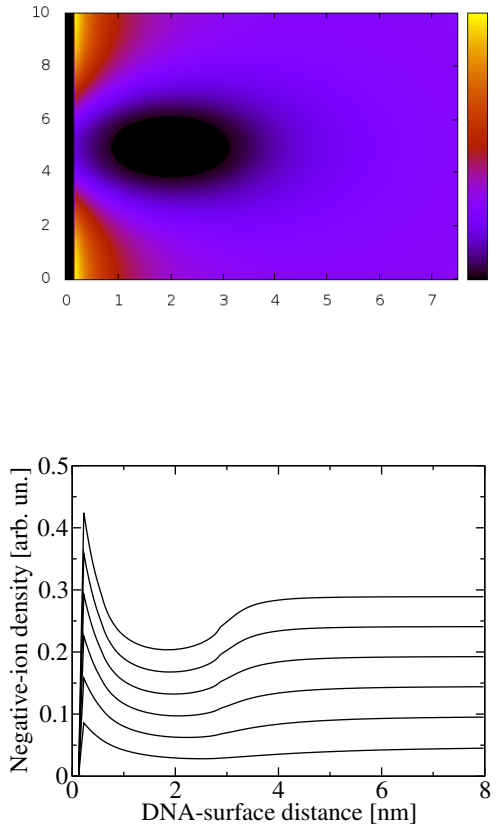


Figure 4.13: Negative-ion densities from classical DFT calculations with DNA-surface distance fixed to 0.7 nm and immersed in a solution of divalent ions and with a surface charge of $0.314\,|e|$ nm^{-2}. Top panel: two-dimensional cut at a concentration of 0.1 M. The axis of the molecule and the funcionalized surface are perpendicular to the figure. Figures on the axes are given in nm; horizontal axis, distance from the functionalized surface; vertical axis, distance along the same surface. The colour scale represents the variation of the charge density in arbitrary units. Bottom panel: integrated density as a function of the distance from the surface. The curves correspond to increasing concentrations from 0.1 to 0.6 M, going from the bottom to the top.

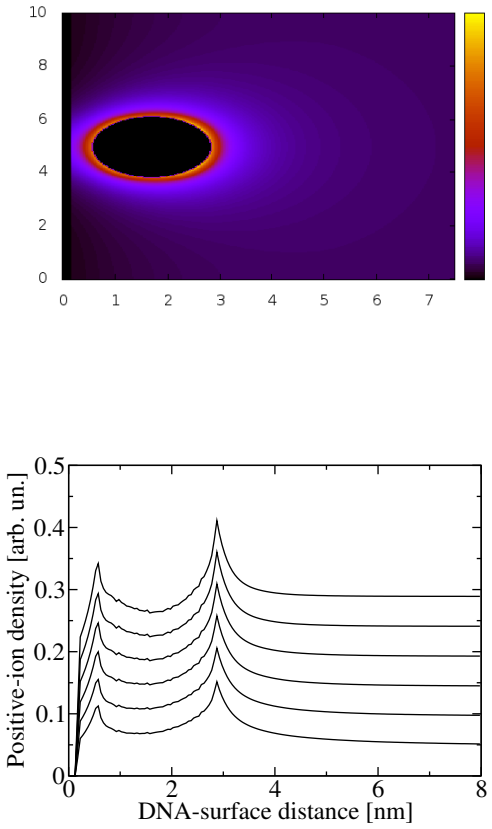


Figure 4.14: Positive-ion densities from classical DFT calculations with the DNA-surface distance fixed at 0.7 nm, immersed in a solution of divalent ions, and with a surface charge of $0.314\,|e|$ nm^{-2}. Top panel: two-dimensional cut at a concentration of 0.1 M. Figures on the axes are given in nm; horizontal axis, distance from the functionalized surface; vertical axis, distance along the same surface. The colour scale represents the variation of the charge density in arbitrary units. Bottom panel: integrated density as a function of the distance from the surface. The curves correspond to increasing concentrations from 0.1 to 0.6 M, going from the bottom to the top.

4.3 Molecular dynamics and umbrella sampling calculations

The DFT analysis of the system shows that the presence of dissolved ions may have a significant impact on the adsorption free-energy of DNA onto an amine-functionalized surface; nevertheless, one doesn't observe anything close to a phenomenon of opposite-charge repulsion, which would allow a efficient control of adsorption and desorption of DNA by varying the concentration of salts.

As a matter of fact, the inherent approximations of the classical DFT approach (where the effect of the solvent is taken into account only through an overall dielectric constant) do not allow a description of the desorption process. To further investigate this matter, we resorted to a fully atomistic modeling of the DNA and of the surface. In our model the surface is a (001) surface of crystalline silica, functionalized with APTES molecules. In Fig. 4.15 we give a simplified picture of our model: the ions that generate the desired value of the concentration, and the water molecules that fill the simulation box when the explicit solvent model is used, are not shown. The silica atoms are fixed throughout all simulations, and act as Lennard-Jones (LJ) force centers. The LJ parameters for Si and O, that have been taken from the DREIDING force field [101], are indicated in Tab. 4.2. The APTES molecules attached to the surface have been described using the GROMOS force field, with a topology generated using the PRODRG program [102].

Finally, we have considered a Dickerson dodecamer of DNA (the sequence is CGCGAATTCGCG), described using the AMBER-99 force-field. Cross LJ interactions between different atoms have been taken into account by the Lorentz-Berthelot mixing rules and all computations have been performed with the GROMACS 4.5.3 Molecular Dynamics (MD) package [11].

	ϵ(kJ mol^{-1})	σ(nm)
Si	2.45	0.339
O	1.72	0.263

Table 4.2: The Lennard-Jones parameters given by the DREIDING force field, used to model the crystalline silica surface.

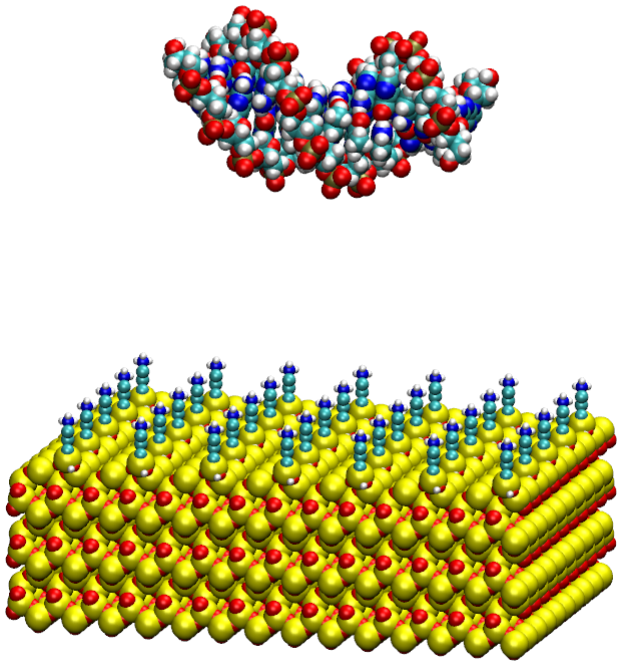
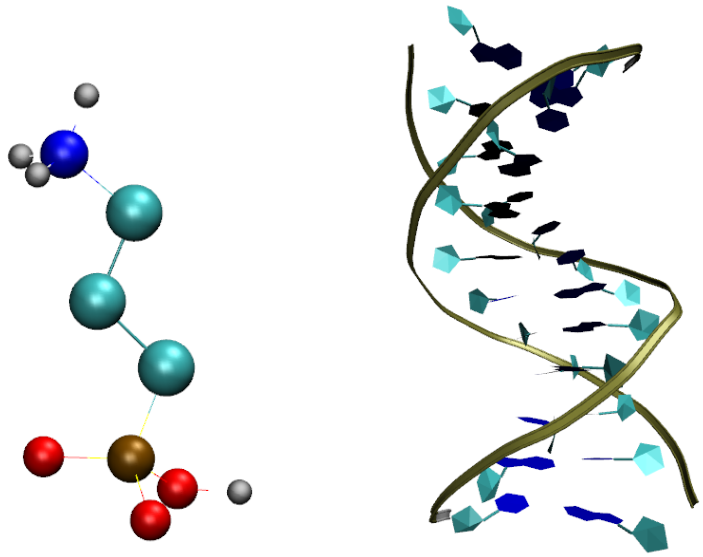


Figure 4.15: Atomistic model of the system constituted by DNA and a surface of crystalline silica functionalized with APTES molecules. When the molecular dynamics simulation is made using the implicit solvent, the box is filled with ions up to the desired concentration. Moreover, when the explicit model of water is used, the box is filled with ions (at the desired concentration) and with water molecules.

4.3.1 Implicit solvent model

In an initial exploratory set of calculations, we adopted the generalized Born implicit solvent model (see section 2.9), as implemented in GROMACS. In this case, one uses simple cut-offs to describe the electrostatic interactions, and we have set the cut-off radius to be $R_{\text{cut}} = 1.5$ nm. We have immersed the system in a continuous medium with relative dielectric constant equal to that of water. The size of the simulation box was $5 \times 7 \times 40$ nm, with the largest dimesion perpendicular to the surface.

We generated the umbrella configurations by a lightly pushing the DNA



(a) (b)

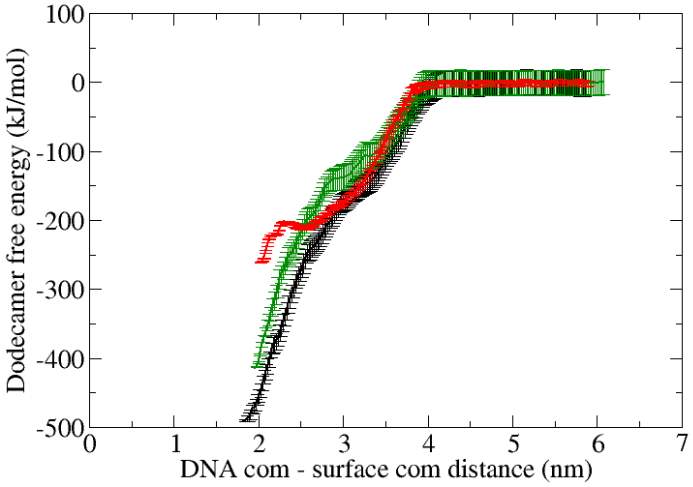
Figure 4.16: Representation of the atomistic model of the APTES molecule [panel (a)] and of the DNA dodecamer [panel (b)]. The DNA visualization displays the backbone (tan), the sugars (cyan) and the bases (blue) of the sequence.

molecule towards the substrate, starting from a distance $d = 6$ nm from the surface. From here on, the distance between the molecule and the surface is defined as the distance between the respective centers of mass; this distance is about 2 nm larger than the distance used in the DFT calculation. From this trajectory we extract a number of configurations uniformly sampling the distance of DNA from the surface, with an average interval of $\Delta z = 0.2$ nm between positions of the center of mass of the molecule. Starting from each of these configurations the center of mass of DNA is restrained by a harmonic potential having an elastic constant $k = 1000$ kJ mol^{-1}nm^{-2} and a new trajectory of 1 ns is generated after 30 ps of equilibration. All of these trajectories are then used as input of the WHAM free-energy calculation.

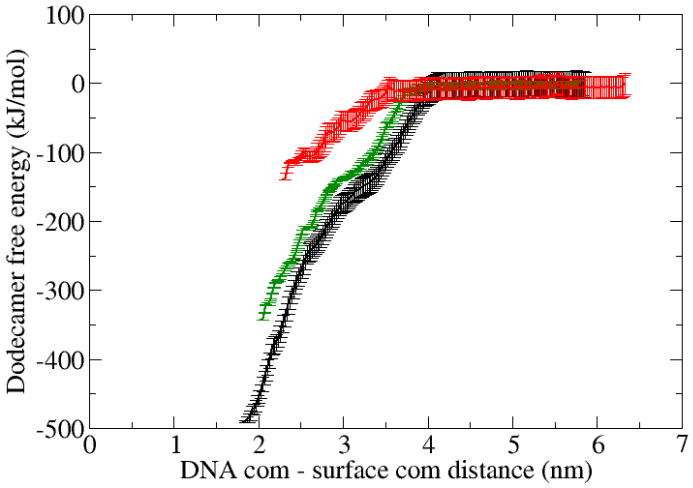
As in our atomistic simulations we are obliged to have a neutral simulation box, to avoid spurious effects due to electrostatic forces we must first introduce in the box a number of counterions (less than or equal to 22) equivalent to the sum of the net charges of the DNA molecule (-22 $|e|$) and

of the APTES molecules. Secondly, we have to add further ions to simulate the various ionic concentration of the solution. The number of these additional ions varies from 0 for the case in which only counterions are present, to 502 (half positive and half negative) for a 0.3 M solution, to 1002 for a 0.6 M solution. In order to compare free-energy profiles obtained with different ionic densities, we have to make sure that the same number of particles are present in the simulation box, independently of the ionic strength of the solution. Otherwise, the free energy of the system would have depended on the number of ions present, due to the different collision rates between ions and DNA in the various cases. This in turn would have made a direct comparison between different conditions quite difficult. To overcome this problem, we added to the system composed by the DNA dodecamer and the functionalized surface a fixed number of neutral atoms. In simulating different ion concentrations, we simply endowed an appropriate fraction of these atoms with the proper charge. In this way, we achieved a different ionic concentration, but not a different collision rate, as the *only* difference between the various states. Notice, however, that in this case we are not implementing the ionic concentration adopted in the DFT simulations, as we introduce real ions: in the atomistic case the ionic concentration is defined as the average density of ions in the simulation box and not, as in the case of DFT calculations, as the boundary condition to be fulfilled far from the adsorbing surface.

The PMF profiles calculated by the WHAM algorithm (see section 3.6.2) are shown in Fig. 4.17 for the case of the Dickerson dodecamer. The general features observed in the DFT analysis are reproduced here. The absolute value of the free energy is reduced when the concentration of dissolved ions is increased, and the interaction between the DNA and the surface becomes quite small a few nanometers away from the surface. Although one would not expect the actual values of the adsorption free energies to be comparable between the continuum (DFT) and the atomistic model, given the different approximations adopted in the two cases, we notice that there are some common features. In the case of monovalent ions and with a surface charge density of $\Sigma = 0.314$ nm^{-2} the minimum of the free energy and its position are similar at 0.3 M [see Fig. 4.12 and Fig. 4.17 (top panel)]; the value at 0.6 M of the free energy has not been calculated with the DFT method, but its position in Fig. 4.12 (top panel) can be inferred to be quite similar to



(a) Black: counterions only; green: 0.3 M of monovalent ions; red: 0.6 M of monovalent ions.



(b) Black: counterions only; green: 0.3 M of divalent ions; red: 0.6 M of divalent ions.

Figure 4.17: Free energy profiles for the adsorption of the Dickerson dodecamer on an APTES-functionalized silica surface with average surface density of $\Sigma = 0.314$ nm^{-2}, as a function of the distance between the center-of-mass of the DNA molecule and the center-of-mass of the surface. The uncertainties of the free energy have been estimated using the *bootstrap* technique [sec. (3.6.2)].

the one that can be seen in Fig. 4.17 (top panel). This similarity of features doesn't hold for the divalent ions. The comparison between experimental results, DFT calculation, and MD simulation is in the section (4.4).

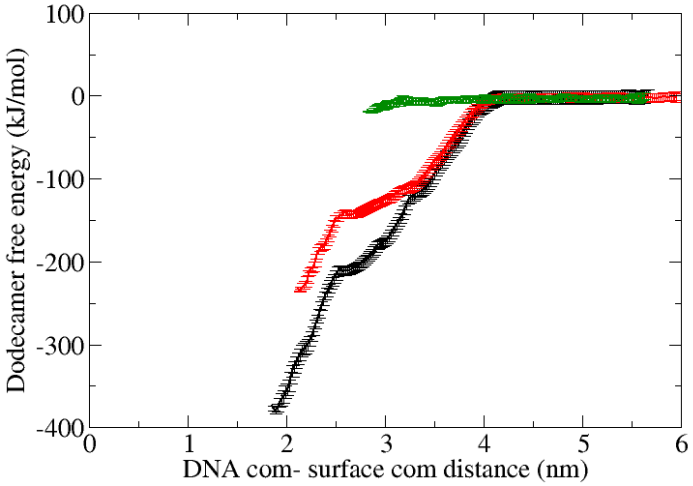
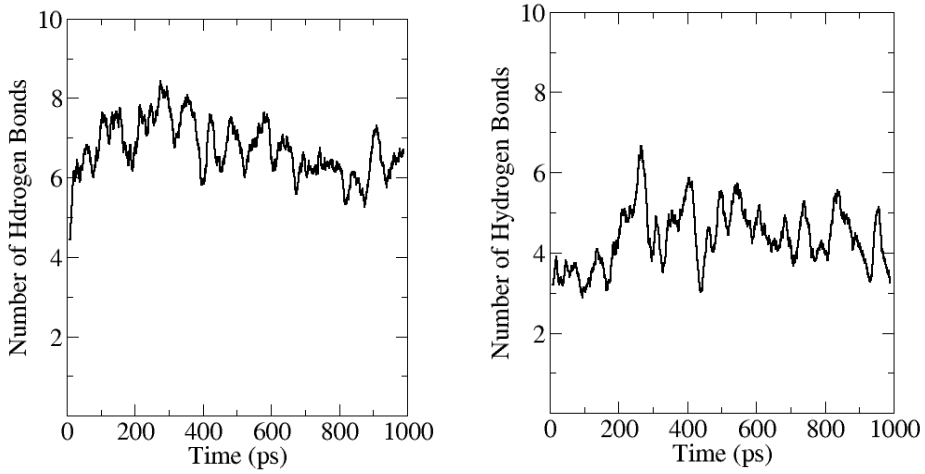


Figure 4.18: Free energy profiles for the adsorption of the antiDickerson dodecamer (see the text for its definition) on an APTES-functionalized silica surface with average surface density of $\Sigma = 0.314\ nm^{-2}$, as a function of the distance between the center-of-mass of the DNA molecule and the surface. The curves are: black: counterions only; red: 0.6 M of monovalent ions; green: 0.6 M of divalent ions. The uncertainties of the free energy have been estimated using the *bootstrap* technique [sec. (3.6.2)].

We have also investigated the dependence of the adsorption free energy on the base-sequence of the DNA. Visual inspection of the MD trajectories indicates that the Dickerson dodecamer interacting with the APTES-functionalized surface shows a pronounced tendency to form hydrogen bonds between the terminal bases (cytosine and guanine) and the -NH^+ groups on the surface. Since guanine and cytosine are characterized by the presence of three hydrogen bonds, we investigated whether substituting them with adenine and thymine would have an impact on the value of the adsorption free energy. To this end, we consider the anti-Dickerson dodecamer, which we define as a DNA oligomer whose sequence is ATATCCGGATAT, that is complementary to that of the Dickerson dodecamer. This oligomer exposes

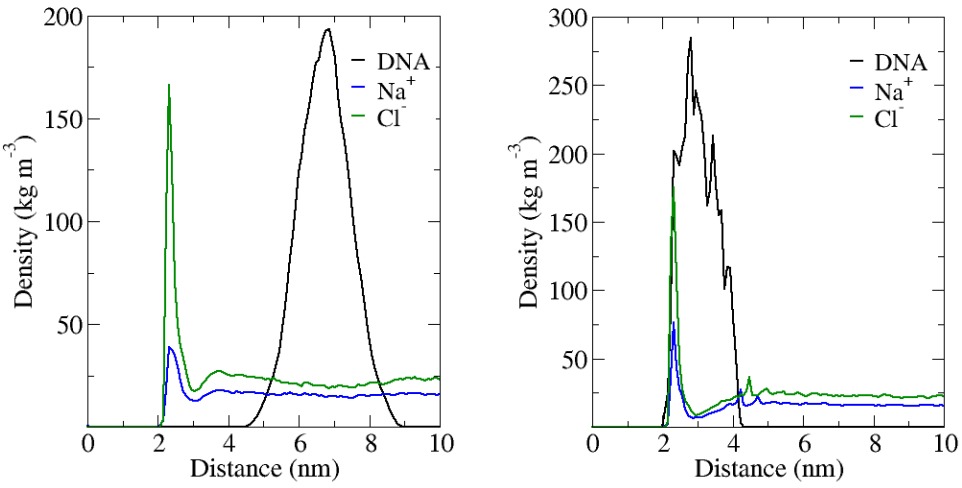
only adenine and thymine at both ending points, and can form only four hydrogen bonds with the surface, instead of the six that are possible when cytosine or guanine are exposed. The results for the free-energy calculations are shown in Fig. 4.18. We notice that the absolute value of the free-energy of adsorption is smaller than in the Dickerson dodecamer case. This is due to the different number of hydrogen bonds between the terminal DNA pairs and the charged groups on the APTES molecules, as shown in Fig. 4.19; there we give the running average of the number of hydrogen bonds for the Dickerson and anti-Dickerson dodecamers, in the molecular dynamics simulation. The analysis of these MD trajectories shows that an adsorbed Dickerson dodecamer forms 7±2 hydrogen bonds with the surface, whereas an anti-Dickerson dodecamer forms 4±2 hydrogen bonds, in presence of a 0.6 M solution of divalent ions. We notice that the adsorption free-energy of an anti-Dickerson dodecamer under these conditions is $\sim$ 10 kJ mol^{-1} , which is comparable to the value of the Boltzmann factor at room temperature, i.e. $k_BT \sim 2.478$ kJ mol^{-1}. This makes a desorption of the molecule quite likely.



(a) Dickerson (b) anti-Dickerson

Figure 4.19: Running average of the number of hydrogen-bonds in the molecular dynamics simulation. The amplitude of the sliding interval is of 100 ps.

The analysis of the MD trajectories shows that the origin of the variation of the free energy of adsorption as a function of the ion density is the different

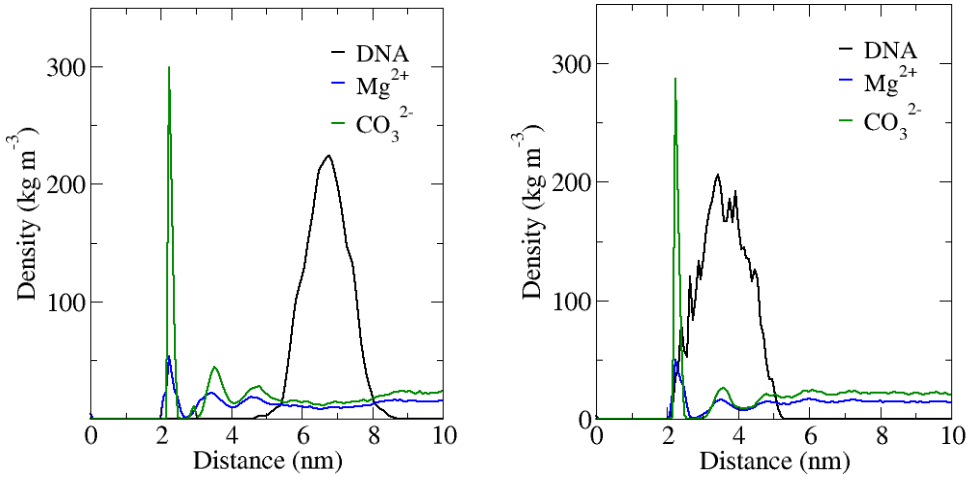


(a) (b)

Figure 4.20: DNA and monovalent ions mass densities for a Dickerson dodecamer constrained to reside far [panel (a)] and close [panel (b)] to the functionalized surface. Black line: DNA; blue line: monovalent cations (+) ; green line: monovalent anions (-). The solution is 0.6 M.

screening of the electrostatic interaction. We report in Fig. 4.20 the densities of the monovalent ions, when the DNA dodecamer is constrained to be either far or close to the surface. In both cases the negative ions accumulate close to the positively charged surface, in agreement with the results obtained using classical DFT calculations (compare with Fig. 4.13). At variance with the DFT results, one sees that in the atomistic model the positively charged ions also display a tendency to condensate on the functionalized surface. The density profiles for divalent ions, reported in Fig. 4.21, show a qualitatively similar picture. Both ionic species are adsorbed on the surface. However, there are a two important differences with the monovalent case worth pointing out.

First, the charge density of the negative divalent ions close to the surface is about two times higher than the density observed in the monovalent case (Fig. 4.21(a) and (Fig. 4.20(a))). This shows that the electrostatic screening is more effective with multivalent ions, as already evidenced from the discussion of classical DFT results, and from the analysis of the free energy profiles of the atomistic model.

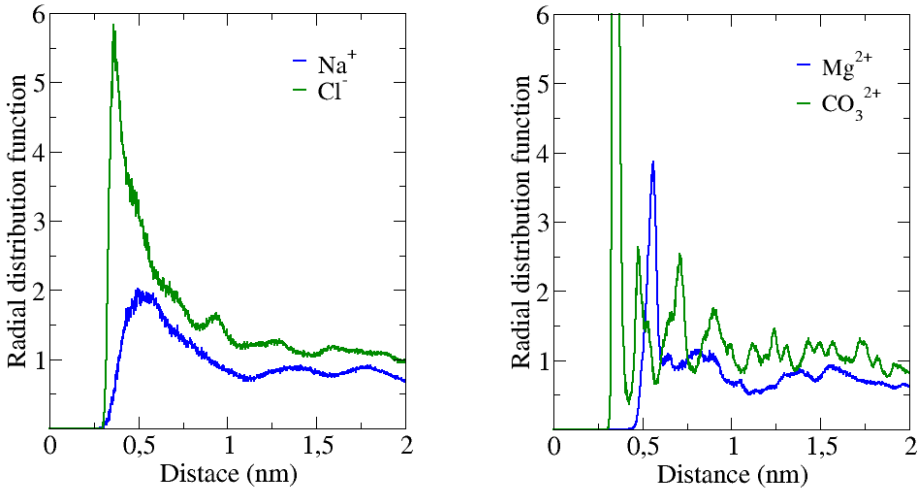


(a) (b)

Figure 4.21: DNA and divalent ions mass densities for a Dickerson dodecamer constrained to reside far [panel (a)] and close [panel (b)] to the functionalized surface. Black line: DNA; blue line: divalent cations (+); green line: divalent anions (-). The solution is 0.6 M.

Secondly, the ionic charge density close to the surface shows a significantly layered structure. When the DNA is far from the surface, one clearly sees the presence of three layers of divalent ions (Fig. 4.21(a)). When the DNA is adsorbed on the surface, this layered structure remains, although the number of measurable layers changes from three to two (Fig. 4.21(b)). The presence of these layers show that there are strong correlations between the ionic positions. We notice that the classical DFT model does not give an indication of this layering structure (see Fig. 4.13). This can be due to the inherent approximations used in developing the classical DFT model: despite the fact that the functional used in the calculation has been developed to include ion-ion correlations, the MD results show that it does not succeed in describing them to full extent.

Another important difference between the results obtained using the DFT approach and the atomistic simulation concerns the detail of the ionic distribution on the surface. In the continuum DFT approach the negative ions tend to accumulate close to the positively charged surface, whereas the positive ions are repelled from it, as can be seen from Figures 4.13 and



(a) (b)

Figure 4.22: Radial distribution functions of the monovalent [panel (a)] and divalent [panel (b)] ions around the charged group centered on -NH_3^+ in the APTES molecule, with a 0.6 M concentration of ions. Blue: cations (+); green: anions (-).

4.14, respectively. On the other hand, the density plots coming from the atomistic simulation, reported in Figures 4.20 and 4.21, show that both the positive and negative ions tend to accumulate close to the positive surface. The origin of this difference lies in the fact that in the atomistic simulation the negative ions tend to accumulate close to the positive APTES, therefore screening their electrostatic interaction, so that the positive ions can also be adsorbed on the surface. To show this, we report in Figure 4.22 the radial distribution function of the dissolved ions around the -NH_3^+ of the APTES molecules, in the case of monovalent and divalent dissolved salts with a concentration of 0.6 M. The radial distribution function $g_{AB}(r)$ is defined in such a way that $\rho_B\ g_{AB}(r)$ is the actual density of the particles of species B at a distance r from species A, where ρ_B is the average density of species B within the simulation box.

We see that generally the anions are found closer to the positive APTES group than the cations, thereby screening their electrostatic interaction. In the case of divalent ions, the radial distribution functions show well defined layered ionic structures. Inspection of the MD trajectories shows that in

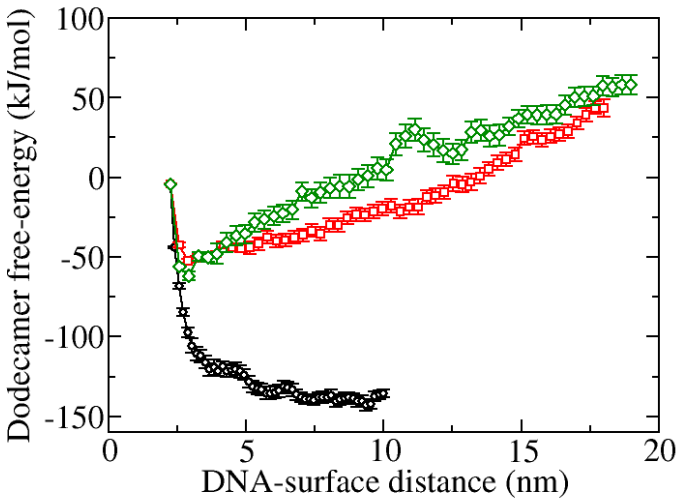
this case the ions tend to crystallize on the surface, occupying well defined positions (within the duration of our simulations). On the other hand, when only monovalent ions are present, they are seen to diffuse more easily along the surface while adsorbed on it.

However, comparison with explicit water calculations showed that the approximations involved in the implicit solvent model are too drastic to study DNA adsorption, and we therefore resorted in our calculations to fully explicit water molecules.

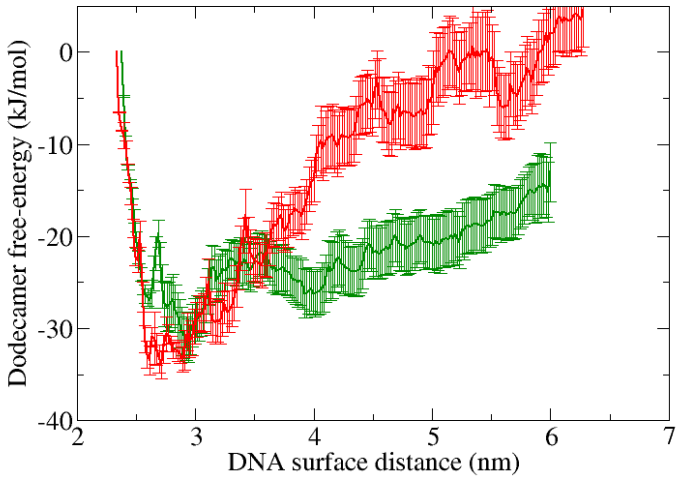
4.3.2 Explicit solvent model

The approximations involved in the implicit solvent model appear to be too drastic to study DNA adsorption under the most general conditions: while it reproduces some effects, the model does not yield a repulsion between the functionalized surface and the DNA, which has been observed, even though not in the same conditions as in the present work [92]. Consequently, as the results obtained with this model cannot be conclusive, we have extended our simulation by including explicit water molecules, which we represent via the widely used SPC/E model. In this way the number of degrees of freedom of the system increases considerably, and longer trajectories are necessary to calculate the free-energy profiles with umbrella sampling.

We have generated umbrella configurations pulling the DNA molecule using a quite strong harmonic potential: the elastic constant value is in the range $k \sim 200 - 500$ kJ mol^{-1}nm^{-2}, depending on the conditions of the system (molarity of the solution and charge on the surface). When water molecules are considered explicitly, we can use a stronger force costant than in the case of implicit solvent, without having a blow-up or a deformation of the DNA; this has been checked by visualizing the trajectories and computing the root mean square displacement. As in the case of implicit water simulation, we have extracted from this pulled trajectory a number of configurations, uniformly sampling the distance between DNA and the surface at intervals of $\Delta z = 0.2$ nm. To restrain the center-of-mass of DNA we have used a harmonic potential having an elastic constant $k = 1000$ kJ mol^{-1}nm^{-2}. At variance with the implicit solvent simulations, in this case we have generated the trajectories of 5 ns after 1 ns of equilibration, and we have treated the electrostatic force with the Reaction-Field method, which is a compromise between using the Particle Mesh Ewald



(a) Concentration: 0.1 M. Black: $\Sigma = 0.1\,|e|$ nm^{-2}; red: $\Sigma = 0.314\,|e|$ nm^{-2}; green: $\Sigma = 0.4\,|e|$ nm^{-2}.



(b) Concentration: 0.3 M. Green: $\Sigma = 0.2\,|e|$ nm^{-2}, red: $\Sigma = 0.314\,|e|$ nm^{-2}.

Figure 4.23: Free energy profiles for the adsorption of the Dickerson dodecamer on an APTES-functionalized silica surface, as a function of the average surface charge density Σ and for constant molar concentration of dissolved monovalent ions. The uncertainties of the free energy have been estimated using the *bootstrap* technique (sec. 3.6.2).

method or a cut-off in the whole box [103].

We report in Fig. 4.23 the free-energy profiles as a function of the distance between the surface and the DNA molecule, calculated analyzing the umbrella sampling trajectories with the WHAM method. Notice that in this and in the following figure the free energy curves are drawn by assigning the zero value at the smallest distance between surface and DNA molecule. This is due to the fact that, at variance with the calculations in implicit water, this simulation with explicit water molecules did not allow to reach the asymptotic behaviour of the free energy far from the surface. Panel (a) and panel (b) in Fig. 4.23 refer to the Dickerson dodecamer for a fixed value of monovalent ions concentration, respectively $c = 0.1$ M and $c = 0.3$ M, and for different values of the charge on the surface. As expected, the interactions between DNA and the functionalized surface become stronger increasing the charge on the surface; in our model a stronger charge on the surface is equivalently to a lower pH value. On the other hand, a higher molarity produces a stronger shielding between surface and DNA molecule, and thus a smaller interaction free energy. This, which is also an expected effect, can be see by a comparison between the red curves of the two panels, even though a conclusive comparison would require matching the asymptotic behaviour of the two curves at large distances.

In the simulation with explicit water molecules, an important feature we can observe is the complete detachment of DNA from the surface. This effect cannot be observed using the DFT calculations, which always produce one free energy minimum near the substrate. On the other hand the implicit water model not only does not allow a complete detachment, but drives the molecule towards the substrate when the latter is not shielded by ions; this might be due to the neglect of entropic effects that would hinder the docking in the real case.

We have also performed MD simulations varying the concentration of the dissolved ions, while keeping a constant value of the surface charge. The results (see Fig. 4.24) show that the interaction between DNA and the surface can indeed be changed by varying the concentration of dissolved ions: keeping a costant surface charge we see that the adsorption free energy can be reduced by increasing the concentration of dissolved ions. This result is in qualitative agreement with the results obtained with DFT and implicit water simulation.

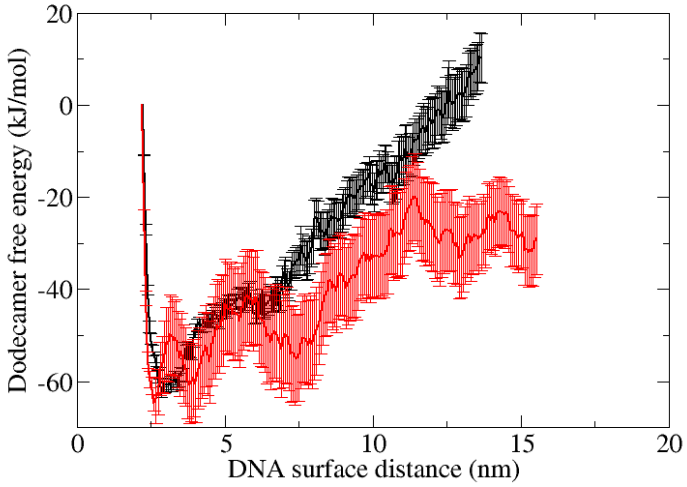
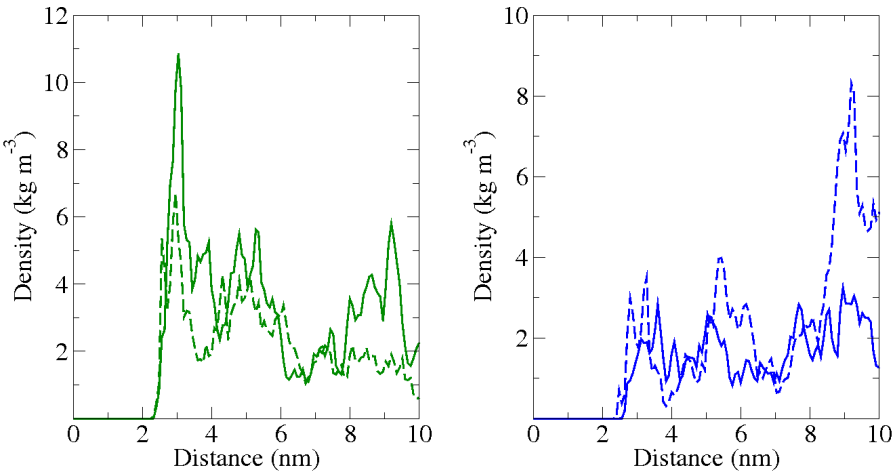


Figure 4.24: Free energy profiles for the adsorption of the Dickerson dodecamer on an APTES-functionalized silica surface with average surface density of $\Sigma = 0.314\,|e|$ nm^{-2}. The curves refer to different monovalent ion concentrations. Black: 0.1 M, red: 0.2 M. The uncertainties of the free energy have been estimated using the *bootstrap* technique (sec. 3.6.2).

The analysis of the MD trajectories shows that the origin of the variation of the free energy of adsorption as a function of the ion density is the different screening of the electrostatic interactions. As an example, in Fig. 4.25 we represent the charge density of monovalent ions in the case of fixed concentration of the solution 0.1 M and varying charge on the surface. As in the implicit water simulation, the negative ions are accumulated near the positively charged surface and their density is dependent on the charge of the surface (see Fig. 4.25(a)). The positive ions also are near the surface (see Fig. 4.25(b)) because they are clustered around the negative ones (see Fig. 4.26), but in this case the density is not strongly dependent on the charge on the surface.



(a) Chlorine (-) (b) Sodium (+)

Figure 4.25: The density of the ions is calculated at fixed concentration 0.1 M, varying the charge on the surface. Dashed line: $\Sigma = 0.314\,|e|$ nm^{-2}; continuous line: $\Sigma = 0.4\,|e|$ nm^{-2}. The DNA molecule is more than 10 nm far from the surface.

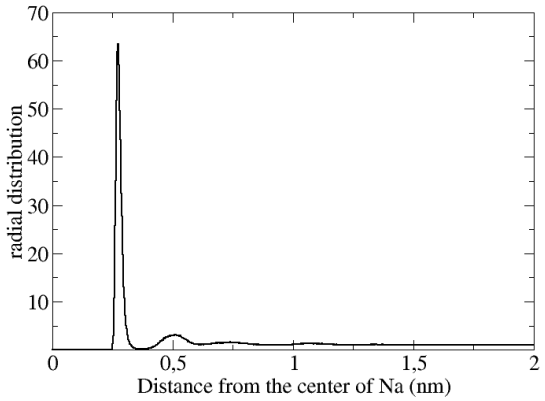
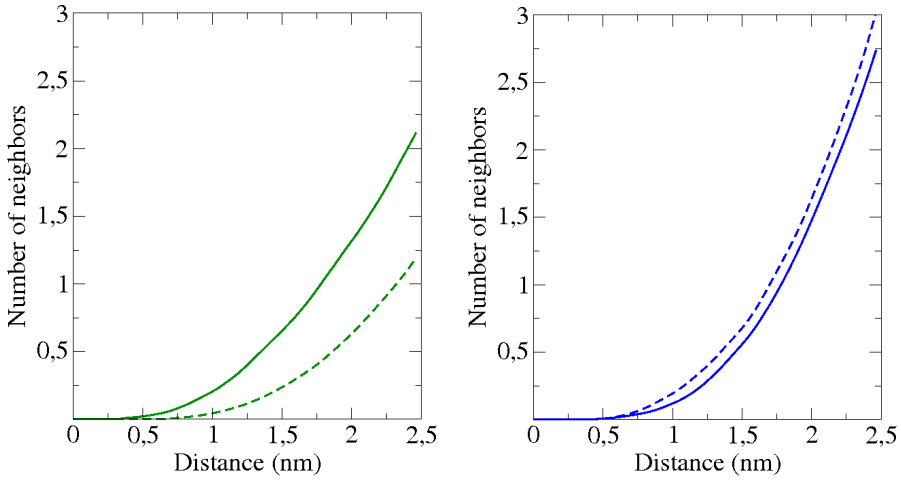


Figure 4.26: Radial distribution function of Cl^- ions around the Na^+ ions. The first maximum is about at ~ 0.3 nm; the second one is at ~ 0.5 nm.

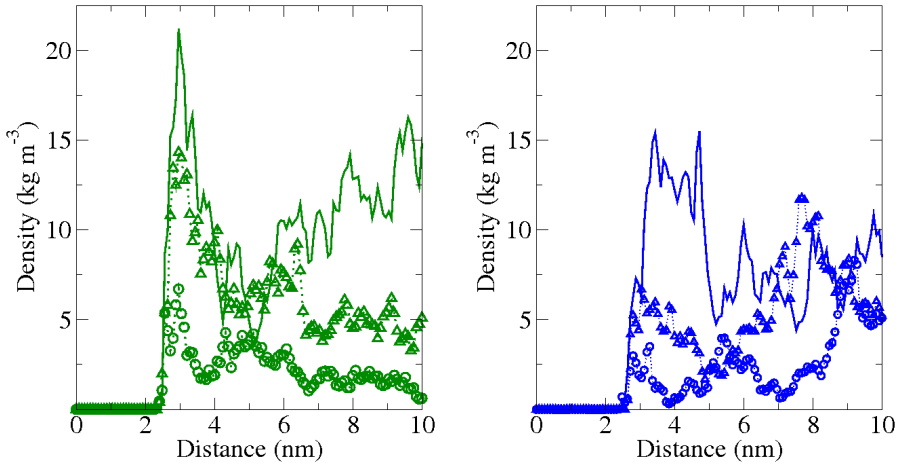


(a) Chlorine (-) (b) Sodium (+)

Figure 4.27: Cumulative radial distribution of dissolved ions relative to the -NH_3^+ moiety of APTES when the DNA is close to the surface, for a fixed concentration 0.1M of dissolved ions. Panel (a): chlorine (-); panel (b): sodium (+). The dashed and solid lines correspond to a surface charge of $\Sigma = 0.1\,|e|$ nm^{-2} and $\Sigma = 0.4\,|e|$ nm^{-2}, respectively.

In Fig. 4.27 we also show the cumulative radial distribution function of the dissolved ions, relative to the positive charge on the APTES moiety, for a fixed value of the concentration and for two surface charges. In the case of the smaller surface charge, the DNA is not adsorbed, whereas it is adsorbed on the surface for larger charges (Fig. 4.23). We notice that when the surface charge is increased the probability of finding negative ions close to the surface also increases. This enhanced distribution of negative charge screens the interaction between the DNA and the surface, but is not sufficient to compensate for the stronger attraction between surface and DNA molecule, and thus does not produce desorption.

The screening effect increases with the charge on the surface but also with the concentration of ions in solution (see Fig. 4.28): the greater the concentration the higher the screening effect between DNA and the functionalized surface.



(a) Chlorine (-) (b) Sodium (+)

Figure 4.28: The density of the ions is calculated at fixed charge on the surface $\Sigma = 0.314\,|e|$ nm^{-2} and varying the concentration of ions in solution. Circles: 0.1 M; triangle up: 0.2 M; continuous: 0.3 M. The DNA molecule is far from the surface.

4.3.3 The experiment by the BioSInt group

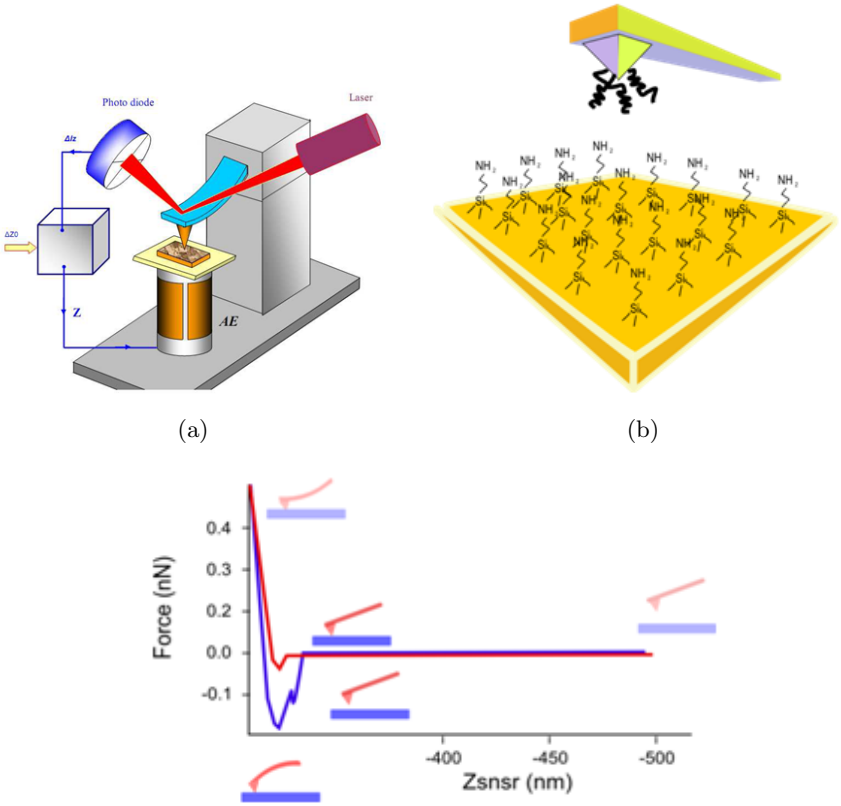


Figure 4.29: Schematic representation of the experimental setup. Panel (a): optical measurement of the deflection of the cantilever. Panel (b): microscopic view of the interacting molecular structures (DNA and APTES) at close distance. Panel (c): distance dependence of the force exerted on the cantilever when it approaches the surface (red curve) and when it is pulled away from it (blue curve).

The AFM force-distance experiments has been performed in liquid environment using AFM tips functionalized with thiolated double-stranded DNAs (dsDNAs) and coated mica surfaces. The coating has been obtained by reaction of aminopropyltrimethoxysilane (APTMS) with the surface. Adhesion forces between the tips and APTMS functionalized mica surfaces have been recorded at varying pH, monitoring the AFM cantilever deflection while

pH	buffer
4.5	acetate
5.6	acetate
6.5	phosphate
7	phosphate
9.2	Tris-HCl
10.5	carbonate/bicarbonate

Table 4.3: Buffers used in force distance AFM experiments.

it retracts from the surface.

A Cypher AFM system (Asylum Research, SantaBarbara, CA, USA) has been used to record force-distance data, utilizing a flexure scanner system in closed-loop configuration, with a range of $40 \times 40 \times 5$ μm^3. Silicon Nitride gold coated cantilevers from Olympus Corporation (TR 400 PB) have been used either without functionalization or after depositing a DNA layer, using 21 base long dsDNAs sequences conjugated with a cyclic dithiol group (DTPA: dithiolphosphoramidite). Before use cantilevers were cleaned by an Argon plasma treatment, using a plasma cleaner PDC-32G (Harrick Scientific Corporation, New York USA) at 40 Watt for 1 minute. The dsDNA solution used for cantilever functionalization has been obtained mixing two complementary strands (5'-DTPA-TAAGTTTGAATGTCATTTCTT-3' carrying the thiol modification at its 5' end and its unmodified complementary sequence), heating up the solution to 95° C for 1 minute and then cooling down to allow sequence hybridization. Functionalized cantilevers were obtained by immersion in 1 μM dsDNAs solution in potassium phosphate buffer (1 M, pH =6.9) for 10 minutes and then extensively rinsed with potassium phosphate buffer.

Functionalized mica surfaces have been obtained immersing them in a APTMS 0.1% water solution for 10 minutes and then rinsing with ultrapure water to remove APTMS in excess. Force curves were acquired at 80 nm/sec using a droplet cantilever holder. The pH of the solution was varied from 4.5 to 10.5 using 20 mM ionic strength buffers, as reported in Table 4.3. Data acquisition and analysis have been performed with Igor 6.2 (Wavemetrics, Oregon, USA) Asylum Research routines.

At least two maps in different sample places, each containing 400 force curves, were acquired at every pH, typically over areas of 500×500 nm^2.

Finally force histograms have been computed from every map, using a bin width of 3 pN.

First, the roughness of APTMS-mica surfaces has beeen characterized with AFM measurements, obtaining a roughness RMS value of 0.14±0.04 nm over areas of 500 × 500 nm^2, to be compared with the value of 0.04 ± 0.01 nm that characterizes freshly cleaved mica surfaces. The morphology of APTMS-mica samples reveals that the treated surfaces are very uniform, where small features characterized by peaks with a height of 0.3 nm and an average distance of 10 nm are visible. Next, force-distance data have been acquired using these APTMS-mica substrates in different buffer conditions. When increasing the pH, the main expected effect is a corresponding decrease of the average charge of the nitrogen group of APTMS molecules (deprotonation) and thus of the interaction force. We report in Fig. 4.30 some typical examples of the adhesion force histograms that can be obtained with DNA tips (panel A) and with neutral gold tips (panel B) at pH 9.2 and 4.5. In panel C of the same figure we report the value of the average force F_{A}, computed as $F_{\mathrm{A}} = \sum_i F_i N_i / \sum_i N_i$, at different pH values when using gold coated tips (open triangles) and DNA tips (circles), where N_i represents the number of events characterized by an adhesion force F_i, and the error bar on F_A represents the distribution of events. It has to be noticed that the experiment was not designed to measure negative forces, which would arise in the case of desorption.

While in both cases a dependence of the average force on pH is detected, this trend is greatly enhanced when a DNA layer is added to the tip. At pH values greater than 9 the average force results to be around $30 - 40$ pN for both kinds of tip, while decreasing pH to 4.5 the average force rises to 200 pN using gold tips, and to over 700 pN with DNA tips.

In the same graph, a pH titration curve (dashed line) has been reported, fitting the experimental data obtained with DNA tips with the formula

$$F(pK; A, B) = A \left(1 - \frac{10^{-pK} \cdot 10^{pH}}{1 + 10^{-pK} \cdot 10^{pH}}\right) + B, \tag{4.4}$$

assuming that the main effect of the pH increase is the deprotonation of APTMS groups. In equation (4.4), A, B and pK are the fitting parameters.

The experimental trend of the DNA-APTMS interaction is well described by this fitting function, that gives a pK value (the definition of which is given in eq. 2.25) of 6.7 for APTMS groups on mica, similar to that found in [104]

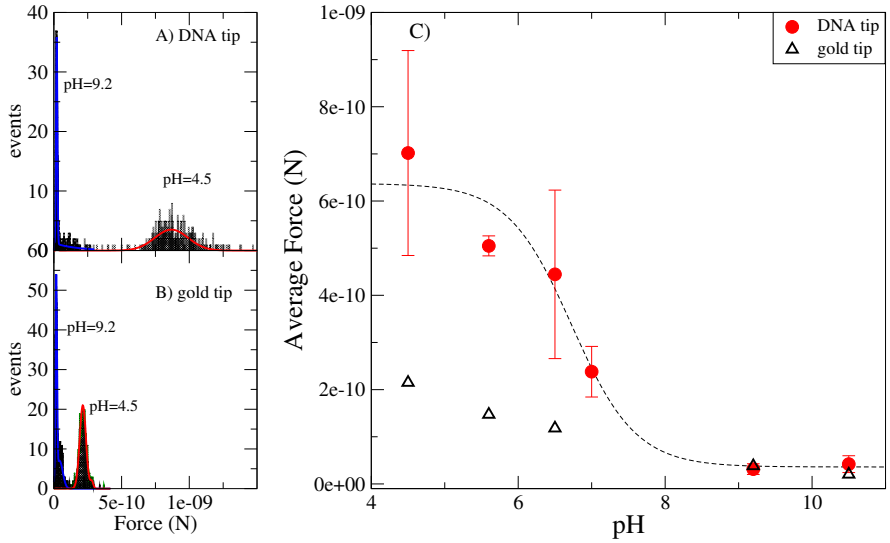


Figure 4.30: In Panels A and B some typical single force histograms are reported, obtained respectively with DNA tips at pH 4.5 and 9.2 and with gold tips at the same pH values. In panel C the pH dependence of the average adhesion force measured with DNA tips (circles) and gold tips (triangles), and the pH titration curve (dashed line) is shown. The average force is computed acquiring at least 800 force curves over two different sample areas of 500x500 nm^2.

for APTES on silicon substrates. These pK values represent a decrease of several units with respect to the value typical of free organic primary amine ($\sim 10-11$) at the same ionic concentration, and produce a decrease of the corresponding pH by several units.

These experimental results indicate that the force that drives the adhesion of DNAs on such functionalized surfaces has an important electrostatic contribution. Recently published results of experiments on DNA adhesion and release from APTES functionalized surfaces and microchip devices [92] are consistent with such interpretation.

4.4 Discussion and conclusions

In this part of thesis we have presented the results of computer simulations aimed at investigating the principal factors involved in the adsorption of

DNA oligomers on surfaces functionalized with positively charged groups. In particular, we have investigated the effect of the concentration and type of dissolved ions, and of the pH of the solution, on the free energy of adsorption.

We showed that a continuum model based on the classical DFT treatment of the electrostatic interaction is able to capture qualitatively the main features of the process. In a solution of monovalent ions, the electrostatic interaction between the surface and a negatively charged DNA oligomer is quite strong, and one can expect a significant adsorption to take place. In the presence of an increasing concentration of divalent ions the screening of the electrostatic interaction becomes stronger, and the free energy of adsorption is progressively reduced. However, in none of the cases investigated by DFT one sees an inversion of the electrostatic interaction between the surface and the DNA (opposite charge repulsion), which could yield a desorption.

The simulation with the atomistic model based on the implicit solvent approach partially confirmed these results: the free energy profiles of adsorption show a more effective screening of the electrostatic interaction by increasing the salt concentration, with the divalent ions being considerably more effective than the monovalent ones. The atomistic description of the latter model allowed us to investigate in some detail the ion distribution in the system. We showed that, at variance with the indications coming from the classical DFT model, both positive and negative ions are adsorbed on the surface. As it might have been expected, the negative ions screen the charge of the functionalizing amine groups allowing adsorption of the positive ones. In the case of divalent ions we found evidence of a layered ionic distribution near the surface, especially at the highest concentrations that have been considered.

We want to point out another interesting result obtained with the atomistic approach: we were able to establish that the sequence of bases of a DNA oligomer influences the free energy of adsorption. In particular, the terminal bases are relevant because of their possibility of forming partial hydrogen bonds with the charged groups on the functionalized surface. As a consequence, oligomers with terminal CG pairs have a larger adsorption free energy than oligomers having a higher terminal density of AT pairs.

Finally, the atomistic simulation using the explicit water model has shown (see Fig. 4.23(a)) that a desorption of the DNA molecule can indeed take place, given the right values of ionic concentration (0.1 M) and

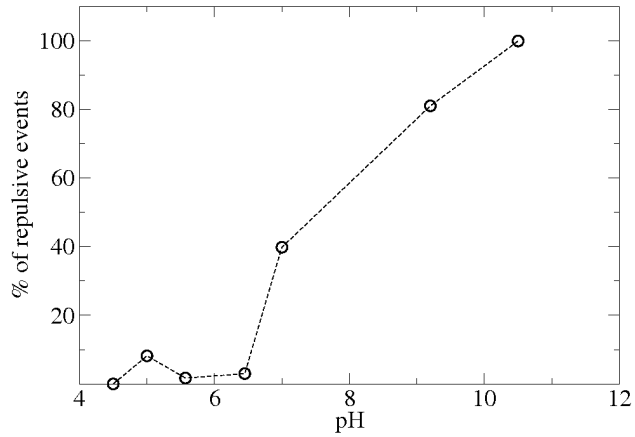
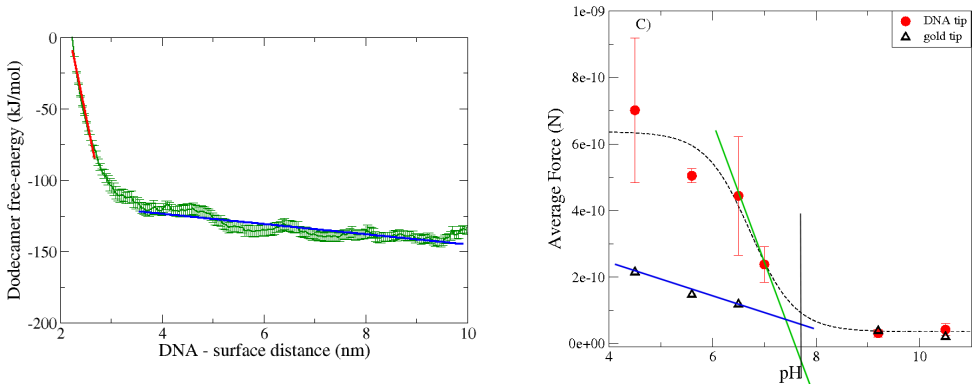


Figure 4.31: Average number of recorded events of repulsive force between the functionalized surface and DNA, as function of pH.

charge density on the functionalized surface (sigma = 0.1 e nm^{-2}). Neither the DFT approximation nor our simulation in implicit water were able to produce this effect. This is possibly related to entropic effects, that are not taken into account in the DFT calculation, and only partially accounted for in the simulation with implicit water. In the latter the DNA molecule is flexible, but the solvent is treated in a mean field approximation, which levels its microscopic interaction with the DNA oligomers, lowering the fluctuations and hence the entropy of the system.

Due to the relevance of the desorption phenomenon, also - as explained at the beginning of the chapter - from the point of view of its application to the purification process, we have looked for experimental data that could confirm our findings. We could not find in the literature results obtained for the same system and/or under the same conditions as ours. The closest results are those of the BioSInt group described before for APTMS. Looking at Fig. 4.30 one sees that increasing pH values correspond to decreasing attraction forces between the functionalized surface and the DNA molecule. The surface charge density at which we find the desorption corresponds to a pH = 7.65. Although occurences of repulsive force between the functionalized surface and DNA have been recorded (see Fig. 4.31), as the experimental setup does not allow a measure of negative forces corresponding to the desorption process, a direct comparison with experiment is not possible. Nevertheless, we have qualitatively checked the reliability of

our simulation in the following way. The experiment has been performed at the molarity 0.02 M, which we compare with our lowest, namely 0.1 M. We have computed average desorption forces deriving two segments of the black free energy curve of Fig. 4.23(a) as a function of the distance between DNA molecule and surface: $F = -df/dx$. The first segment corresponds to the almost linear decrease between 2 and 3 nm, the second segment corresponds to the almost linear decrease between 4 and 10 nm 4.32(a). In this way we obtain two limits for the force, which turn out to be, respectively, $-3.0 \cdot 10^{-10}$ N and $-0.02 \cdot 10^{-10}$ N. In order to make a rough comparison with experiment, we have to infer from Fig. 4.30 (c) the value of the force that would have been measured at pH = 7.65 if the apparatus had been able to record negative values. We do this by linearly extrapolating the two experimental data found just below the desired pH value (Fig. 4.32(b), green line) which yields a negative force of about $-0.5 \cdot 10^{-10}$ N (intersection of the green and of the vertical black lines). Correcting this value to take into account the contribution to the force due to the pure gold tip (blue line), one finds a negative force of about $-1.2 \cdot 10^{-10}$ N. This value falls right in the middle of the interval between the two limits estimated for the force in the atomistic simulation; given our interpolation, extrapolation, and further approximations of our model, this correspondence is surprising and encouraging at the same time. It also means that our simulation can be a valid tool to explore future applications of this phenomenon.



(a) PMF

(b) Experimental result

Figure 4.32: Estimates of the mean desorption force at pH = 7.65. Panel (a) Maximum (red interpolation) and minimum (blue interpolation) values: see text. Panel (b) Extrapolation of force values (green line); background contribution of the gold tip (blue line); intersection at pH = 7.65 (vertical black line).

Chapter 5

Tau, an intrinsically disordered protein

5.1 Intrinsically disordered proteins

Intrinsically Disordered Proteins (IDPs), or Intrinsically Unstructured Proteins, or Natively Unfolded Proteins, are proteins that in their native state do not have an average stable structure but fluctuate between many conformations, and thus resemble highly denaturated globular proteins [105, 106]; IDPs are presumed today to constitute up to one fourth of all proteins [105]. Due to their flexibility, and at variance with the well-known lock and key biomolecular paradigm, they perform tasks that cannot be carried out by globular proteins [105, 107–110]. In some cases they can switch between apparently unrelated functions; this ability of IDPs is called moonlighting [111].

IDPs entail at least an extended disordered region, but can also entail globular domains alternating with flexible linkers or disordered domains. These proteins are therefore characterized by different degrees of disorder, from those formed by globular domains connected by disordered segments to those totally disordered [105, 106]. Even the latter may entail segments endowed, albeit temporarily, with secondary structures like α-helices, β-sheets, or PPII helices [112].

The main functions of IDPs are not structural, but regulatory: control, modulation, and signalling. Their biological function is characterized by an interaction energy among residues that is significantly lower than

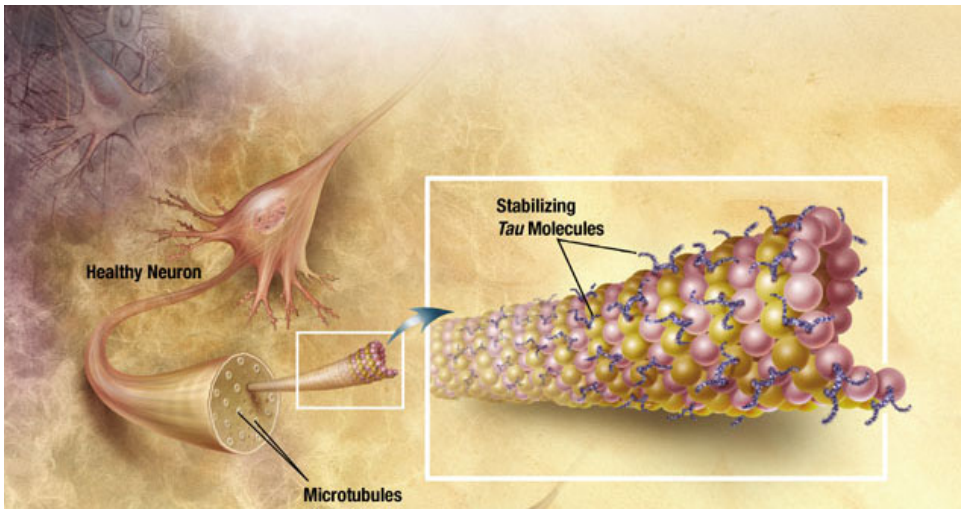
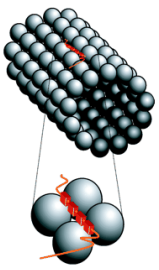


Figure 5.1: Structure of a healthy neuron. The tau protein, a microtubule-associated protein, play a large role in microtubule stability. http://www.highschoolbioethics.org/briefs/nfl.asp

for globular proteins [105, 110]; this favours fast shifts between extended conformations, that generally accompany the binding to other molecules, and disordered molten-globule-like conformations. The formation and dissolution of bound states is probably faster than in the case of globular proteins[105, 108].

The biochemical functions of IDPs are related to their structure, which varies in time. Given the speed at which transitions between conformations are supposed to take place, the simulation of their dynamics seems to be a promising tool to understand their behaviour. But the dynamical simulation of an IDP is a computational challenge, because by definition there are no experimentally determined 3D structures of the whole molecule, like a protein data bank (pdb) file, from which to start. In Section 5.3 we describe a procedure to overcome this obstacle.

5.2 The tau protein



J. Avila *et al.*, Physical Rev. **84,** 361-384 (2004)

The tau protein, one of the largest totally disordered IDPs [107], is involved in the nucleation and stabilization of the microtubules (MTs) in the ax-

ons of the neurons (see Fig. 5.1). Stabilization is achieved through the bonding of the repeats domain (see Tables 5.1 and 5.2) of tau to the α- and β-tubulines forming the MTs [113].

But the same tau can aggregate in paired helical filaments (PHFs) and form fibrils which, in their turn, form insoluble tangles as shown in Fig. 5.2 [105–107, 113, 114]. The aggregation of tau proteins prevents them from carrying on their physiological stabilization role of the MTs and this pathological deviation from its physiological function - together with other factors - is at the origin of the development of the Alzheimer disease and of other neurodegenerative diseases [113].

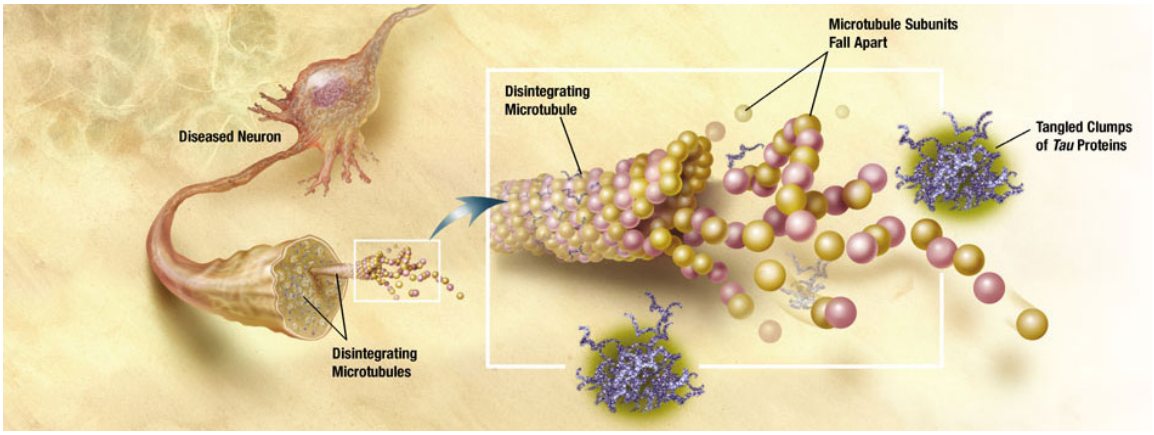
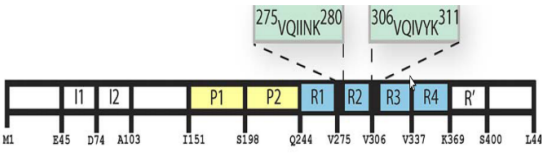


Figure 5.2: Structure of a diseased neuron. The tau proteins are chemically altered, forming tangles by coupling with other tau proteins. These altered tau structures collapse causing microtubules to disintegrate and the cells to die. http://www.highschoolbioethics.org/briefs/nfl.asp

The process of formation of the PHFs is not entirely known, but some of its factors and stages have been investigated. A precursor stage of tau polymerization has been related to specific conformations of the protein, in which the C terminal is in the neighbourhood of the repeats domain that binds to a MT, and the N terminal is folded near the C terminal. The pathological aggregation in the form of amyloids has been attributed to a local transition from the unfolded state to a β-structure [112, 115–117]. The aggregation process is supposed to start from a nucleus entailing the VQIVYK motif, a tau segment with high propensity for a β structure [116], or the VQIINK motif.

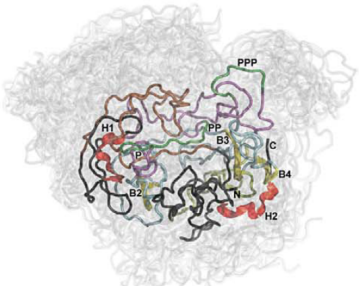


M.D. Mukrasch *et al.*, PLoS Biology **7**, 399-414 (2009)

These two hexapeptides are, respectively, at the beginning of the third and of the second repeat [112, 116, 118], and have been identified as components of steric zippers formed by β-sheets parallel to the axis of the fibril [117]. Also the propensity of a polyproline-II motif toward the formation of β-sheets has been listed among the possible causes of the aggregation of tau proteins, and hence of the origin of tau-pathologies [108, 119].

The gyration radius R_g (see eq. (5.1)) of a molecule measures the average size of its overall conformation. The R_g of tau has been measured by Small-Angle X-ray Scattering (SAXS), and on average $R_g = 6.6$ nm [109]. This value may be compared to that of α-amylase [1], a globular protein that has 448 residues and a gyration radius $R_g = 2.3$ nm. If tau were a globular protein with a spherical shape, its volume of 56 nm^3 would imply a radius of 2.4 nm. The larger R_g of tau is due to its non-globular structure, and its R_g is about that of a random coil of a polypeptide with the same number of residues (6.9 nm [109]). Nevertheless, when R_g is measured in partial domains of tau that entail the repeats, its value turns out to be larger than the value estimated for a random coil; this hints at a propensity of these domains to acquire an extended structure [109, 120]. Because these structures would very likely be transient, there is a definite interest in a dynamical simulation of tau, in order to gather information on the existence and probability of segments of tau endowed with a secondary structure.

Tau exists in several isoforms [121]; we have chosen to study with MD simulation its htau40 isoform which is found in the human central nervous system; it has 441 residues (see Tab. 5.1) and a molecular weight of 45.85 kDa [114]. As said before, the dynamical simulation of an IDP is challenging because there are no available structures of the whole molecule from which to start. Up till now crystallography has yielded only the structure of short segments of tau in the following cases: as



M.D. Mukrasch *et al.*, PLoS Biology **7**, 0399-0414 (2009).

[1] For α-amylase the pdb entry is 1AQH.

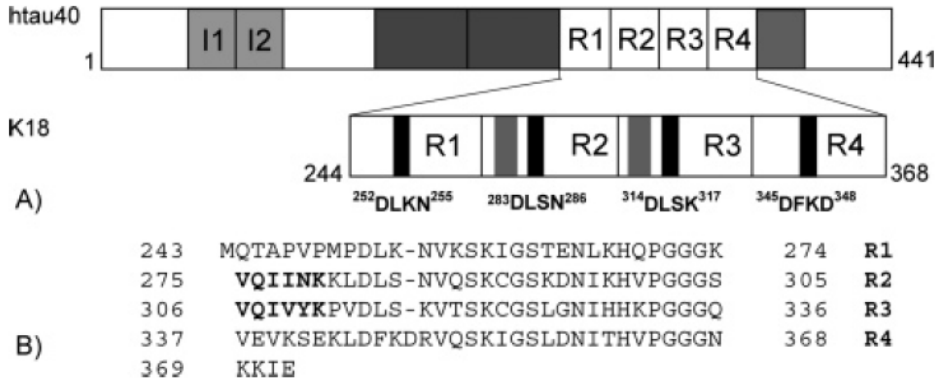


Figure 5.3: Tau protein and K18. (A) Position of the K18 construct with respect to the htau40 isoform of tau. Tau contains four pseudorepeats (R1–R4) involved in binding microtubules and forming the core of PHFs. (B) Primary sequence of the K18 construct, showing the position of repeats 1-4, containing 30-31 residues each [112].

a free molecule [2], or complexed with another structure [110], or in a docking process [122]. There are no available structures of the whole free tau.

A static simulation has been previously used to study the 3D structure of the K18 construct, a region of tau containing the four repeats, as shown in Fig. 5.3; this region of about 120 residues is known to be involved in the binding to microtubules, but also in the formation of PHFs. This simulation has been done using the Flexible-Meccano method, a statistical sampling from structural databases of folded proteins, combined with a biased-potential dynamical simulation, to produce thousands of conformers; the ensemble is then filtered by an Ensemble Optimization Method (EOM) that uses NMR experimental data [109, 112]. A series of static structures of the whole tau has been created using the XPLOR-NIH package (NMR molecular structure determination package); they correspond to energy minima complying with structural energy data from steric repulsion, bond lengths, bond angles, and dihedral angles [118]. An extended dynamical simulation of tau in water has been performed only for a segment of 19 residues, assuming an α-helix structure of the segment as the initial configuration of

[2]For the hexapeptide VQIVYK the pdb entries are 2ON9 and 3FQP.

10	20	30	40	50	60	
‖	‖	‖	‖	‖	‖	
MAEPRQEFEV	MEDHAGTYGL	GDRKDQGGYT	MHQDQEGDTD	AGLKESPLQT	PTEDGSEEPG	- 60
SETSDAKSTP	TAEDVTAPLV	DEGAPGKQAA	AQPHTEIPEG	TTAEEAGIGD	TPSLEDEAAG	- 120
HVTQARMVSK	SKDGTGSDDK	KAKGADGKTK	IATPRGAAPP	GQKGQANATR	IPAKTPPAPK	- 180
TPPSSGEPPK	SGDRSGYSSP	GSPGTPGSRS	RTPSLPTPPT	REPKKVAVVR	TPPKSPSSAK	- 240
SRLQTAPVPM	PDLKNVKSKI	GSTENLKHQP	GGGKVQIINK	KLDLSNVQSK	CGSKDNIKHV	- 300
PGGGSVQIVY	KPVDLSKVTS	KCGSLGNIHH	KPGGGQVEVK	SEKLDFKDRV	QSKIGSLDNI	- 360
THVPGGGNKK	IETHKLTFRE	NAKAKTDHGA	EIVYKSPVVS	GDTSPRHLSN	VSSTGSIDMV	- 420
DSPQLATLAD	EVSASLAKQG	L				

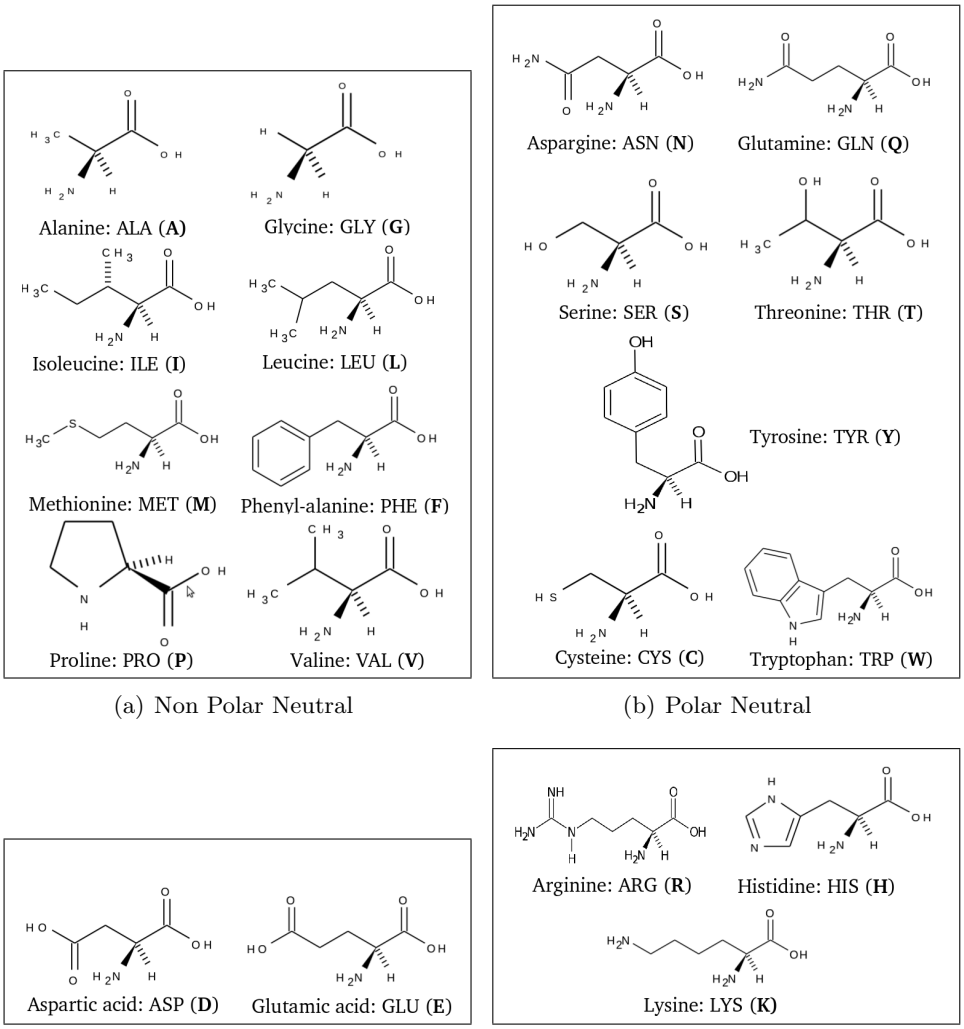
Table 5.1: Primary sequence of htau40.

the simulation, and studying its stability [123].

5.3 The dynamical simulation

In order to produce a generic 3D structure of the whole htau40 to start a Molecular Dynamics (MD) simulation, we have implemented a protocol which requires only some hundred picoseconds of dynamical simulation [124]. Before describing our method we want to highlight a second problem in the simulation of an IDP, namely the choice of a suitable force field. Molecular mechanics force fields have been parametrized on folded protein structures, and therefore may not correctly reproduce the structure of disordered proteins. On the other hand, there is no alternative to the use of one of the known force fields, because an *ab initio* calculation of a large molecule like tau is unfeasible. For this simulation we have chosen the ffG53a6 force field, implemented in the GROMACS package. The simulation has been carried out at neutral pH (pH = 7), close the physiological value (that is in the 7.2 - 7.4 range). Accordingly, amino acids were set to their default protonation states at pH = 7, i.e., with Lys, Arg carrying a +1 and Glu, Asp a −1 net charge.

In order to produce a generic 3D structure of the tau protein we start from its primary sequence of aminoacids, given in Tab. 5.1. The correspondence between the amino acids (AA) symbols and the relevant chemical structures is given in Fig. 5.4.



(a) Non Polar Neutral

(b) Polar Neutral

(c) Polar Acid

(d) Polar Basic

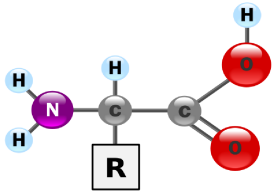


Figure 5.4: Structure of 20 amino acids (AA) encoded by the universal genetic code. The structures shown have a group specific to each AA, generally referred to as **R** group. The carbon atom next to the carboxyl group ($-COOH$) is called the α-carbon (C_α). The amino group ($-NH_2$) is attached to the C_α. In all structure the **R** group appear on the left of the ($-NH_2$).

We use this sequence as an input of the Visual Molecular Dynamics (VMD) program [12]; this program lines up the amino acids following their primary sequence, departing from a straight line only when obliged by stereochemical incompatibility of neighbouring amino acids. The output of VMD is thus a 3D sequence of straight segments of amino acids that bears little resemblance to a real protein, as shown in Fig. 5.5.

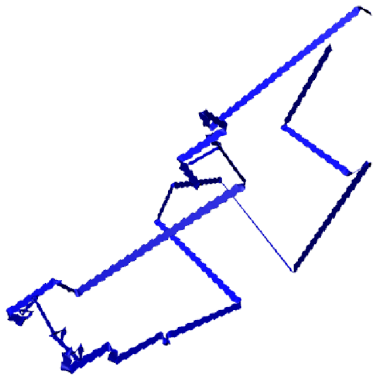


Figure 5.5: Initial shape of tau protein, as produced by the VMD program from the primary sequence.

When the MD simulation program GROMACS [125] is used to evolve this structure *in vacuo*, the molecule's configuration collapses in a short time. This can be monitored by measuring the gyration radius

$$R_g = \left(\sum_i r_i^2 m_i / \sum_i m_i \right)^{1/2} \tag{5.1}$$

where $\mathbf{r}_i$ are the positions of the atoms with respect to the centre of mass of the molecule, and m_i are their masses; R_g measures the overall size of a molecule. The curve #1 in Fig. 5.6 shows that a collapse of R_g, from a value of 10.4 nm to a value of 2.5 nm, takes place in about 100 ps, and yields a very compact and entangled configuration (Fig. 5.7). This fast evolution of the molecule is due to the absence of the solvent, which would prevent the collapse and the formation of a high number of intramolecular H-bonds that can be seen in Fig. 5.8 (curve #1).

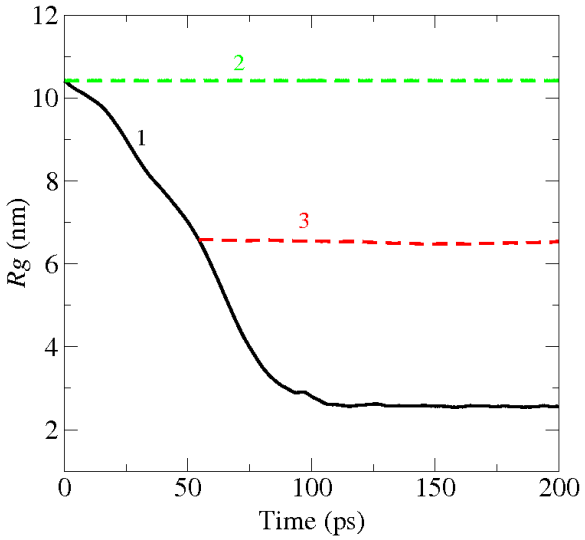


Figure 5.6: Evolution of the radius of gyration at $T = 300$ K in the first 200 ps, starting from the configuration of Fig. 5.5. The continuous black line (#1) shows the rapid collapse of the protein *in vacuo*. The dashed green line (#2) shows the evolution after addition of water molecules to the initial configuration of tau. The dashed red line (#3) shows the evolution after addition of water molecules to the conformation extracted at $t = 56$ ps and $R_g = 6.57$ nm.

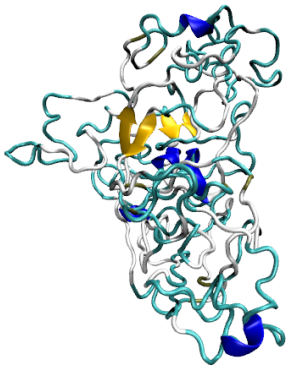


Figure 5.7: Collapsed shape of the tau protein after a 100 ps evolution *in vacuo* at $T = 300$ K. Two short β-sheets (yellow) and some short 3-helices (blue) are highlighted.

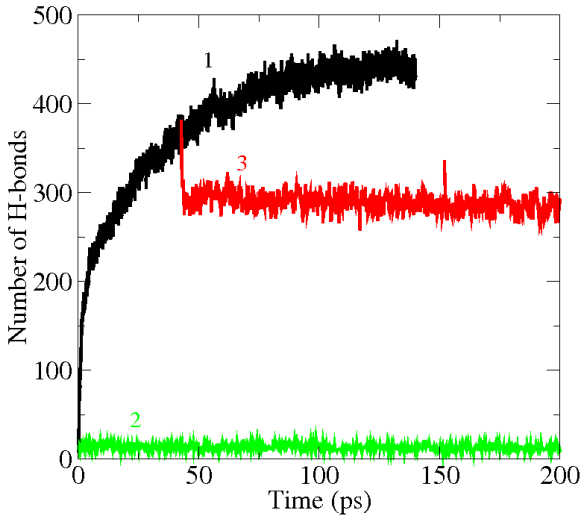


Figure 5.8: Time evolution of the number of intramolecular H-bonds at $T =$ 300 K, first 200 ps. The black line (#1) shows their rapid increase during the evolution *in vacuo*. The green line (#2) shows their slow change if water molecules are added to the initial configuration of tau. The red line (#3) shows a sharp drop when water molecules are added to the conformation extracted at $t = 56$ ps, and the stable subsequent evolution.

A straightforward way of letting the initial configuration evolve towards a native-like state of the molecule is to embed its initial structure in the solvent. The configuration produced by VMD from the primary sequence is very extended, and the simulation box has to be accordingly large. A totally disordered protein like the tau can be considered to be approximately a random coil, fluctuating in an ensemble of alternative conformations. Therefore, in a MD simulation of an IDP, one has to pay attention to the flexibility of the molecular structure: when a periodic box is used, its size must be large enough to avoid that during the dynamics the protein interacts with one of its periodic images, if it extends its shape to the region bordering the walls of the box. A large box entails a very large number of solvent molecules; this is a relevant obstacle for the workability of a simulation. Using the box-to-molecule size relation usually adopted in this kind of MD simulation, this means in the present case that one has to use a box filled with about 1.5×10^6 water molecules. Moreover, the evolution of the molecule from the initial configuration is very slow, as shown by curves #2 in Fig. 5.6 and in Fig. 5.8. Taking into account the experimental value for the average gyration radius $R_g = 6.57$ nm [109], one can foresee for the overall configuration of the molecule and for R_g an equilibration time in excess of several tens of ns. All in all, this would be a computationally very expensive procedure.

We propose an alternative and efficient way of creating a 3D structure of the protein. We start with the atomic positions produced by the VMD program, let them evolve dynamically *in vacuo*, and stop the evolution when the configuration has reached a value of the gyration radius equal to the experimental average $R_g = 6.57$ nm. The structure obtained in this way is then embedded in water, where the simulation box and hence the number of water molecules is significantly reduced (to about 1/3) with respect to the initial extended state. After a short minimization of the total energy, the system (tau + water molecules) is ready to start a dynamical evolution in a region of the phase space corresponding to realistic conformations of the molecule. The stabilizing effect of the introduction of explicit water on the dynamical evolution of the protein is a sudden interruption of the collapse of the molecule, and the beginning of a slow fluctuation of the structure; this can be seen in Fig. 5.6, curve #3. As for the intramolecular H-bonds, curve #3 of Fig. 5.8 shows that the introduction of the water molecules causes a sudden decrease in the number of those bonds, about one fourth of which is

replaced by H-bonds between tau and the water molecules. Fig. 5.9 shows more in detail this instantaneous decrease (curve #1), and the first 20 ps of a NVT simulation. After the introduction of the solvent, the molecule assumes in a short time a native-like configuration, as shown in Fig. 5.10; the structure displays short transient secondary structures like β-sheets and α-helices [126].

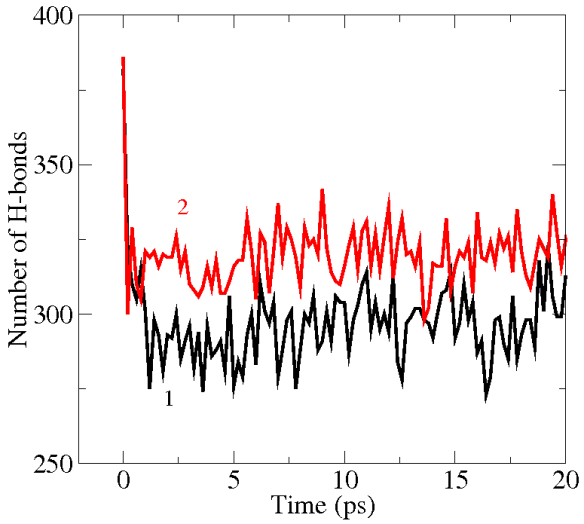


Figure 5.9: Time evolution of the number of intramolecular H-bonds at $T =$ 300 K, beginning at the time when solvent is introduced in the simulation. Black line (#1): explicit water molecules. Red line (#2): implicit water solvent.

The same procedure could be followed embedding either the initial structure of the molecule, or the one extracted midway during the collapse, in implicit water [127]. The use of implicit solvent seems to be a convenient choice for a simulation run beginning when the molecule has already shrunk to a natural size. On the other hand, it is known that the two types of solvent are not equivalent in the prediction of secondary structures [128, 129]. Indeed, as shown in Fig. 5.9, the effect of implicit water on the replacement of intramolecular H-bonds is not the same as that of explicit water molecules: the latter seem to be more efficient in competing with intramolecolar H-bonds and replacing them with solvent-solute ones. This indicates that after the evolution *in vacuo* of the molecule, possible spurious H-bonds, that would not have been formed if the solvent had been included in the simulation from the beginning of the dynamics, are better removed by putting the molecule

Figure 5.10: Shape of the tau protein after a 100 ps evolution, 56 ps *in vacuo* and 44 ps in explicit water, at $T = 300$ K. Two transient short β-sheets (yellow) and a transient short α-helix (purple) are highlighted.

in explicit solvent.

5.4 The simulation procedure

The method used to find an initial dynamical state of the tau protein can be used for any IDP. In order to start a Molecular Dynamics simulation of a protein of unknown 3D structure, one begins from its primary sequence of aminoacids. One then implements the following procedure to produce a 3D structure to start the simulation:

1. A first 3D structure is created by feeding the VMD program with the primary sequence of the whole protein.

2. The energy of this structure - a sequence of sticks - is shortly minimized to avoid stereochemical incompatibility.

3. The resulting structure is put in a large box with periodic boundary conditions, and taken as the initial one for a dynamical evolution *in vacuo* at the chosen temperature, performed with the package GROMACS or with any other simulation program; this step produces a rapid contraction of the protein.

4. The evolution is stopped when the decreasing gyration radius R_g has reached its average experimental value. This yields a starting point for the simulation in a more realistic environment, i.e., with the addition of solvent.

5. The simulation box is reduced to fit the reduced size of the protein, and filled with solvent. This allows a significant reduction of the volume of box, and hence of the number of solvent molecules needed to fill it.

6. The energy of the system (protein + solvent) is minimized to allow the water molecules to adapt to the shape of the solute molecule.

7. A short equilibration (100 ps) is performed at constant temperature.

8. Another short equilibration (100 ps) is performed at constant temperature and pressure.

9. The last conformation of the previous step is used to start an extended simulation at constant temperature and pressure.

The gyration radius used in step 4 is known for many molecules, being measured either by light scattering, or SAXS, or small angle neutron scattering. But one could instead use any other measure of the overall configuration of the molecule to monitor its evolution *in vacuo*, like asphericity, that combines shape and compactness, and is measured by fluorescence microscopy [130].

5.5 Domain patterns

In our first dynamical simulation of the complete tau (htau40), we have studied the time evolution of the molecule in water over a time of 30 ns. Fig. 5.11 shows the gyration radius; R_g is not stabilized around its experimental value, as it progressively decreases to about 4.3 nm. Even though the latter

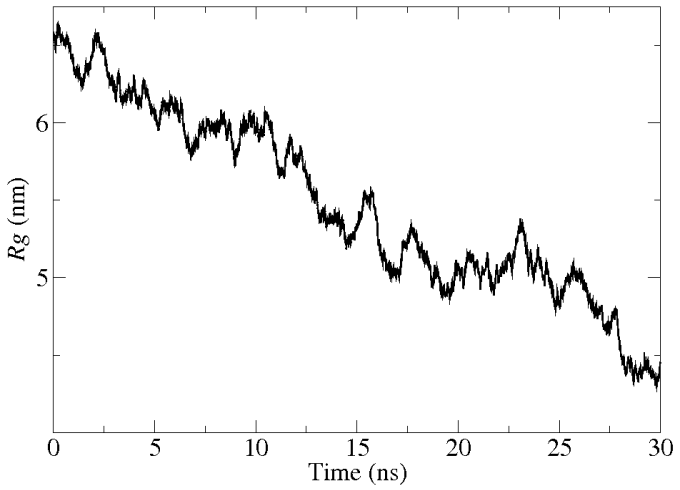


Figure 5.11: Time evolution of the radius of gyration of protein tau during the dynamics at $T = 300$ K. The experimental average value of R_g is 6.6 nm, the standard deviation 0.3 nm [109].

value is within the range of values computed from a set of static conformers of tau produced by the EOM method [109, 112], the continuous decrease of R_g hints at a possible shortcoming of the force field in reproducing the overall shape of the molecule. In order to clarify this dynamical behaviour we have computed the time evolution of the gyration radius of four domains corresponding to morphologically different sections of the molecule: the N-terminal projection domain (residues 1-150); a proline-rich segment (residues 151-243); the repeats domain (residues 244-368); the C-terminal domain (residues 369-441). The results are reported in Fig. 5.12, and show that all four domains reach an equilibrium stage: first the C-terminal domain, after about 10 ns; second the repeats domain, shortly before 20 ns; then the proline-rich segment and the N-terminal domain, after 22 ns. The final decrease of the total R_g visible in Fig. 5.11 after an apparent stabilization between 18 and 24 ns can thus be attributed to a reduction of the distances among domains, rather than to a further shrinking of one or more of them.

The simulation shows a higher flexibility of the N-terminal compared with the region entailing the repeats [118]. If the four domains of tau to which Fig. 5.12 refers were random coils, their gyration radius could be evaluated by the formula $R_g = 0.1927 n^{0.588}$ nm, where n is the number of residues [109]. This formula yields $R_g = 3.7$ nm and 3.3 nm for

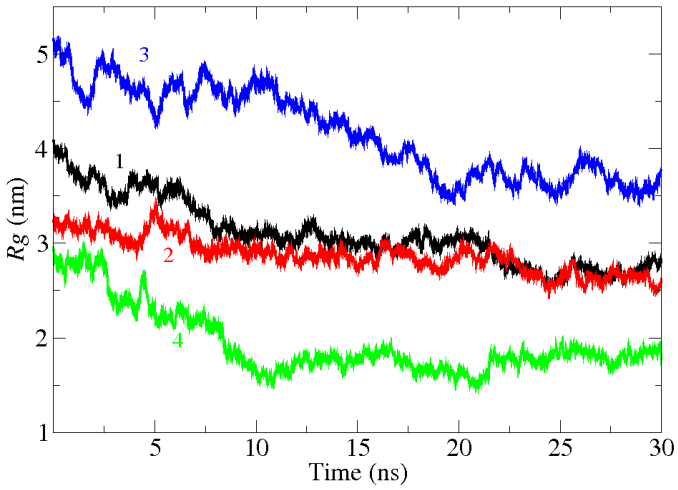


Figure 5.12: Time evolution of the radius of gyration of four domains of tau protein. Curve #1 (black): residues 1-150, N-terminal domain; curve #2 (red): residues 151-243, proline-rich segment; curve #3 (blue): residues 244-368, repeats domain; curve #4 (green): residues 369-441, C-terminal domain.

the N-terminal domain and the repeats domain, respectively; the stabilized stretches of the dynamical evolution of these quantities shown in Fig. 5.12 (the last 8 ns and 10 ns) yield an average gyration radius of 2.8 nm and 3.7 nm, respectively. The smaller value of R_g found for the N-terminal domain with respect to the random coil estimate highlights its high flexibility, i.e., a shorter persistence length of the chain; on the other hand, the higher value of R_g found for the repeats domain with respect to the random coil estimate hints at an extended conformation [109]. The value of 3.7 nm found for the repeats domain in our simulation practically coincides with the value of 3.8 nm obtained in a SAXS experiment for construct K18, which includes residues 244-372 [109], i.e., four residues more. The random coil value of R_g for the proline-rich domain is 2.8 nm, which corresponds to the average value shown in Fig. 5.12 in the stabilized stretch of the last 8 ns. Also the C-terminal, like the N-terminal, turns out to be more compact than a random coil: its average R_g, computed over the last 20 ns of the simulation is 1.9 nm, clearly smaller than the random coil estimate of 2.4 nm.

A further understanding of the conformations assumed by tau domains can be obtained by the projection of the dynamics on pairs of partial R_g. The curves thus obtained show the instantaneous R_g value of one of the

Domain	residues	Rg (nm)	
		random-coil	MD simulation
N-terminal	1-150	3.7	2.8
Proline-rich	151-243	2.8	2.8
Repeats	244-368	3.3	3.7
C-terminal	369-441	2.4	1.9

Table 5.2: R_g values of the four domains of tau; random-coil estimates and MD results.

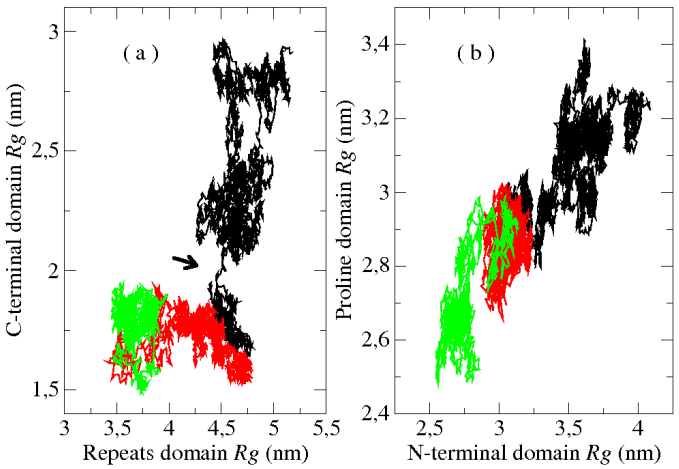


Figure 5.13: Time evolution of the radius of gyration of pairs of domains of tau. Black: 0-10 ns; red: 10-20 ns; green: 20-30 ns. Panel (a): Repeats domain and C-terminal domain; the arrow points at the transition taking place around $t = 8.3$ ns. Panel (b): N-terminal domain and proline-rich domain.

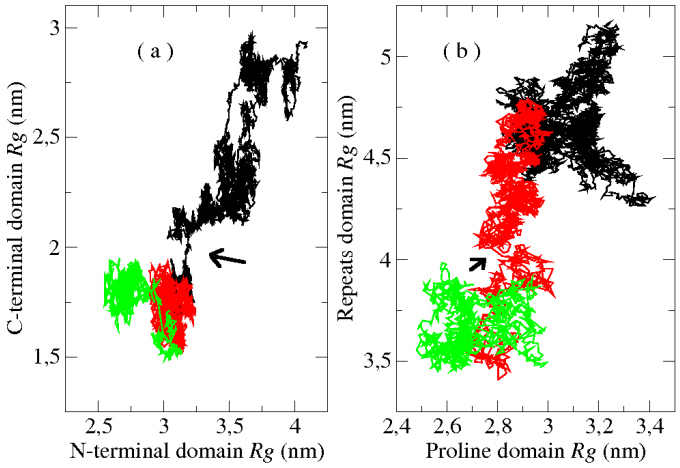


Figure 5.14: Time evolution of the radius of gyration of pairs of domains of tau. Black: 0-10 ns; red: 10-20 ns; green: 20-30 ns. Panel (a): N-terminal domain and C-terminal domain; the arrow points at the transition taking place around $t = 8.3$ ns. Panel (b): repeats domain and proline-rich domain; the arrow points at the transition taking place around $t = 17$ ns.

domains as a function of the R_g value of one of the other domains. The graphs have been represented using a color code: black in the first 10 ns, red in the 10 to 20 ns interval, and green in the last 10 ns. Fig. 5.13 and Fig. 5.14 show four different projections; in all panels the dynamics begins in the upper right quarter and ends in the lower left quarter. The patterns in the two figures show a step-wise evolution of the four domains, each step representing the dynamics in a confined, almost separated basin of attraction. Fig. 5.13(a) displays three separate basins, corresponding to three different ranges of R_g values of the C-terminal domain; during the dynamics the domain moves back and forth some times between the first and the second domain, and eventually shifts rapidly (in the time interval between 8.3 and 8.4 ns) to the third domain, below 2 nm, and stays there for the rest of the dynamics. As for the repeats domain, Fig. 5.14(b) shows a transition of R_g between two ranges, above and below 4 nm, at a time around 17 ns. The N-terminal domain and the proline-rich domain display a similar evolution, as displayed in Fig. 5.13(b) and Fig. 5.14(a). It appears that each basin of attraction represents a temporary local equilibrium for the domains that are confined in it.

The simulation provides evidence that the N-terminal approaches the

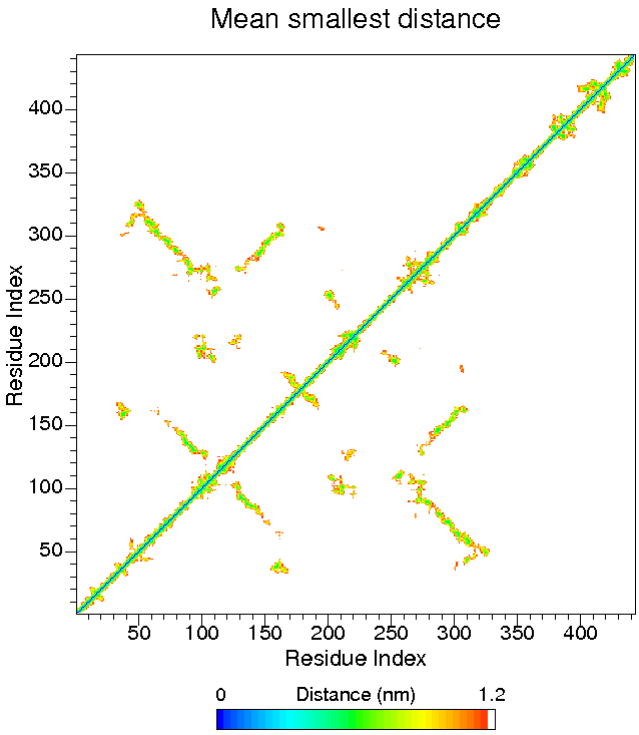


Figure 5.15: Contact map (mean smallest distance between pairs of residues) averaged over 30 ns of dynamics at $T = 300$ K.

central region of the molecule, as observed in [131]. Fig. 5.15 shows a contact map of the protein computed by the average C_α-C_α pair distance of all residues in the structure during the MD trajectory. Only distances smaller than 1.2 nm are considered. The distance map clearly shows that a large segment of the N-terminal region (residues 30-115) is found near the repeats domain (residues 255-325). This region crumples and temporarily folds in a globular-like shape. Furthermore, patterns of distances perpendicular and parallel to the main diagonal hint at stretches of local secondary structures.

5.6 SAXS experiment

We have used the data produced by the MD simulation to extract useful information on the equilibrium behaviour of tau, by comparing them with experimental SAXS results obtained by Gabriele Ciasca [132] from a specimen of tau in solution. The SAXS experiment has been performed using full length htau40 purchased from Sigma Aldrich (product code: T0-576). The powder was reconstituted in 50 mM MES, pH 6.8, 100 mM NaCl and 0.5 mM EGTA and concentrated at the nominal concentration of 2 mg/ml in 0.1X Phosphate Buffer Saline solution (ionic strength ∼0.02 M; 10X PBS: 1.3 M NaCl, 0.07 M Na_2HPO_4 and 0.03 M NaH_2PO_4, pH 7.4) by the QuickSpin protein concentration/buffer exchange (Dualsystem Biotech AG, Schlieren, Switzerland). Subsequently, the solution was centrifuged for 10 minutes at 10000 g and the supernatant passed through a 20 nm-pore size syringe filter to eliminate aggregates. Protein quality was assayed by SDS-PAGE in 12% (w/v) polyacrylamide, according to Laemmli [133]. The gels were stained with Coomassie brilliant blue R-250. The SDS-PAGE analysis revealed the occurrence of a major protein band with the expected size (∼ 45 kDa), a purity $> 90\%$ and the absence of a significant amount of aggregates.

SAXS measurements were acquired on the BioSAXS beamline (ID 14-3) at the Synchrotron Radiation Facility ESRF (Grenoble, France) [134], at the constant temperature of 303 K, for two solute concentrations, namely 1 mg/ml and 2 mg/ml. A volume of 50 μl of solution has been placed in a 1.8 mm diameter quartz capillary with a few tens of micron wall thickness. Data acquisition has been performed with a $Pilatus1M$ detector in the scattering range 0.01 - 5.8 nm^{-1}. Ten 2 seconds exposures were compared, without observing any radiation damage. SAXS data reported in the following were obtained with an exposure time of 3 seconds. Solvent scattering was measured to allow an accurate subtraction of the background scattering. The scattering patterns measured at different concentrations can be well scaled to each other, pointing out the absence of a significant aggregation phenomenon, as confirmed by the SDS-PAGE analysis. The result of this experiment is shown in Fig. 5.16(a).

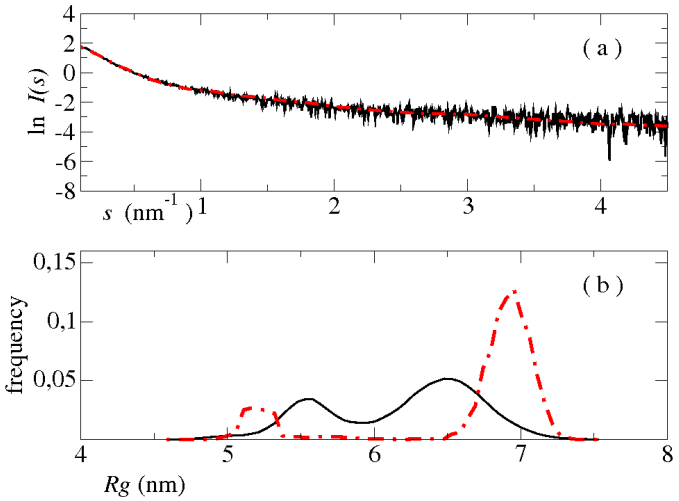


Figure 5.16: Panel (a): experimental SAXS curve (black continuous line); fit by an ensemble of conformers selected by the genetic algorithm GAJOE (red dashed line). Panel (b): distribution of R_g values. Conformers produced by the simulation (black continuous line) and conformers selected by the genetic algorithm (red line), after addition of the coordinated water layer.

5.7 Selection of equilibrium conformers

We have used the SAXS curve to extract from our simulation an ensemble of conformations that give the best fit of the experimental data, as shown in Fig. 5.16. The fitting procedure is as follows. We have extracted about 9000 regularly spaced conformers from our 30 ns MD simulation. This pool of conformers has been processed by the program CRYSOL to obtain the theoretical SAXS pattern of each conformer, taking properly into account the scattering from the hydration shell of the water layer coordinated with the molecule [135]. The Ensemble Optimization Method (EOM) with the Genetic Algorithm Judging Optimisation of Ensembles (GAJOE) was then employed to select from the pool of theoretical SAXS curves an ensemble of 162 conformers that provided on the average the best fit of the experimental SAXS curve [136]. As shown in Fig. 5.16(a), the selected ensemble corresponding to these curves fits quite well the experimental SAXS results, when each conformer is weighted with the appropriate multiplicity determined by the genetic algorithm; the accuracy of the fit is attested by its $\chi^2 = 1.4$. The distribution of R_g values of both the original pool (black

line) and of the selected ensemble (red line) are shown in Fig. 5.16(b); the range of R_g values of the original pool is shifted to the right with respect to the R_g values visible in Fig. 5.11, because the former take into account the hydration shell, which adds about 0.2 nm to the R_g values [137].

The R_g values corresponding to the two peaks of the distribution selected by the genetic algorithm, shown in Fig. 5.16(b), encompass the upper part of the distribution of R_g values derived by the same method from a pool of static conformers of tau [109]. It may be noted that while most selected conformers belong to the first 5 ns of the trajectory, where the value of R_g is near to the initial equilibrium value, there is a significant presence of conformers with R_g values between 5.1 and 5.4 nm, belonging to the temporarily stabilized trajectory stretch between 18 and 24 ns (Fig. 5.11).

5.8 Secondary structures

The selected ensemble of conformers has been analyzed with the DSSP program [138, 139], as implemented in GROMACS, to identify the presence of secondary structures like coils, β-sheets, β-bridges, bends, turns, and α-helices. We report in the A row of Table 5.3 the propensity of the molecule to form these secondary structures, as measured by the number of residues in each structure; the numbers are averages over the 162 selected conformers, each weighted with its multiplicity.

Structure	coil	β-sheet	β-bridge	bend	turn	α-helix
A	268	7	8	123	25	6
B	257 ± 11	12 ± 6	14 ± 5	123 ± 7	30 ± 6	5 ± 2

Table 5.3: Average number of residues found in coils and in various secondary structures. A row: averages over the set of 162 conformers selected by the genetic algorithm. B row: time averages over the 30 ns dynamics; the errors are one standard deviation.

In order to assess the validity of these propensities, which should correspond to an equilibrium state of the protein, we have also extracted a similar information on the formation of temporary secondary structures from the whole 30 ns dynamics. We have measured the time evolution of the number of residues found in coils, β-sheets, β-bridges, bends, turns, and α-helices.

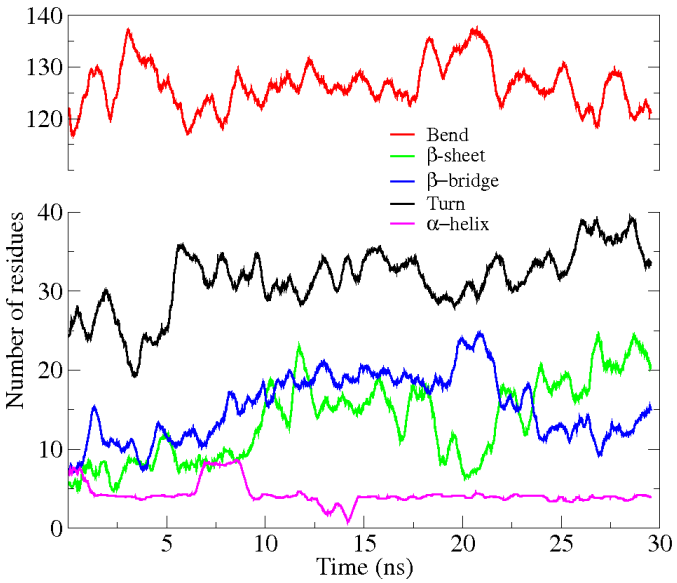


Figure 5.17: Number of residues found in secondary structures during the 30 ns dynamics. The curves have been smoothed by averaging the data over a sliding 1 ns interval.

These quantities are shown in Fig. 5.17, with the exception of the residues in a coil-like conformation, not included in the figure.

While the majority of residues are found in a coil-like conformation and in bends, there is a significant presence of secondary structures like turns, β-bridges, β-sheets, and α-helices. We report in the B row of Table 5.3 the time averaged number of residues found in these various conformations. The number of coil-like residues oscillates in a stable way during the dynamics, its average value of 257 being almost the same as the value found by the EOM/GAJOE selection procedure, taking the standard deviation of 11 as an estimate of the error affecting the average. The number of residues in β-sheets and β-bridges gradually increases during the first 15 ns, reflecting the shrinking size of the molecule; in the second half of the simulation these numbers undergo stable oscillations, in which a fraction of residues switches back and forth between the two structures: this can be clearly seen in Fig. 5.17 in the time interval 18-22 ns, and around 27 ns. The average number of residues found in β-sheets and β-bridges during the whole dynamics, respectively about 12 and 14, is higher than the corresponding number given in Table 5.3; this could mean that the progressive shrinking of the molecule,

shown in Fig. 5.11 and analyzed in Sec. 5.5, favors the formation of β-structures in excess of the equilibrium ones. On the other hand, the strong oscillations of these numbers during the dynamics produce a relatively high standard deviation, as shown in Table 5.3; taking this as the error affecting the time averages, the two estimates for the β-sheets fall in the same range, and the two estimates for the β-bridges nearly do the same. The number of residues forming bends oscillates in a stable way during the dynamics, its average value of 123 being the same as the value found by the EOM/GAJOE selection procedure. The number of residues forming turns stabilizes after about 6 ns, oscillating around an average value of 30, slightly higher than the value of 25 found with the selection procedure, but again in the range of one standard deviation. The number of residues forming a helix oscillates around 5 during the whole dynamics (mostly a α-helix, with some short shifts to a 3-helix or a 5-helix), about the value found in the set of selected conformers. The comparison of Table 5.3 and Fig. 5.17 shows that while the formation of temporary β-structures depends on the overall shape of the molecule, the frequency and extension of other secondary structures do not depend significantly on it, due to their localized nature. Paying attention to this warning about the β-structures, Fig. 5.17 shows a pattern of extension and time dependence of temporary secondary structures in tau that should be representative of the equilibrium state. It should be stressed that the numbers given in Table 5.3 are ensemble or time averages; given the temporary nature of secondary structures in tau, when they actually arise their spatial extension may greatly exceed these numbers.

We show in Fig. 5.18 a snapshot of tau taken at time $t = 28.7$ ns, corresponding to a conformer of the selected set. Some of the secondary structures accounted for in Table 5.3 are highlighted.

5.9 Discussion

We have already mentioned that the force fields used in molecular simulation have been optimized to reproduce the structures of folded proteins; therefore, a *caveat* is necessary when they are used to simulate disordered proteins. But, as more precisely tailored force fields are not available, previous simulations of segments of tau or other IDPs have been done using the ones computed to reproduce known globular proteins [112, 123, 137].

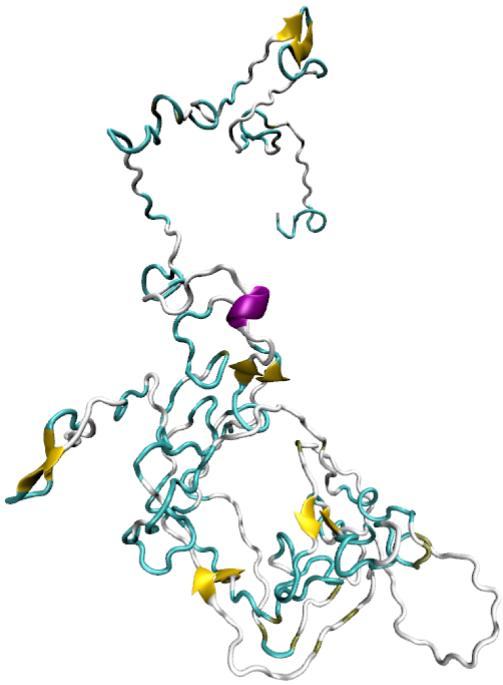


Figure 5.18: Snapshot of tau at time $t = 28.7$ ns, corresponding to a conformer of the selected ensemble. Some secondary structures are highlighted: turns (cyan), β-sheets (yellow) and α-helices (purple). The terminal visible in the upper half of the figure is the C-terminal.

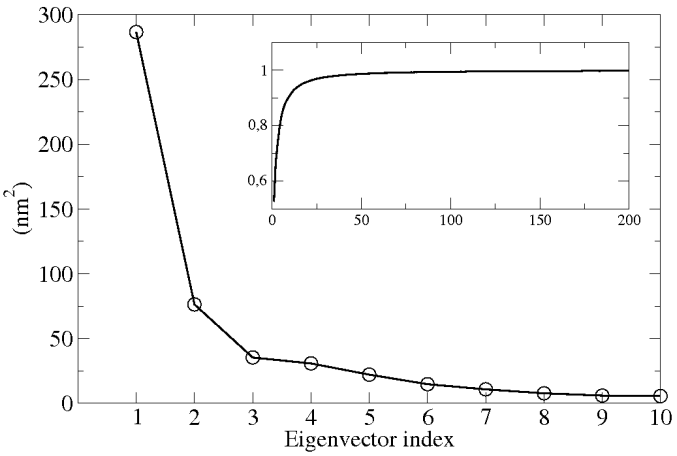


Figure 5.19: First ten eigenvalues of the covariance matrix computed over 30 ns of dynamics at $T = 300$ K. The insert shows the cumulative sum of the eigenvalues: the first ten eigenvectors account for 91% of the total fluctuation.

Static 3D structures of segments or of the whole tau have been produced by a random choice of angles between neighbouring amino acids taken from a database [112, 118, 136, 140]; this base was derived from known 3D structures of globular proteins and is therefore subject to a similar bias as the mentioned force fields.

The analysis in Sec. 5.5 of the time evolution of the four domains of tau shows that while the overall structure of the molecule progressively shrinks, each domain is most of the time in a basin of local equilibrium, moving from time to time from one basin to another. A further insight into the equilibrium-like behavior of our tau model is provided by a Principal Component Analysis (PCA) [58, 141, 142] of the 30 ns dynamics (see section [3.7]). In Fig. 5.19 the eigenvalues corresponding to the first ten eigenvectors of the positional fluctuations covariance matrix are reported, showing that most of the dynamics is entailed in the first ten eigenvectors, which account for 91% of the total fluctuation of the molecule; all further eigenvalues have negligible magnitude. It is worth noting that, also for the intrinsically disordered protein htau40, the sum of the eigenvalues of the essential subspace (defined as the first 10 eigenvectors of the atomic covariance fluctuation matrix) is in the range that has been computed for stable folded proteins [58]. This is likely to reflect the transient formation in htau40 of folded domains entailing secondary structure regions with a flexibility typical of natively

folded proteins.

In Fig. 5.20 the time evolution of the amplitude of the dynamical components along the first ten eigenvectors is displayed. The amplitude of the component along the first eigenvector, which constitutes the largest component of the dynamics, increases over the first 20 ns, while it appears to be stabilized in the last 10 ns. The amplitude of the component along the second eigenvector displays an increasing trend after 15 ns. As for the following eigenvectors, the corresponding components of the dynamics seem to be stabilized, i.e. to have reached an irregular oscillatory behaviour after 10-15 ns. It appears that most principal components of the dynamics have reached an equilibrium state during the dynamics, with the possible exception of the component along the second eigenvector. As these components encompass all four domains of tau, this also supports the validity of our simulation for the study of temporary, local secondary structures.

The set of conformers selected from the simulated dynamics by fitting the SAXS data is the best approximation of an equilibrium ensemble that can be extracted from our 30 ns dynamics. Even if the force field chosen for this simulation may not be able to reproduce all the features of the overall tertiary structure of tau, as shown by the progressive decrease of R_g in Fig. 5.11, the results on the formation of local transient secondary structures appear to be sound, albeit with the warning regarding the number of residues in β-structures shown in Fig. 5.17. In this regard, it is interesting to compare our results with those obtained by Mukrasch and coworkers [118], who used NMR analysis to assign to various segments of tau accurate propensities to form a β-structure or an α-helix. Weighing the number of residues entailed in each of these segments with its propensity (fraction of time spent in the secondary structure), one finds an average number of 12 residues in β-structures and of 4 residues in a α-helix. Comparing these results with Table 5.3 shows that while the first number is of the order of (but 20% less than) the sum of the average numbers of residues found in β-sheets (7) and β-bridges (8) in the selected ensemble of conformers, the second agrees with the time average over the 30 ns dynamics, within one standard deviation.

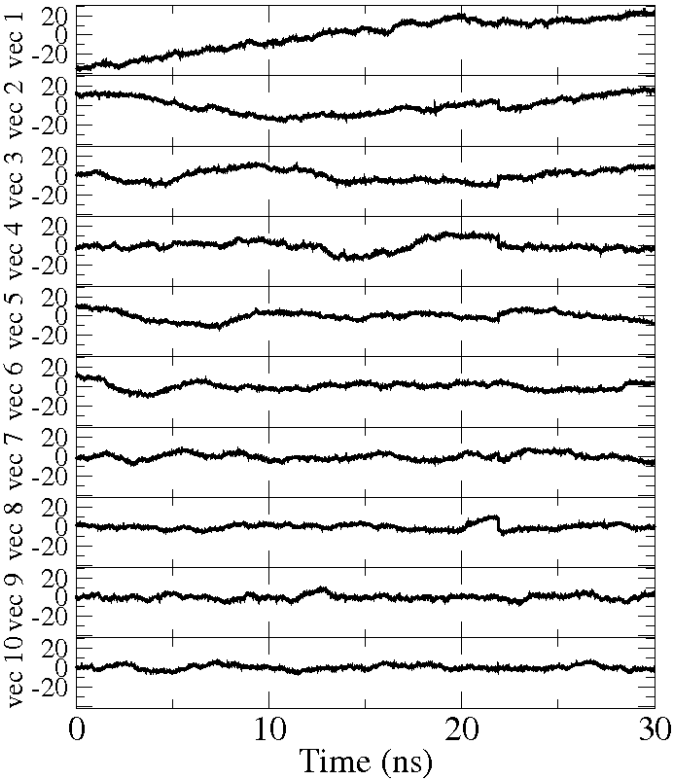


Figure 5.20: Time evolution of the dynamical components along the first ten eigenvectors of the covariance matrix during the dynamics at $T = 300$ K.

5.10 Conclusions

We propose a procedure to start the molecular dynamics simulations of intrinsically disordered proteins, that overcomes the lack of a known 3D structure of these molecules, and only requires the knowledge of the primary sequence. We have tested this method on tau, a large totally disordered protein, and verified its effectiveness, as it requires only few hundred picoseconds of dynamical simulation. This procedure should be useful to the community of researchers interested in the atomistic simulation of the dynamical behaviour of the wide, and yet little known, set of intrinsically disordered proteins.

We have analyzed a 30 ns interval of conformational dynamics on the htau40 protein; this provides a significant observation of secondary structures and of the overall fold of the molecule. The simulated trajectory indicates that the protein samples a limited number of almost stable secondary structure motifs (mainly short α and β structures), whereas the overall preferred conformation is a random coil. We find that the size of the radius of gyration is mainly due to a large collapse of the N-terminal domain that drives the initial fold of the polypeptide chain from its starting extended conformation. Finally, the combination of our simulation data and SAXS measurements yields an ensemble of conformers that represent a 3D data basis from which further simulations can be started.

The transient nature of the secondary structures in protein tau is relevant, and has biochemical implications *in vivo*; their topological and temporal details is the object of further research on our side.

General conclusion

This thesis investigates two biological systems using *atomistic modeling* and *molecular dynamics simulation*: we have studied the interaction between a segment of a DNA molecule and a functionalized surface, and the dynamical modeling of protein tau, an intrinsically disordered protein.

Our study highlights that the model used to describe the system plays a critical role in the results observed. Therefore, validation of the model through comparison with the literature or with experimental measures is very important in any computational study; this is especially true in computational studies of biological systems, where the complexity of the system requires a very large number of parameters to define the force field. On the other hand, when one has verified the range of validity of the model, an atomistic description and a molecular dynamics simulation are very powerful tools: computer simulation can complement experiment by providing not only averages, but also distributions and time series of any definable quantity, for example, conformational distributions or interactions between parts of systems.

Acknowledgements

I would like to thank my tutor Dr. Maurizio Dapor for giving me the resources and the possibility to reach the results obtained. Thanks for your friendly attitude.

Thanks to Dr. Giovanni Garberoglio for introducing me to the *complex* area of the biological system simulations; for being both a point of reference and comparison in these years. Thanks for the DFT calculations included in this work, for the laughter (I will never forget the day you were born!) and also for the heated discussions.

I also wish to thank the BioSint Group: Dr. Cecilia Pederzolli, Dr. Lorenzo Lunelli, Lorenza Marocchi, Dr. Laura Pasquardini e Dr. Cristina Potrich. Thanks for the results produced and included in this work.

Thanks to KORE cluster and to the relevant Customer Support GSC.

I wish to thank Prof. Alexander Tenenbaum, extremely competent, capable and reliable. Thanks for involving me in the Project relevant to Protein TAU, for all the works and the discussions, also relevant to the work on DNA, always interesting and challenging. Thanks for improving my English too.

Thanks to Dr. Gabriele Ciasca for the work made together, for the experimental data included in this thesis and for the friendship.

Thanks to Dr. Alessandro Grottesi for the suggestions on the use of Gromacs, for the useful discussions and for the support given on the use of CASPUR cluster.

Thanks to Prof. Giuseppina Orlandini for the very interesting discussions relevant to the course of *Many Body Theory* and for the opportunity of making a teaching experience.

Thanks to Simone Taioli for the exchange of ideas and for the friendship.

Thanks to my PhD colleagues. A special thanks to Emmanuel for the exchanged ideas. Thanks to Roberto, Diego, Giorgia and Enrico with whom

I shared the PhD room with pleasure.

I wish to thank to all the members of LISC.

Thanks to Micaela Paoli, always very available and kind.

Thanks to my friends: Giorgina, Lucia, Amedeo, Nadia and Zsuzsanna with whom I had nice breaks.

Thanks to Sergio that always amaze me with his ability in using LaTeX.

Thanks to Stefania, Valeria, Manuela, Ramona and Effie whose friendship has been really important, in particular at the beginning of my time in Trento.

A special thanks to my family, my cornerstone. Thanks to my mommy and daddy. Thanks to Nicola, Edoardo, Giulio and Maria Rosaria.

Thanks also to Gabriella, Enrico, Luca, Rosanna and Primo, always affectionate.

Thanks to *my love* Renato.

Bibliography

[1] Watson, J.D. and Crick, F.H.C., Molecular structure of nucleic acids; Nature, **171**, 737-738 (1953).

[2] Behr, J.P., The lock-and-key principle: the state of the art-100 years on; **1** (1994).

[3] Ramachandran, GN and RAMAKRISHNAN, C. and Sasisekharan, V., Stereochemistry of polypeptide chain configurations; Journal of molecular biology, **7**, 95 (1963).

[4] Cremer, D., Møller-Plesset perturbation theory; Encyclopedia of computational chemistry (1998).

[5] Goddard III, W.A. and Dunning Jr, T.H. and Hunt, W.J. and Hay, P.J., Generalized valence bond description of bonding in low-lying states of molecules; Accounts of Chemical Research, **6**, 368-376 (1973).

[6] Sherrill, C.D. and Schaefer, H.F., The configuration interaction method: Advances in highly correlated approaches; Advances in quantum chemistry, **34**, 143-269 (1999).

[7] Hermann, G.K., A biography of the Coupled Cluster Method.

[8] Kawashima, N, Quantum Monte Carlo Methods; Progress of Theoretical Physycs-Supplement, 138-149 (2002).

[9] Gross, E.K.U. and Dreizler, R.M., Density functional theory; Springer **337** (1995).

[10] Lindahl, E. and Hess, B. and Van Der Spoel, D., GROMACS 3.0: a package for molecular simulation and trajectory analysis; Journal of Molecular Modeling, **7**, 306-317 (2001).

[11] http://www.gromacs.org/

[12] W. Humphrey, A. Dalka, and K. Schulten, Visual Molecular Dynamics, Journal of Molecular Graphics **14**, 33-38 (1996).

[13] Smith, W. and Forester, TR and Todorov, IT and Leslie, M., The DL poly 2 user manual; STFC Daresbury Laboratory Daresbury, Warrington WA4 4AD Cheshire, UK, Version **2** (2008).

[14] Jorgensen, W.L., Quantum and statistical mechanical studies of liquids. 10. Transferable intermolecular potential functions for water, alcohols, and ethers. Application to liquid water; Journal of the American Chemical Society, **103**, 335-340 (1981).

[15] Hermans, J. and Berendsen, H.J.C. and Van Gunsteren, W.F. and Postma, J.P.M., A consistent empirical potential for water–protein interactions; Biopolymers, **23**, 1513-1518 (1984).

[16] Mark, P. and Nilsson, L., Structure and dynamics of the TIP3P, SPC, and SPC/E water models at 298 K; The Journal of Physical Chemistry A, **105** 9954-9960 (2001).

[17] Wang, J. and Wolf, R.M. and Caldwell, J.W. and Kollman, P.A. and Case, D.A., Development and testing of a general amber force field; Journal of computational chemistry, **25** , 1157-1174 (2004).

[18] Momany, F.A. and Rone, R., Validation of the general purpose QUANTA® 3.2/CHARMm® force field; Journal of computational chemistry, **13**, 888-900 (1992).

[19] Jorgensen, W.L. and Maxwell, D.S. and Tirado-Rives, J., Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids; Journal of the American Chemical Society, **118**, 11225-11236 (1996).

[20] Zhu, J. and Shi, Y. and Liu, H., Parametrization of a generalized Born/solvent-accessible surface area model and applications to the simulation of protein dynamics; The Journal of Physical Chemistry B, **106** 4844-4853 (2002).

[21] Postma, J.P.M. and Berendsen, H.J.C. and Haak, J.R., Thermodynamics of cavity formation in water. A molecular dynamics study; Faraday Symp. Chem. Soc., **17**, 55-67 (1982).

[22] Ferrara, P. and Apostolakis, J. and Caflisch, A., Evaluation of a fast implicit solvent model for molecular dynamics simulations; Proteins: Structure, Function, and Bioinformatics, **46**, 24-33 (2002).

[23] Roux, B., Implicit solvent models; Eastern Hemisphere Distribution, 133 (2001).

[24] Born, M., Volumes and heats of hydration of ions; Z. Phys **1**, 45-48 (1920).

[25] Kirkwood, JG, Theoretical Studies upon Dipolar Ions; Chemical Reviews, **24**, 233-251 (1939).

[26] Onsager, L., Electric moments of molecules in liquids; Journal of the American Chemical Society, **58**, 1486-1493 (1936).

[27] http://www.charmm-gui.org/?doc=input/pbeqsolver

[28] Fenley, A.T. and Gordon, J.C. and Onufriev, A., An analytical approach to computing biomolecular electrostatic potential. I. Derivation and analysis; The Journal of chemical physics, **129**, 075101 (2008).

[29] Po, H.N. and Senozan, NM, The Henderson-Hasselbalch equation: Its history and limitations; Journal of Chemical Education, **78**, 1499 (2001).

[30] Neves-Petersen, M.T. and Petersen, S.B., Protein electrostatics: a review of the equations and methods used to model electrostatic equations in biomolecules–applications in biotechnology; Biotechnology Annual Review, **9**, 315-395 (2003).

[31] http://www.gromacs.org/Documentation/How-tos/Constant_pH_Simulation?highlight=delphi

[32] Tsui, V. and Case, D.A., Theory and applications of the generalized Born solvation model in macromolecular simulations; Biopolymers, **56**, 275-291 (2000).

[33] Di Qiu, and Shenkin, P.S. and Hollinger, F.P. and Still, W.C., The GB/SA continuum model for solvation. A fast analytical method for the calculation of approximate Born radii; The Journal of Physical Chemistry A, **101**, 3005-3014 (1997).

[34] Hawkins, G.D. and Cramer, C.J. and Truhlar, D.G., Parametrized models of aqueous free energies of solvation based on pairwise descreening of solute atomic charges from a dielectric medium; The Journal of Physical Chemistry, **100** 19824-19839 (1996).

[35] Onufriev, A. and Bashford, D. and Case, D.A., Exploring protein native states and large-scale conformational changes with a modified generalized born model; Proteins: Structure, Function, and Bioinformatics, **55**, 383-394 (2004).

[36] Hornak, V. and Abel, R. and Okur, A. and Strockbine, B. and Roitberg, A. and Simmerling, C., Comparison of multiple Amber force fields and development of improved protein backbone parameters; Proteins: Structure, Function, and Bioinformatics, **65**, 712-725 (2006).

[37] Mu, Y. and Kosov, D.S. and Stock, G., Conformational dynamics of trialanine in water. 2. Comparison of AMBER, CHARMM, GROMOS, and OPLS force fields to NMR and infrared experiments; The Journal of Physical Chemistry B, **107**, 5064-5073 (2003).

[38] Martin, M.G., Comparison of the AMBER, CHARMM, COMPASS, GROMOS, OPLS, TraPPE and UFF force fields for prediction of vapor-liquid coexistence curves and liquid densities; Fluid phase equilibria, **248**, 50–55 (2006).

[39] Smith, J.C., X-Ray and Neutron Scattering as Probes of the Dynamics of Biological Molecules; Eastern Hemisphere Distribution, 237 (2001).

[40] Becker, O.M. and Mackerell Jr, A.D. and Roux, B. and Watanabe, M., Computational biochemistry and biophysics.; CRC Press, chapter 13, 14 (2001).

[41] Chothia, C. and Lesk, A.M., The relation between the divergence of sequence and structure in proteins; The EMBO journal, **5**, 823 (1986).

[42] Frenkel, D. and Smit, B., Understanding molecular simulation: from algorithms to applications; Elsevier (formely published by Academic Press) (1996).

[43] Ryckaert, J.P. and Ciccotti, G. and Berendsen, H.J.C., Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes; Journal of Computational Physics, **23**, 327-341 (1977).

[44] Hess, B. and Bekker, H. and Berendsen, H.J.C. and Fraaije, J.G.E.M., LINCS: a linear constraint solver for molecular simulations; Journal of computational chemistry, **18**, 1463-1472 (1997).

[45] Berendsen, H.J.C. and Postma, J.P.M. and Van Gunsteren, W.F. and DiNola, A. and Haak, JR, Molecular dynamics with coupling to an external bath; The Journal of Chemical Physics, **81**, 3684 (1984).

[46] Nosé, S., A molecular dynamics method for simulations in the canonical ensemble; Molecular Physics, **52**, 255-268 (1984).

[47] Hoover, W.G., Canonical dynamics: Equilibrium phase-space distributions; Physical Review A, **31**, 1695 (1985). The Journal of Chemical Physics, **81**, 3684 (1984).

[48] Parrinello, M. and Rahman, A., Polymorphic transitions in single crystals: A new molecular dynamics method; Journal of Applied Physics, **52**, 7182-7190 (1981).

[49] Auffinger, P. and Beveridge, D.L., A simple test for evaluating the truncation effects in simulations of systems involving charged groups; Chemical physics letters, **234**, 413-415 (1995).

[50] Cheng, H. and Greengard, L. and Rokhlin, V., A fast adaptive multipole algorithm in three dimensions; Journal of Computational Physics, **155**, 468-498 (1999).

[51] Darden, T. and York, D. and Pedersen, L., Particle mesh Ewald: An N log (N) method for Ewald sums in large systems; Journal of Chemical Physics, **98**, 10089-10089 (1993).

[52] Laio, A. and Gervasio, F.L., Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science; Reports on Progress in Physics, **71**, 126601 (2008).

[53] Laio, A. and Parrinello, M., Escaping free-energy minima; Proceedings of the National Academy of Sciences, **99**, 12562 (2002).

[54] Kumar, S. and Rosenberg, J.M. and Bouzida, D. and Swendsen, R.H. and Kollman, P.A., The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method; Journal of Computational Chemistry, **13**, 1011-1021 (1992).

[55] Roux, B., The calculation of the potential of mean force using computer simulations; Computer Physics Communications, **91**, 275-282 (1995).

[56] Hub, J.S. and De Groot, B.L. and Van Der Spoel, D., g_wham-A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates; Journal of Chemical Theory and Computation, (2010).

[57] Chernick, M. R., Bootstrap methods: A guide for practitioners and researchers; Wiley-Interscience, **619** (2008).

[58] Amadei, A. and Linssen, A. and Berendsen, H.J.C., Essential dynamics of proteins; **17**, 412-425 (1993).

[59] Van Aalten, DMF and Amadei, A. and Linssen, ABM and Eijsink, VGH and Vriend, G. and Berendsen, HJC, The essential dynamics of thermolysin: Confirmation of the hinge-bending motion and comparison of simulations in vacuum and water; Proteins: Structure, Function, and Bioinformatics, **22**, 45-54 (1995).

[60] an Aalten, DMF and De Groot, BL and Findlay, JBC and Berendsen, HJC and Amadei, A., A comparison of techniques for calculating protein essential dynamics; Journal of computational chemistry, **18**, 169-181 (1997).

[61] M. Zwolak and M. Di Ventra, Colloquium: Physical approaches to DNA sequencing and detection; Rev. Mod. Phys. **80**, 141 (2008).

[62] Aksimentiev, Aleksei and Brunner, Robert and Cohen, Jordi and Comer, Jeffrey and Cruz-Chu, Eduardo and Hardy, David and Rajan, Aruna and Shih, Amy and Sigalov, Grigori and Yin, Ying and Schulten, Klaus, Computer Modeling in Biotechnology; Nanostructure Design, Methods in Molecular Biology, **474**, 181 (2008).

[63] Ise, N., Ordering of ionic solutes in dilute solutions through attraction of similarly charged solutes - A change of paradigm in colloid and polymer chemistry; Angew. Chem. Int. Ed., **25**, 323 (1986).

[64] Larsern, A.E. and Grier, D.G., Like-charge attractions in metastable colloidal crystallites; Nature, **385**, 16 (1997). A. Larserr and D. Grier, Nature 385, 16 (1997).

[65] Zixiang Tang and L. E. Scriven and H. T. Davis, Interactions between primitive electrical double layers; J. Chem. Phys., **97**, 9258 (1992).

[66] Messina, R., Electrostatics in soft matter; J. Phys.: Condens. Matter, **21**, 113102 (2009).

[67] Chan, V. and Graves, D.J. and Fortina, P. and McKenzie, S.E., Adsorption and surface diffusion of DNA oligonucleotides at liquid/solid interfaces; Langmuir, **13**, 320 (1997).

[68] Solberg, S.M. and Landry, C.C., Adsorption of DNA into mesoporous silica; J. Phys. Chem. B, **110**, 15261 (2006).

[69] Yoza, B. and Matsumoto, M. and Matsunaga, T., DNA extraction using modified bacterial magnetic particles in the presence of amino silane compound; J. Biotech., **94**, 217 (2002).

[70] Nakagawa, T. and Tanaka, T. and Niwa, D. and Osaka, T. and Takeyama, H. and Matsunaga, T., Fabrication of amino silane-coated microchip for DNA extraction from whole blood; J. Biotech., **116**,, 105 (2005).

[71] Aranda-Espinoza, H. and Chen, Y. and Dan, N. and Lubensky, T.C. and Nelson, P. and Ramos, L. and Weitz, D.A., Electrostatic repulsion of positively charged vesicles and negatively charged objects; Science, **285** 394 (1999).

[72] Trulsson, M. and Jönsson, B. and Åkesson, T. and Forsman, J. and Labbez, C., Repulsion between oppositely charged surfaces in multivalent electrolytes; Phys. Rev. Lett., **97**, 68302 (2006).

[73] Trulsson, M. and Jönsson, B. and Åkesson, T. and Forsman, J. and Labbez, C., Repulsion between oppositely charged macromolecules or particles; Langmuir, **23**, 11562 (2007).

[74] Mengistu, DH and May, S., Debye-Hückel theory of mixed charged-zwitterionic lipid layers; Eur. Phys. J. E, **26**, 251 (2008).

[75] Mengistu, D.H. and Bohinc, K. and May, S., Binding of DNA to zwitterionic lipid layers mediated by divalent cations; J. Phys. Chem. B, **113**, 12277 (2009).

[76] Mengistu, D.H. and Bohinc, K. and May, S., A model for the electrostatic contribution to the pH-dependent nonideal mixing of a binary charged-zwitterionic lipid bilayer; Biophys. Chem., **150**, 112 (2010).

[77] Patra, C.N. and Yethiraj, A., Density functional theory for the nonspecific binding of salt to polyelectrolytes: Thermodynamic properties; Biophys. J., **78**, 699 (2000). C. Patra and A. Yethiraj, Biophys. J. 78, 699 (2000).

[78] Patra, C.N. and Chang, R. and Yethiraj, A., Structure of polyelectrolyte solutions at a charged surface; J. Phys. Chem. B, **108**, 9126 (2004). C. Patra, R. Chang, and A. Yethiraj, J. Phys. Chem. B 108, 9126 (2004).

[79] Goel, T. and Patra, C.N., Structure of spherical electric double layers: A density functional approach; J. Chem. Phys., **127**, 034502 (2007). T. Goel and C. Patra, J. Chem. Phys. 127, 034502 (2007).

[80] Goel, T. and Patra, C.N. and Ghosh, S.K. and Mukherjee, T., Molecular solvent model of cylindrical electric double layers: A systematic study by Monte Carlo simulations and density functional theory; J. Chem. Phys., **129**, 154707 (2008).

[81] Patra, C.N., Molecular Solvent Model of Spherical Electric Double Layers: A Systematic Study by Monte Carlo Simulations and Density Functional Theory; J. Phys. Chem. B, **113**, 13980 (2009).

[82] B. Cappella and G. Dietler, Force-distance curves by atomic force microscopy; Surface Science Reports **34**, 1-104 (1999).

[83] Dammer, U. and Hegner, M. and Anselmetti, D. and Wagner, P. and Dreier, M. and Huber, W. and Güntherodt, H.J., Specific antigen/antibody interactions measured by force microscopy; Biophysical journal **70**, 2437-2441 (1996).

[84] Roy, D. and Kwon, S.H. and Kwak, J.W. and Park, J.W., Seeing and Counting Individual Antigens Captured on a Microarrayed Spot with Force-Based Atomic Force Microscopy; Analytical chemistry, **82**, 5189-5194 (2010).

[85] Dufrêne, Y.F. and Hinterdorfer, P., Recent progress in AFM molecular recognition studies; Pflugers Archiv European Journal of Physiology, **456**, 237-245, (2008).

[86] Clowney, L. and Westbrook, J.D. and Berman, H.M., CIF applications. XI. A La Mode: a ligand and monomer object data environment. I. Automated construction of mmCIF monomer and ligand models; Journal of applied crystallography, **32**, 125-133 (1999).

[87] Berman, H.M., Clowney, L., Gelbin, A., Zardecki, C., and Westbrook, J.D., The network interface to the Nucleic Acid Database.

[88] Berman, H.M., Crystal studies of B-DNA: the answers and the questions, Biopolymers **44**, 23-44 (1997).

[89] Parkinson, G., Vojtechovsky, J., Clowney, L., Brnger, A.T., and Berman, H.M., New parameters for the refinement of nucleic acid containing structures; Acta Crystallographica Section D: Biological Crystallography, **52**, 57-64 (1996).

[90] Bloomfield, V.A. and Crothers, D.M. and Tinoco, I., Nucleic acids: structures, properties, and functions; Univ Science Books (2000).

[91] Pearlman, D.A. and Kim, S.H., Atomic charges for DNA constituents derived from single-crystal X-ray diffraction data; Journal of molecular biology, **211**, 171-187 (1990).

[92] Pasquardini, L. and Lunelli, L. and Potrich, C. and Marocchi, L. and Fiorilli, S. and Vozzi, D. and Vanzetti, L. and Gasparini, P. and Anderle, M. and Pederzolli, C., Organo-silane coated substrates for DNA purification; Applied Surface Science (2011).

[93] Ebner, C. and Saam, W.F. and Stroud, D., Density-functional theory of simple classical fluids. I. Surfaces; Phys. Rev. A, **14**, 2264 (1976).

[94] Barrat, J.L. and Hansen, J.-P., Basic concepts for simple and complex liquids; Cambridge University Press (2003).

[95] Hansen, J.-P. and McDonald, I. R., Theory of simple liquids; Elsevier, third ed. (2006).

[96] Rosenfeld, Y., Free-energy model for the inhomogeneous hard-sphere fluid mixture and density-functional theory of freezing; Phys. Rev. Lett., **63**, 980 (1989). Y. Rosenfeld, Phys. Rev. Lett. 63, 980 (1989), ISSN 1079-7114.

[97] Rosenfeld, Y. and Schmidt, M. and Löwen, H. and Tarazona, P., Fundamental-measure free-energy density functional for hard spheres: Dimensional crossover and freezing; Phys. Rev. E, **55**, 4245 (1997).

[98] Roth, R. and Evans, R. and Lang, A. and Kahl, G., Fundamental measure theory for hard-sphere mixtures revisited: the White Bear version; J. Phys: Condens. Matter, **14**, 12063 (2002).

[99] Battisti, A., Marocchi, L., Lunelli, L., and Garberoglio, G., Adsorption of DNA oligomers on amine-functionalized surface: a combined experimental and computational investigation (work in progress).

[100] Release 3.1, http://software.sandia.gov/tramonto/.

[101] S.L. Mayo and B.D. Olafson and W.A. {Goddard III}, DREIDING: A generic force field for molecular simulations; J. Phys. Chem., **94**, 8897 (1990).

[102] A.W. Schuettelkopf and D.M.F. van Aalten, PRODRG - a tool for high-throughput crystallography of protein-ligand complexes; Acta Cryst., **D60**, 1355 (2004).

[103] Tironi, I.G. and Sperb, R. and Smith, P.E. and van Gunsteren, W.F., A generalized reaction field method for molecular dynamics simulations; The Journal of chemical physics, **102**, 5451 (1995).

[104] Cedric Volcke and Ram Prasad Gandhiraman and Vladimir Gubala and Colin Doyle and G. Fonder and Paul A. Thiry and Attilio A. Cafolla and Bryony James and David E. Williams, Plasma functionalization of AFM tips for measurement of chemical interactions; Journal of Colloid and Interface Science, **348**, 322 - 328 (2010).

[105] P. Tompa, 'Intrinsically Disordered Proteins', in 'Structural Proteomics and its Impact on the Life Sciences'; J. Sussman and I. Silman eds., World Scientific (2008), pp 153-158.

[106] A.K. Dunker, C.J. Brown, J.D. Lawson, L.M. Iakoucheva, and Z. Obradovic, Intrinsic Disorder and Protein Function; Biochemistry **41**, 6573-6582 (2002).

[107] M. Sickmeier, J.A. Hamilton, T. LeGall, V. Vacic, M.S. Cortese, A. Tantos, B. Szabo, P. Tompa, J. Chen, V.N. Uversky, Z. Obradovic, and A.K. Dunker, DisProt: the Database of Disordered Proteins, Nucl. Acids Res. **35**, D786-793 (2007).

[108] P. Tompa, Intrinsically Unstructured Proteins; Trends Biochem. Sci. **27**, 527-533 (2002).

[109] E. Mylonas, A. Hascher, P. Bernad, M. Blackledge, E. Mandelkow, and D.I. Svergun, Domain Conformation of Tau Protein Studied by Solution Small-Angle X-ray Scattering; Biochemistry **47**, 10345-10353 (2008).

[110] P. Tompa, The interplay between structure and function in intrinsically unstructured proteins; FEBS Lett. **579**, 3346-3354 (2005).

[111] Tompa, P. and Szasz, C. and Buday, L., Structural disorder throws new light on moonlighting; Trends in biochemical sciences, **30** 484-489 (2005).

[112] M.D. Mukrasch, P. Markwick, J. Biernat, M. von Bergen, P. Bernad, C. Griesinger, E. Mandelkow, M. Zweckstetter, and M. Blackledge,

Highly Populated Turn Conformations in Natively Unfolded Tau Protein Identified from Residual Dipolar Couplings and Molecular Simulation; J. Am. Chem. Soc. **129**, 5235-5243 (2007).

[113] J. Avila, J.J. Lucas, M. Prez, and F. Hernndez, Role of Tau Protein in Both Physiological and Pathological Conditions; Physiol. Rev. **84**, 361-384 (2004).

[114] O. Schweers, E. Schoenbrunn-Hanebeck, A. Marx, and E. Mandelkow, Structural Studies of Tau Protein and Alzheimer Paired Helical Filaments Show No Evidence for p-Structure; J. Biol. Chem. **269**, 24290-24297 (1994).

[115] M. von Bergen, S. Barghorn, J. Biernat, E.-M. Mandelkow, and E. Mandelkow, Tau aggregation is driven by a transition from random coil to beta sheet structure; Biochimica et Biophysica Acta **1739**, 158-166 (2005).

[116] M. von Bergen, P. Friedhoff, J. Biernat, J. Heberle, E.-M. Mandelkow, E. Mandelkow, Assembly of tau protein into Alzheimer paired helical filaments depends on a local sequence motif ((306)VQIVYK(311)) forming beta structure; Proc. Natl. Acad. Sci. USA **97**, 5129-5134 (2000).

[117] M.R. Sawaya, S. Sambashivan1, R. Nelson, M. I. Ivanova, S. A. Sievers,M. I. Apostol, M. J. Thompson, M. Balbirnie, J. J. W. Wiltzius, H. T. McFarlane,A. . Madsen, C. Riekel, and David Eisenberg, Atomic structures of amyloid cross- spines reveal varied steric zippers; Nature **447**, 453-457 (2007).

[118] M.D. Mukrasch , S. Bibow, J. Korukottu, S. Jeganathan, J. Biernat, C. Griesinger, E. Mandelkow, and M. Zweckstetter, Structural Polymorphism of 441-Residue Tau at Single Residue Resolution; PLoS Biology **7**, 0399-0414 (2009).

[119] P. Friedhoff, M. von Bergen, E.-M. Mandelkow, E. Mandelkow, Structure of tau protein and assembly into paired helical filaments; Biochimica et Biophysica Acta **1502**, 122-132 (2000).

[120] S. Jeganathan, M. von Bergen, H. Brutlach, H.J. Steinhoff, and E. Mandelkow, Global Hairpin Folding of Tau in Solution; Biochemistry **45**, 2283-2293 (2006).

[121] www.uniprot.org/uniprot/P10636.

[122] H. Singh, S.S. Marla, and M. Agarwal, Docking studies of Tau Protein; IAENG Int. J. of Computer Science **33**, 36-42 (2007).

[123] J. Mendieta, M.A. Fuertes, R. Kunjishapatham, I. Santa-Mara, F.J. Moreno, C. Alonso, F. Gago, V. Munoz, J. Avila, and F. Hernndez, Phosphorylation modulates the alpha-helical structure and polymerization of a peptide from the third tau microtubule-binding repeat; Biochimica et Biophysica Acta **1721**, 16-26 (2005).

[124] Battisti, A. and Tenenbaum, A., Molecular dynamics simulation of intrinsically disordered proteins; Taylor & Francis (2011).

[125] GROMACS release 4.5.3, www.gromacs.org; box volume = 15253 nm^3; spce water model; time step 2 fs; modified Berendsen thermostat, Parrinello-Rahman pressure coupling.

[126] Battisti, A. and Ciasca, G. and Grottesi, A. and Bianconi, A. and Tenenbaum, A., Temporary secondary structures in tau, an intrinsically disordered protein; Taylor & Francis (2011).

[127] Still, W.C. and Tempczyk, A. and Hawley, R.C. and Hendrickson, T., Semianalytical treatment of solvation for molecular mechanics and dynamics; Journal of the American Chemical Society, **112**, 6127-6129 (1990).

[128] Roe, D.R. and Okur, A. and Wickstrom, L. and Hornak, V. and Simmerling, C., Secondary structure bias in generalized Born solvent models: comparison of conformational ensembles and free energy of solvent polarization from explicit and implicit solvation; The Journal of Physical Chemistry B, **111**, 1846-1857 (2007).

[129] Tan, C. and Yang, L. and Luo, R., How well does Poisson-Boltzmann implicit solvent agree with explicit solvent? A quantitative analysis; The Journal of Physical Chemistry B, **110**, 18680-18687 (2006).

[130] Arteca, G.A. and Reimann, CT and Tapia, O., Proteins in vacuo: Denaturing and folding mechanisms studied with computer-simulated molecular dynamics; Mass spectrometry reviews, **20**, 402-422 (2001).

[131] T.C. Gamblin, R.W. Berry, and L.I. Binder, Tau Polymerization: Role of the Amino Terminus; Biochemistry **42**, 2252-2257 (2003).

[132] G. Ciasca, G. Campi, A. Battisti, G. Rea, P. Pernot, M. Rodio, A. Tenenbaum, and A. Bianconi, Continuous temperature-induced compaction of the intrinsically disordered tau protein (Preprint).

[133] U.K. Laemmli, Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4; Nature, **227**, 680-685 (1970).

[134] P. Pernot, P. Theveneau, T. Giraud, R.F. Nogueira, . Nurizzo, D. Spruce, J. Surr, S. McSweeney, A. Round, F. Felisaz, L. Foedinger, A. Gobbo, J. Huet, C. Villard, F. Cipriani, New beamline dedicated to solution scattering from biological macromolecules at the ESRF; J. Phys.: Conf. Ser. **247**, 012009 (2010).

[135] D.I. Svergun, C. Barberato, and M.H.J. Koch, CRYSOL a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates; J. Appl. Crystallogr. **28**, 768-773 (1995).

[136] P. Bernadò, E. Mylonas, M. V. Petoukhov, M. Blackledge, and D. I. Svergun, Structural Characterization of Flexible Proteins Using Small-Angle X-ray Scattering; J. Am. Chem. Soc. **129**, 5656-5664 (2007).

[137] T. Oroguchi, M. Ikoguchi, and M. Sato, Towards the Structural Characterization of Intrinsically Disordered Proteins by SAXS and MD Simulation; J. of Phys. Conference Series **272**, 012005 (2011).

[138] W. Kabsch and C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features; Biopolymers **22**, 2577-2637 (1983).

[139] R.P. Joosten, T.A.H. te Beek, E. Krieger, M.L. Hekkelman, R.W.W. Hooft, R. Schneider, C. Sander, and G. Vriend, A series of PDB related databases for everyday needs; Nucleic Acids Res. **39**, D411 (2011).

[140] P. Bernadò, L. Blanchard, P. Timmins, D. Marion, R.W.H. Ruigrok, and M. Blackledge, A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering; PNAS **102**, 17002-17007 (2005).

[141] A.E. Garcia, Large-amplitude nonlinear motions in proteins; Phys. Rev. Lett. **68**, 2696-2699 (1992).

[142] A. Amadei, B.L.de Groot, M.-A. Ceruso, A. Di Nola, and H.J.C. Berendsen, A kinetic model for the internal motions of proteins: Diffusion between multiple harmonic wells; Proteins: Struct. Funct. Gen. **35**, 283-292 (1999).