Scuola Internazionale Superiore di Studi Avanzati - Trieste

Molecular Mechanics/ Coarse–Grained hybrid model for investigating protein function

Thesis submitted for the degree of

Doctor Philosophiæ
in Statistical and Biological Physics

Candidate:
Marilisa Neri

Supervisors:
Prof. Paolo Carloni
Prof. Amos Maritan
Prof. Cristian Micheletti

28th february , 2007

SISSA - Via Beirut 2-4 - 34014 TRIESTE - ITALY

Contents

Introduction and Motivation of the present work

One of the most powerful tools in the theoretical study of biological molecules is the method of molecular dynamics simulations (MD) [1]. This computational approach provides detailed information on structures, energetics, dynamics and thermodynamics of biomolecules [1, 2, 3].

Unfortunately, however, MD simulations are restricted to relatively short time scales (typically few tens of nanoseconds). In Fig. (1), the characteristic time scale of protein motions is compared with the approximate range of time that MD simulation is able to cover. It is evident that most functional processes, such as for instance conformational changes, ligand binding and protein folding (Fig. (1)), which are much slower, cannot be followed by MD.

In the last decades, coarse–grained (CG) approaches have attempted to bridge the gap between MD and biologically relevant time scales by using simplified potentials. The development of CG models has a long history in the study of protein structure prediction [4,5,6], thermal fluctuations [7,8,9,10,11] and kinetics [12], in the understanding of DNA supercoiling [13], and in the study of RNA dynamics within the ribosome [14]. For more sophisticated representations, on the other hand, it has become possible to study reliably the insertion and assembly of membrane proteins [15, 16] or the folding of small proteins for longer time scales with acceptable levels of detail [17, 18, 19, 20]. In these approaches, groups of atoms are represented by pseudo–atoms instead of explicitly representing every atom of the system. In such way, the atomistic details are neglected and the spatial degrees

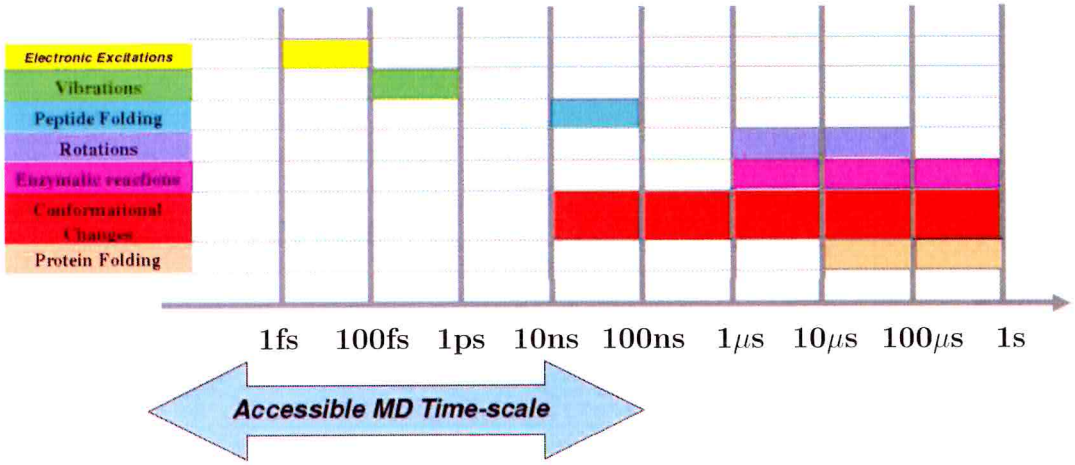


Figure 1: Schematic representation of some protein events and the corresponding range of time in which they occur. The big arrow represents approximately the time scale covered by MD simulation.

of freedom of the protein are reduced. These approaches have been shown to provide an accurate tool for characterizing the vibrational motions of proteins near their native state [11, 21], as well as to investigate protein folding [22, 23], requiring only a modest investment of computational resources. However, since the atomistic details are neglected, these models can not be generally used to investigate all those processes that involve molecular recognition (such as ligand binding or protein–protein interactions). These processes are of outmost relevance in biology, as they are present in metabolic pathway, signaling and protein biosynthesis. In addition, they are central for drug design.

In summary, MD is able to capture the local microscopic interactions between atoms, yet it is restricted to short time scales. CG model is apt to study the mesoscopic dynamics of proteins, such as large scale fluctuations or mechanical deformations along distant sites in protein molecules but CG cannot deal with molecular recognition events.

Based on the observation that, generally, such recognition events involve usually a small portion of protein (*e.g.* in the active site) while the rest can be consid-

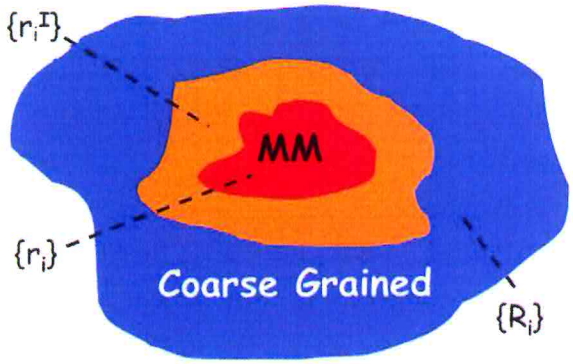


Figure 2: Schematic representation of the regions described by the MM/CG model.

ered as a "scaffold", we implemented a novel hybrid model, MM/CG (Molecular Mechanics/ Coarse Grained model) [24]. The idea is to take the advantages coming from MD and CG overcoming some of their drawbacks.

The MM/CG model uses a hybrid force field in which is included an atomistic–detailed representation of the amino acid residues involved in the ligand binding (*i.e.* the active site) using the MD potential (MM region in Fig. (2)), whereas the rest of the protein is treated at the CG level of detail (CG region in Fig. (2)). An additional region is located between the two MM and CG regions, bridging the discontinuity between full–atom and CG descriptions: the interface region I (orange region in Fig. (2)). The latter is also treated with the MD force field. Chapter 1 is devoted to a detailed description of this hybrid scheme.

One of the main result of this thesis is the construction, the validation and the applications of an MM/CG approach. The novel approach is presented to explore the conformational space of globular and membrane proteins near their native state with low computer resources. We focus on aspartyl proteases [25], which constitute a very large and pharmacologically important enzymatic family, studied for long time in the biomolecular simulation group in SISSA.

The first application [24] deals with two selected cytoplasmatic proteins of the aspartic protease family (AP) [25]: human β–secretase (BACE) [26], a key protein in the Alzheimer's disease development [27] and HIV type 1 virus aspartic pro-

tease (HIV-1 PR) [28], a major target for anti–AIDS therapy [29]. Our MM/CG calculations turn out to reproduce well both the local and the mesoscopic features of these enzymes as obtained by all–atom MD simulations. In particular, the calculations reproduce the structural fluctuations of the substrate in the binding cavity, which play a fundamental role for the enzymatic activity [30, 31].

Next, we investigate the MM/CG dynamic evolution of a membrane protein. The protein chosen as a test system is a membrane protease, the outer membrane protease T (OmpT) [32]. As for HIV-1 PR and BACE, we compare the large scale fluctuations computed with the available MD simulation [33] of the free state [34]. MM/CG approach turns out to be suitable not only for globular proteins, which are immersed in a homogeneous medium (*i.e.* water) but also for this membrane protein, whose environment features are characterized by discontinuities at the water–polar head and apolar chain–polar head interfaces.

Finally, we investigate the dynamics of a membrane protein, OmpT, in complex with its substrate (OmpT/ARRA) [35] over 1 μs, that is a much longer time–scale than those usually covered with MD (Fig. (1)). Our calculations are suggestive of a functional role of large–scale fluctuations of complex [36] (as in BACE and HIV-1 PR). In addition, they point to a polarization of the reactants as a key factor for the reaction. Our MM/CG simulations have also permitted to provide a rationale for the low catalytic efficiency of two mutants of the enzyme (H212A and S99A).

Thus, our MM/CG approach emerges as a powerful predictive tool to investigate μs (potentially function–related) dynamics of enzymes, that may impact on function, as well as an efficient and fast tool for computational structural biology.

Outline of the thesis

Chapter 1 deals with computational methods used in this thesis. Firstly, the MD simulation method and a short overview of Go model based approach [22] are introduced, then our hybrid model, MM/CG [24], is discussed.

Chapter 2 presents the validation of the MM/CG model [24]. The test systems

we chose are two cytoplasmatic Aspartic Protease enzymes: BACE which belongs to pepsin family and HIV-1 PR which belongs to retropepsin family. After a short introduction of the biochemistry of these two proteins of AP class, the validation of our MM/CG model is presented [24]. MM/CG potential is calibrated so as to reproduce the MD data, then we show that our computational approach is able to reproduce both the mesoscopic (*i.e.* large scale fluctuations) and the local microscopic details (*i.e.* chemistry in the active site) of HIV-1 PR and BACE, suggesting that MM/CG can be conveniently applied to those systems for which either the size or the necessity of long–time sampling prevents the application of standard MD techniques.

Chapter 3 applies the MM/CG method to the membrane protease OmpT [34, 36]. The comparison between MD data [33] with those computed by MM/CG simulations on the OmpT in its free state suggests that this approach is well suitable to investigate also membrane proteins [34]. From the MM/CG simulation analysis, it seems that OmpT large–scale conformational fluctuations might play a role for its biological function, as for HIV-1 PR [30] and BACE [31]. In order to elucidate the role of large scale fluctuations, we study the dynamics of OmpT in complex with its substrate using the MM/CG approach [36]. We find that large scale motions and fluctuations of the electric field in the μs time scale may impact on the biological function. Such conclusion can not be drawn within the time scale typical of molecular dynamic simulations and the MM/CG approach is further shown to be a fast and useful tool to provide structure/ function relationships of mutants affecting the enzymatic activity [36].

Two Appendices are present: Appendix A and B.
Appendix A is devoted to the investigation of a particular CG model, the βGaussian model (βGM) [11], which is compared to MM/CG model in Section 2.3.
Appendix B provides further information of our simulation of OmpT in complex with its substrate.

Chapter 1

From MD to MM/CG simulations

This Chapter is devoted to the construction of the MM/CG method [24]. Part of the protein (the MM and I regions) is treated by classical MD, the rest with a simplified Go model (Fig. (2)). Thus, the first part of this Chapter summarizes the principles of all–atom MD simulations [1] and a short introduction to the Go–model [22].

1.1 Principles of molecular dynamics simulations

MD provides the description of the structure and dynamic of molecular system based on the the following assumptions:

(i) The nuclei can be treated as classical particles

(ii) The Born-Oppenheimer approximation holds

(iii) The electronic degrees of freedom can be integrated out

Under these greatly simplifying assumptions, the dynamics of the system can

be described by the second law of mechanics:

$$m_i \frac{d^2\vec{r}_i}{dt^2} = -\nabla_{\vec{r}_i} E_{MM}(\vec{r}_1, ..., \vec{r}_N), \tag{1.1}$$

where $\vec{r}_i$ is the position of the i^{th} atom and E_{MM} is the total energy of the system of N particles. Thus, if one knows the initial structure (by experiments or by computer modeling) and provides a velocity distribution consistent with the temperature simulated, one can provides the time-evolution of the system. MD average values of several properties can be evaluated from the resulting trajectory.

The general flowchart of an MD run is:

1. **Reading the initial conditions**
   Potential interaction as a function of atom positions; position $\vec{r}$ and velocities $\vec{v}$ of all atoms in the system

2. **Compute the force**
   The force of any atom:
   $$\vec{F}_i = -\frac{\partial E_{MM}}{\partial \vec{r}_i} ,$$
   is computed by calculating the force between atom pairs

3. **Update configuration**
   The movements of the atoms is simulated by numerically solving Newton's equations of motion:
   $$\frac{d^2\vec{r}_i}{dt^2} = \frac{\vec{F}_i}{m_i} ,$$
   $$\frac{d\vec{r}_i}{dt} = \vec{v}_i \; ; \; \frac{d\vec{v}_i}{dt} = \frac{\vec{F}_i}{m_i}$$

4. **Output step**
   Write positions, velocities, energies, temperature, pressure, *etc...*

5. **back to point 2**

In the followings, the form of E_{MM} used for proteins and some details of MD algorithms are given.

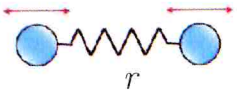
1.1.1 Force Field

The potential energy function (E_{MM}, also called force field) consists of additive parameterized two–bodies terms that can be obtained by fitting experimental and/or high-level quantum chemical data into simple functional forms. Specifically, the potential function usually takes the form of the summation of different additive terms that correspond to bond distances: E_{bonds}; bond angles: E_{angles}; bond dihedral: $E_{dihedrals}$; van der Waals: E_{vdw} and electrostatic interaction: E_{elec}:

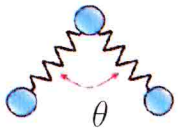
$$E_{MM} = E_{bonds} + E_{angles} + E_{dihedrals} + E_{vdw} + E_{elec} \, . \tag{1.2}$$

Fig. (1.1) shows each term of the force field in Eq. (1.2) in a graphical way. Bond stretching and angle bending are represented as harmonic energy terms where r_{eq} and θ_{eq} refer to equilibrium bond lengths and angles, k_r and k_θ are the vibrational constants. The equilibrium values of the bond and angle parameters are usually derived from structural databases, while the force constants are derived from infrared spectroscopy. In the third term, V_n is the torsional barrier corresponding to the n^{th} barrier of a given torsional angle with phase γ: dihedral parameters are calibrated on small model compounds, comparing the energies with those obtained by quantum chemical calculations. Improper dihedral angles are added to take into account quantum effects that are not present in E_{MM} in Eq. (1.2) as, for example, to preserve planarity in aromatic rings. The last two terms refer to non-bonded van der Waals and electrostatic interactions: the first are described by a Lennard-Jones potential, containing an attractive and a repulsive term, and parameters are defined so as to reproduce chemical-physical properties. The electrostatic energy is evaluated by assuming the dielectric constant ϵ equal to 1, and using the restrained electrostatic potential model [37,38] to define partial atomic point charges: in this model, charges are assigned to the atom-centered points so as to fit the electrostatic potential derived from quantum chemistry calculations for a set of small representative molecules. Van der Waals and electrostatic interactions are calculated between atoms belonging to different molecules or for atoms in the same molecule separated by at least three bonds.

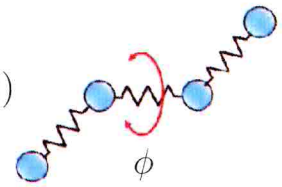
$$E_{bonds} = \sum_{bonds} \frac{1}{2} k_r (r - r_{eq})^2$$



$$E_{angles} = \sum_{angles} \frac{1}{2} k_\theta (\theta - \theta_{eq})^2$$



$$E_{dihedrals} = \sum_{dihedrals} \frac{1}{2} V_n (1 + \cos(n\phi - \gamma))$$



$$E_{vdw} = \sum_{pairs} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right)$$

$$E_{elec} = \sum_{pairs} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

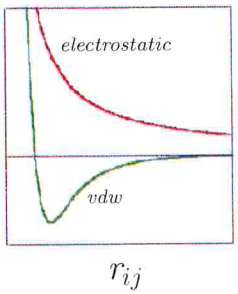


Figure 1.1: Pairwise terms of the MD potential energy in Eq. (1.2).

1.1.2 Integration of the equation of motion: Verlet algorithm

During the MD simulations, the system evolves according to the classical Newton equation of motion:

$$\frac{d^2 \vec{r}_i(t)}{dt^2} = \frac{\vec{F}_i(t)}{m_i} \,, \tag{1.3}$$

where

$$\vec{F}_i(t) = -\vec{\nabla}_{r_i} E_{MM}(t) \,. \tag{1.4}$$

The Eq. (1.3) is refereed to a system of coupled differential equations and it is unrealistic to expect to perform step–by–step numerical integration of them using some numerical methods to accurately follow the true trajectory. In order to

overcame the drawbacks coming from the exact integration of Eq. (1.3), suitable approximated algorithms are implemented. For instance, to numerically integrate the equation of motions, one can use the approximation known as the Verlet algorithm in which the positions of each atom are expressed by Taylor expansions whose combination yields an expression for $\vec{r}_i(t+\Delta t)$. The lack of explicit velocities in the Verlet algorithm is remedied by the leap-frog algorithm [39]. Positions at times $t+\Delta t$ and t are given by the Taylor expansions around $t+\Delta t/2$ below:

$$\vec{r}_i(t+\Delta t) = \vec{r}_i(t+\Delta t/2) + \vec{v}_i(t+\Delta t/2)\frac{\Delta t}{2} + \frac{1}{2}\vec{a}_i(t+\Delta t/2)\frac{\Delta t^2}{4} + \dots , \quad (1.5)$$

$$\vec{r}_i(t) = \vec{r}_i(t+\Delta t/2) - \vec{v}_i(t+\Delta t/2)\frac{\Delta t}{2} + \frac{1}{2}\vec{a}_i(t+\Delta t/2)\frac{\Delta t^2}{4} + \dots . \quad (1.6)$$

The difference between Eq. (1.5) and Eq. (1.6) gives:

$$\vec{r}_i(t+\Delta t) = \vec{r}_i(t) + \vec{v}_i(t+\Delta t/2)\Delta t . \quad (1.7)$$

Analogously, an expression for the velocities at half time step is obtained from the Taylor expansions around t:

$$\vec{v}_i(t+\Delta t/2) = \vec{v}_i(t) + \vec{a}_i(t)\frac{\Delta t}{2} + \dots , \quad (1.8)$$

$$\vec{v}_i(t-\Delta t/2) = \vec{v}_i(t) - \vec{a}_i(t)\frac{\Delta t}{2} + \dots , \quad (1.9)$$

whose difference between Eq. (1.8) and Eq. (1.9) gives:

$$\vec{v}_i(t+\Delta t/2) = \vec{v}_i(t-\Delta t/2) + \vec{a}_i(t)\Delta t . \quad (1.10)$$

The leap-frog algorithm [39] updates both velocities and positions using the forces $\vec{F}_i(t)$ determined by the positions at time t:

$$\vec{v}_i(t+\Delta t/2) = \vec{v}_i(t-\Delta t/2) + \frac{\vec{F}_i(t)}{m_i}\Delta t , \quad (1.11)$$

$$\vec{r}_i(t + \Delta t) = \vec{r}_i(t) + \vec{v}_i(t + \Delta t/2)\Delta t \ , \tag{1.12}$$

respectively.

First, the velocities are calculated at half time step with Eq. (1.11), then these are used to calculate the positions at one time step with Eq. (1.12). In this way, positions and velocities leap each other. The advantage of this algorithm is that velocities are explicitly calculated, and the disadvantage is that they are not calculated at the same time as the positions. As a consequence, kinetic and potential energy cannot directly compute the total energy at time t, but energy evaluation is possible using the following approximate value of velocities at time t:

$$\vec{v}_i(t) = \frac{\vec{v}_i(t - \Delta t/2) + \vec{v}_i(t + \Delta t/2)}{2} \ . \tag{1.13}$$

1.1.3 Constraint algorithm

In biomolecular simulations often we are not interested in describing bond stretching vibrations. Because those are the modes with highest frequency, constraining these bond lengths allows to increase Δt with small affects on the accuracy of the simulation, under the assumption that bonds are almost exclusively in their vibrational ground state.

The SHAKE algorithm [40] changes a set of unconstrained coordinates to a set of coordinates that fulfill a list of distance constraints. It is possible solving a set of Lagrange multipliers in the constrained equations of motion.

If:

$$\sigma_k(\vec{r}_1, ..., \vec{r}_N) = 0 \qquad k = 1, ..., N_c \ , \tag{1.14}$$

are N_c holonomic constrains (one for each constrained bond, *e.g.* $(\vec{r}_1 - \vec{r}_2)^2 - b^2 = 0$, where $\vec{b}$ is the equilibrium distance), the forces are defined as:

$$\vec{F}_i = -\vec{\nabla}\left(E_{MM} + \sum_{k=1}^{N_c} \lambda_k(t)\sigma_k(\vec{r}_N)\right) \qquad i = 1, ..., N \ , \tag{1.15}$$

where λ_k are Lagrange multipliers which must be solved to fulfill the constraint equations. The second part of the sum in Eq. (1.15) determines the constraint forces $\vec{G}_i$, defined by:

$$\vec{G}_i = -\sum_{k=1}^{K} \lambda_k \frac{\partial \sigma_k}{\partial \vec{r}_i} \,. \tag{1.16}$$

The displacement due to the constraint forces in the leap frog or Verlet algorithm is equal to $(G_i/m_i)(\Delta\ \mathrm{t})^2$. Solving the Lagrange multipliers (and hence the displacements) requires the solution of a set of coupled equations of the second degree and these are solved iteratively by SHAKE [40].

1.1.4 Thermostats: Berendsen temperature coupling

To perform simulations at constant temperature, one can use the weak coupling method (Berendsen thermostat [41]) or the coupling to an external bath algorithm (Nosé Hoover thermostat [42, 43]).

In the weak coupling method [41] used here, the temperature of the system $T(t)$ is kept close to the target temperature T_0 by the equation:

$$\frac{dT(t)}{dt} = \frac{T_0 - T(t)}{\tau_T} \,, \tag{1.17}$$

where $T(t)$ is the instantaneous temperature, T_0 is the reference temperature and τ_T is the coupling time constant. Eq. (1.17) means that a temperature deviation decays exponentially with a time constant τ_T.

The instantaneous temperature $T(t)$ of a system with N_{df} degrees of freedom is related to the kinetic energy $E_{kin}(t)$:

$$E_{kin}(t) = \sum_{k=1}^{N} \frac{1}{2} m_k v_k^2(t) = \frac{1}{2} N_{df} k_B T(t) \,. \tag{1.18}$$

Starting from Eq. (1.18), the atomic velocities can than be scaled by a factor $\lambda(t)$

as follow:

$$\delta E_{kin}(t) = [\lambda(t)^2 - 1]\frac{1}{2}N_{df}k_B T(t) \ . \tag{1.19}$$

Since:

$$\delta E_{kin}(t) = N_{df}c_V \delta T(t) \ , \tag{1.20}$$

then:

$$\lambda(t) = \sqrt{1 + \frac{2c_V dt}{k_B \tau_T}\frac{T_0 - T(t)}{T(t)}} \ , \tag{1.21}$$

where c_V is the heat capacity per degree of freedom. The factor $\lambda(t)$ in Eq. (1.21) is used to scale the velocities v_k at each integration step dt, in order to relax the temperature toward the target temperature value T_0. The relaxation rate is controlled by the time coupling constant τ_T. The Berendsen algorithm is stable up to $\tau_T \sim dt$, being dt the integration time step.

1.2 The Go model

Go–type models [22] provide minimal yet fairly realistic CG models of proteins [23, 44, 45].

There are several variants of the Go models that are set in the continuum space. For instance, Zhou and Karplus [12] and Dokholyan et al. [46] have considered models with a square well potential which allow for a simplified discretized time evolution. Wolynes et al. [47] have implemented an associative memory Hamiltonian in which the contact potentials are assumed to be the Gaussian functions. Clementi et al. [48], on the other hand, have studied the 12 10 power law potentials. Here, we focus on the Go–like approach based on the Lennard-Jones potentials [49].

The target native conformation is represented by beads on a chain. The coordinates of the beads are taken as positions of the C_α. The potential energy of the system has the following form:

$$E_p(\vec{r}_i) = E^{BB} + E^{NAT} + E^{NON} + E^{ST} \ . \tag{1.22}$$

The first term represents rigidity of the backbone potential, the second term corresponds to interactions in the native contacts and the third term to those in the non-native contacts. Finally, the last term corresponds to the steric constraints. Two monomers are assumed to be in a native contact if their distance in the native conformation is smaller than some value d_{nat}. Usually, one chooses $d_{nat} = 7.5$ Å.

The backbone potential takes on the form of a sum over harmonic [50] and anharmonic [51] interactions:

$$E^{BB} = \sum_{i=1}^{N-1} [k_1(r_{i,i+1} - d_0)^2 + k_2(r_{i,i+1} - d_0)^4] \,, \tag{1.23}$$

where $r_{i,i+1} = |\vec{r}_i - \vec{r}_{i+1}|$ is the distance between two consecutive beads; $d_0 = 3.8$ Å, $k_1 = \epsilon$ and $k_2 = 100\epsilon$, where ϵ is the Lennard–Jones energy parameter corresponding to a native contact.

The interaction between residues that form a native contact in the target conformation is taken to be of the Lennard–Jones form:

$$E^{NAT} = \sum_{i<j}^{NAT} 4\epsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \,, \tag{1.24}$$

where the sum is over all pairs of residues i and j (but those which are immediate neighbors along the chain) which form the native contacts in the given target conformation. $r_{ij} = |\vec{r}_i - \vec{r}_j|$ is the monomer to monomer distance. The parameters σ_{ij} are chosen in a way that each contact in the native conformation is stabilized at the minimum of the potential. Essentially, $\sigma_{ij} = 2^{-1/6} \cdot d_{ij}$, where d_{ij} is the corresponding native contacts length. Residues that do not form the native contacts interact via a repulsive soft core potential [52], where the potential falls to 0 after some cut–off distance, d_{cut}, which improves foldability. The purpose of introducing the cut–off distance is to make sure that the target conformation is

in fact the ground state of the system.

$$E^{NON} = \sum_{i<j}^{NON} E_{ij}^{NON} \ , \tag{1.25}$$

$$E_{ij}^{NON} = \begin{cases} 4\epsilon \left[\left(\frac{\sigma_0}{r_{ij}} \right)^{12} - \left(\frac{\sigma_0}{r_{ij}} \right)^{6} \right] + \epsilon \ , & r_{ij} < d_{ij} \\ 0 \ , & r_{ij} \geq d_{ij} \end{cases} \ . \tag{1.26}$$

Here, $\sigma_0 = 2^{-\frac{1}{6}} \cdot d_{cut}$. For distances shorter than d_{cut} the potential is purely repulsive. $d_{cut} = \langle d_{ij} \rangle$ [52] which is a mean value of the lengths of the contacts.

The steric constraints can be represented by:

$$E^{ST} = E_{angle} + E_{dihedrals} \ , \tag{1.27}$$

where the terms correspond to the bond angle and dihedral angle potentials respectively, and they take the same shape as in Eq. (1.2)(see E_{angles} and $E_{dihedrals}$ in Fig. (1.1)).

1.3 Development of MM/CG method

In this section it is presented a novel simulation method for globular proteins that is a hybrid approach: Molecular Mechanics /Coarse–Grained (MM/CG) [24]. Within this model, the small biologically relevant region of the protein is treated at the level of detail allowed by classical MD, while the rest of the protein is treated at the CG level, by only considering C_α centroids. An interface region is located between the two MM and CG regions, bridging the large discontinuity between full–atom and CG descriptions (see Fig. (2)).

In the MM/CG method, the MM region (set of atoms $\{\vec{r}_i\}$) is treated using a standard molecular force-field, the CG region (set of atoms $\{\vec{R}_i\}$) using a simplified Go potential [22] and the interface region (I, set of atoms $\{\vec{r}_i{}^I\}$), located between the first two (see Fig. (2)), is also treated with the MM force

field. Solvent-protein interactions are treated with implicit model in terms of viscosity and environment random forces in the framework of stochastic dynamics (SD) [53], that is a langevin dynamic [1]. If the I and MM regions are solvent exposed, the solvent is treated in a explicit way, in the same framework of the MM/CG approach: a drop of water is centered around the MM and I regions and if a molecule exits from the water shell, its velocity is reflected towards the inside. Within this approach, water properties are very similar to those of the bulk water in proximity of the all–atom region, but approaching the drop border located approximately at the interface region, the water density lowers, providing a rough approximation of the bulk behaviour.

Within MM/CG approach, the total potential energy of the system reads:

$$V = E_{MM} + E_{CG} + E_I + E_{MM/I} + E_{CG/I} + E_{SD} , \tag{1.28}$$

where the first three terms represent, respectively, the interactions within the MM, CG and I regions, whereas the fourth and fifth represent the cross-terms potentials. The last term, E_{SD}, mimics the stochastic and frictional forces acting on the system, due to the solvent, proportional to the particle velocity and mass [53,54], as in Eq. (A.6). E_{MM} is represented by the GROMOS96 43a1 force field, with only polar hydrogens explicitly considered [55]. The shape of E_{MM} is the same of that in Eq. (1.2). Details of the force field are discussed in Section 1.1.

E_{CG} takes the following form:

$$\begin{aligned} E_{CG} &= \frac{1}{4}\sum_i K_b(|\vec{R}_i - \vec{R}_{i+1}|^2 - b^2_{i\ i+1})^2 + \\ &+ \sum_{i>j} V_0 \left[1 - \exp\left(-B_{ij}(|\vec{R}_i - \vec{R}_j| - b_{ij})\right)\right]^2 . \end{aligned} \tag{1.29}$$

The first term in Eq. (1.29) takes into account bonded interactions between consecutive CG (C_α) centroids, identified by the position vectors $\vec{R}_i$ and $\vec{R}_{i+1}$, and K_b is the relative bond force constant. b_{ij} is the equilibrium distance, corresponding to the native distance between CG atoms. The second term in Eq. (1.29)

[1] More details about the langevin equation, written in Eq. (A.6), are in Section A.1.

describes the non-bonded interactions between CG atoms. V_0 is the interaction well depth. B_{ij} is the modulating exponent of the Morse potential.

In region I, all atoms are explicitly considered, as in the MM part, and both E_I and $E_{MM/I}$ energy terms have the same form of E_{MM} in Eq. (1.2).

At the interface between I and CG regions, we have to ensure the protein backbone integrity. Thus, we impose bonds between consecutive C_α belonging to I and CG regions. Similarly to the first term in Eq. (1.29), we take

$$E_{CG/I}^{bonded} = \frac{1}{4}\sum_{i,j} K_b(|\vec{r}_{i,C_\alpha}^{\,I} - \vec{R}_j|^2 - b_{ij}^2)^2 \,. \tag{1.30}$$

To this, a term describing the non-bonded interactions is added. It reads:

$$E_{CG/I}^{non-bonded} = \frac{1}{2}\sum_{i\in[C_\alpha,C_\beta],j} V_0\left[1 - \exp\left(-B_{ij}(\left|\vec{r}_i^{\,I} - \vec{R}_j\right| - b_{ij})\right)\right]^2 \,, \tag{1.31}$$

where the interface ith atom is either a C_α or a C_β atom and the factor 1/2 stands for the interaction energy equally distributed between the two types of atoms. All the coefficients are chosen similarly to those in Eq. (1.29).

1.3.1 Analysis of MM/CG trajectories

In the following, data from MM/CG trajectories used to calculate several structural and dynamic properties are exposed.

Displacement

The conformational stability of a macromolecule can be estimated by the Root Mean Square Displacement (RMSD). The RMSD of a set of N atoms at time t, with respect to the initial conformation, reads:

$$\mathrm{RMSD}(t) = \sqrt{\frac{\sum_{i=0}^{N} |\vec{r}_i(t) - \vec{r^0}_i|^2}{N}} \,, \tag{1.32}$$

where $|\vec{r}_i(t) - \vec{r^0}_i|$ is the displacement of the i^{th} atom at time t from the reference position $\vec{r^0}_i$. An increase of the RMSD indicates that the protein moves to a conformation different from the initial structure and thus suggests an incomplete sampling or a conformational change.

Fluctuations

The Root Mean Square Fluctuation (RMSF) is computed for each atom i:

$$\mathrm{RMSF}_i = \sqrt{\langle (\vec{r}_i - \langle \vec{r}_i \rangle)^2 \rangle} \,, \tag{1.33}$$

where $\vec{r}_i$ is the position vector of the i^{th} atom and the $\langle \, . \, \rangle$ symbol stands for a temporal average. This quantity can be compared to crystallographic B–factor B_i through the relation:

$$B_i = \frac{8}{3}\pi^2 \mathrm{RMSF}_i^2 \,. \tag{1.34}$$

The RMSF analysis provides information about the atomic fluctuations throughout the simulation and thus indicates the most flexible regions of the protein.

Large-scale collective motions

Functional proteins are generally stable mechanical constructs that allow certain types of internal motion to enable their biological function. The internal motions may allow the binding of a substrate or coenzyme, or the transmission of a conformational adjustment to affect the binding or reactivity at remote site, as in allosteric effects. Such functional internal motions may be subtle and involve complex correlations between atomic motions, but their nature is inherent in the structure and interactions within the molecule. It is a challenge to derive such motions from the molecular structure and interactions, to identify their functional role, and to reduce the complex protein dynamics to its essential degrees of freedom.

The description of the protein dynamics in terms of collective motions is particularly useful to dissect from an MD trajectory the structural low frequency vibrations, which are often relevant for the protein function. To identify the

concerted large-scale structural fluctuations occurring in a protein, the degree of correlation of pairs of residues is calculated in the covariance matrix $\mathcal{C}$, whose elements are:

$$\mathcal{C}_{ij} = \langle (\vec{r}_i - \langle \vec{r}_i \rangle) \cdot (\vec{r}_j - \langle \vec{r}_j \rangle) \rangle \ , \tag{1.35}$$

where $\vec{r}_i$ is the i^{th} atom position and $\langle \ . \ \rangle$ indicates temporal average.

Large–scale motions can be calculated as eigenvectors of the C_α atoms covariance matrix in Eq. (1.35). The symmetric matrix can be diagonalized by an orthonormal coordinate transformation R which transforms the matrix $\mathcal{C}$ into a diagonal matrix Λ of $3N$ eigenvalues λ_i:

$$R^T \mathcal{C} R = \Lambda \ . \tag{1.36}$$

The columns of R are the eigenvectors or Essential Modes (EM) and the eigenvalues λ_i are the variance in the direction of the corresponding EM. The covariance analysis defines a new coordinate system for the data set, in which the new coordinates are uncorrelated [56,57,58]. The six roto–translational degrees of freedom can be eliminated by performing a RMSD fit between each MD snapshot and the initial configuration before calculating $\mathcal{C}$.

Eigenvectors are ordered with decreasing eigenvalues, which are proportional to the percentage of total fluctuation the corresponding EM describes ($\lambda_i / Tr\ \Lambda$). Usually, it turns out that most of the total motion is spanned over the first few eigenvectors. The amount of motion covered by a subset of EMs is the summation of the corresponding eigenvalues λ_i divided by the trace of Λ.

The principal component of a EM is the projection of the eigenvector of covariance matrix $\mathcal{C}$ on the trajectory.

Chapter 2

Testing the MM/CG model: Cytoplasmatic Aspartic Proteases

This Chapter is devoted to the validation of the MM/CG method. We test the method on two cytoplasmatic protease enzymes: β–secretase (BACE, PDB code: 1FKN1) [26], belonging to pepsin family (clans AA; family A1; Table (2.1)) and HIV type 1 virus aspartic protease (HIV–1 PR, PDB code: 5HVP) [28], belonging to retropepsin family (clans AA; family A2; Table (2.1)). These enzymes have been largely studied in our group in SISSA by all–atom MD simulations [31, 59, 30, 60, 61] and feature the two known folds of cytoplasmatic aspartic proteases.

After a short overview of the biological function and the catalytic pathway of these enzymes, we show our application of the MM/CG model [24]. First, we compare our results with all–atom MD simulations. Specifically, we show that our computational approach is able to reproduce both the mesoscopic (*i.e.* large scale fluctuations) and the local microscopic details (*i.e.* molecular interaction in the active site) of HIV-1 PR and BACE, suggesting that MM/CG can be conveniently applied to those systems for which either the size or the necessity of long–time sampling prevents the application of standard MD techniques. Then, we compare our method with a pure coarse–grained approach, the β–Gaussian

Clan	**Catalytic residues: Asp, Asp (or His)**	**Fold: double β–barrel (or dimer of single β–barrel)**
AA		
Family	A1	Pepsin A (*Homo sapiens*)
	A2	HIV-1 retropepsin (human immunodeficiecy virus type 1)
	A3A	Cauliflower mosaic virus–type endopeptidase (cauliflower mosai virus)
	A3B	Bacilliform virus putative protease (rice tungro bacilliform virus)
	A9	Spumapepsin (human spumaretrovirus)
	A11	Copia transponson (*Drosophila melanogaster*)
	A12	Retrotransponson bs1 endopeptidase (*Zea mays*)
	A16	Tas retrotransponson peptidase (*Ascaris lumbricoides*)
	A17	Pao retrotransponson peptidase (*Bombyx mori*)
	A18	Putative proteinase of Skippy retrotransponson (*Fusarium oxysporum*)
AB	**Catalytic residues: Asp (or Glu), Asn**	**Fold: β–sandwich**
Family	A6	Nodavirus endopeptidase (flock house virus)
	A21	Tetravirus endopeptidase (*Nudaurelia capensis* omega virus)
AC	**Catalytic residues: Asp, Asp**	**Fold: unknown**
Family	A8	Signal peptidase II (*Escherichia coli*)
AD	**Catalytic residues: Asp, Asp**	**Fold: unknown**
Family	A22	Presenilin 1 (*Homo sapiens*)
	A24A	Type IV prepilin peptidase type M1 (*Pseudomonas aeruginosa*)
	A24B	Preflagellin peptidase (*Methanococcus maripauldis*)
AF	**Catalytic residues: Asp, Asp, Asp, His**	**Fold: β–cylinder**
Family	A26	Omptin (*Escherichia coli*)

Table 2.1: Clans and family of aspartic proteases. For each family, the peptidase named is the type example used as the foundation for the family in the MEROPS database (http://merops.sanger.ac.uk) [25].

model (βGM) [62] (more details of βGM are discussed in Appendix A).

2.1 The Aspartic Protease Enzymatic class

Five classes of proteases (also called peptidases) are known, namely, serine proteases, threonine proteases, cysteine proteases, aspartic proteases and metalloproteases [25]. Examples of processes in which protease activity is important include cell growth, cell death, blood clotting, immune defense and secretion [63]. Moreover, pathogenic viruses and bacteria use proteases for their life cycle and

for infection of host cells. For these reasons, the field of protease research is enormous: approximately 8,000 papers related to this field are published each year [25].

Aspartic proteases (AP) are so named because a catalytic Asp dyad or an Asp residue activates a water molecule for the hydrolysis reaction:

$$R - CO - NH - R' + H_2O \rightarrow R - COOH + NH_2 - R' \,. \tag{2.1}$$

Whilst the reaction does not occur in water solutions it happens in the $\sim$ ms in the enzyme.

AP are assigned to clans AA, AB, AC, AD and AF. Tertiary structures solved for members of clans AA, AB and AF each show a unique protein fold unrelated to that of any other protease. For clans AC and AD, there is as yet no crystal structure . The clans and families of aspartic proteases are summarized in Table (2.1), accordingly to MEROPS database classification (http://merops.sanger.ac.uk) [25].

The substrate–binding cleft of the cellular aspartic peptidases extends several amino acid residues. For example in family A1 from P5 to P3' approximately (Fig. (2.1)) [1]. This cleft is formed where the N–terminal domain and the C–terminal domain meet.

2.2 Cytoplasmatic Aspartic Proteases

2.2.1 Structure and biological function

Two folds have been discovered so far for cytoplasmatic AP found in eukaryotes. The first is the **pepsin family**. Enzymes belonging to this class have a molecular weight of $\approx$ 35 kd. The enzymes are approximately 330 amino acids long, with

[1] In the descriptions of the specificity of the proteases, the symbol "+" has been used to mark the bond that is hydrolyzed, the scissile bond, in the formula of substrate molecules. The subsites are numbered from the catalytic site, S1....Sn towards the N–terminus of the substrate, and S1'....Sn' towards the C–terminus. The residues they accommodate are numbered P1...Pn, and P1'...Pn', respectively, as follows (the catalytic site of the enzyme being marked "*"):
Substrate: – P3 – P2 – P1 + P1' – P2' – P3' –
Enzyme: – S3 – S2 – S1 * S1' – S2' – S3' –.

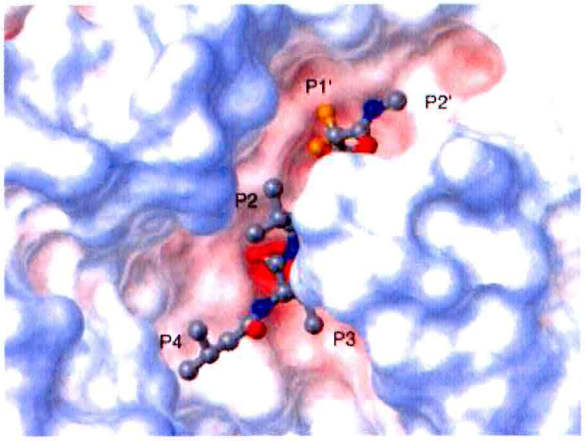


Figure 2.1: A view of the substrate binding cleft. The residues from the flap make intimate contact with the P3 and P1 residues of the substrate (or the inhibitor). Hydrogen bonding and van der Waals interactions position the scissile peptide appropriately for nucleophilic attack by the general-base activated water molecule. Atomic color coding: gray, carbon; blue, nitrogen; red, oxygen; orange, fluorine.

only $\approx 5\%$ of sequence identity between all members of the family.

Pepsins are bi–lobed monomeric proteins, with the active-site cleft located between the lobes (see Fig. (2.2)). The two lobes are linked by a short esapeptide and each of them contributes one of the pair of aspartic acid residues that is responsible for the catalytic activity [65, 66]. The two lobes have similar mainly β–structures, and a similar position in the sequence of the loop that displays the catalytic aspartate. These features strongly indicate that the two lobes are homologous, despite very little amino acid sequence similarity.

In almost all the members of the pepsin family, the catalytic Asp are contained in an Asp–Thr–Gly–Xaa motif (S1–S2–S3–S4)[1], where Xaa is a serine or a threonine, the side-chain of which H–bonds directly to the Asp.

An extended β-hairpin on the N–terminal lobe surface projects across the binding cleft at the junction of the two lobes to form a "flap" that encloses ligands (substrates or inhibitors) into the active site (see Fig. (2.3)). The majority of the members of the family shows specificity for the cleavage of bonds in peptides of at least six residues with hydrophobic amino acids in both the P1 and P1'

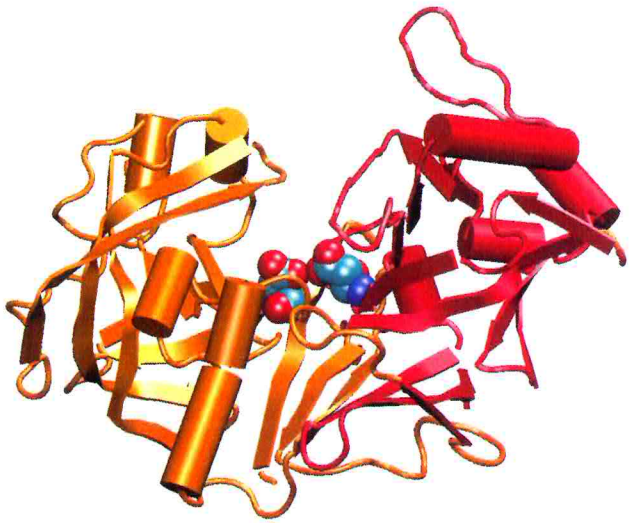


Figure 2.2: 3-D structure of porcine pepsin [64]. The N– and C–lobes are colored in orange and red, respectively. The catalytic aspartic acids are drawn in spheres.

positions[1] [67]. The specificity sub–sites are formed by hydrophobic residues surrounding the catalytic Asp dyad, and by the residues in the flap–turn.

There are seven identified human aspartic proteases. Pepsin and gastricsin participate in digestion in the stomach, whereas cathepsin D and cathepsin E function in intracellular protein degradation. Renin that is clinically important in hemostasis. Cathepsin D has been implicated in the metastasis of breast cancer and in Alzheimer's disease. Two new human aspartic proteases, memapsin 1 and 2, have been cloned [68]. Memapsin 2 has been identified as BACE, a key protein in the Alzheimer's disease development. The intimate involvement of human aspartic proteases in physiology and diseases illustrates their central role in biology and medicine.

The second fold is that of **retropepsin**. It is typically expressed in retroviruses where the genetic information of the virus is stored on a double filament of RNA, in contrast to the other organisms which use DNA. The RNA filament, which is inserted in to the cytoplasm, is protected by a shell of proteins called

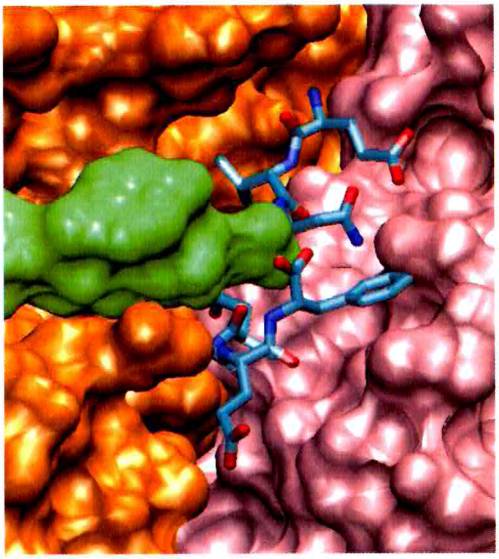


Figure 2.3: Example of ligand binding cleft of pepsins (BACE, in particular). The surfaces of the two lobes are colored in orange and rose, respectively, while the flap is colored in green. The ligand is represented in cylinders.

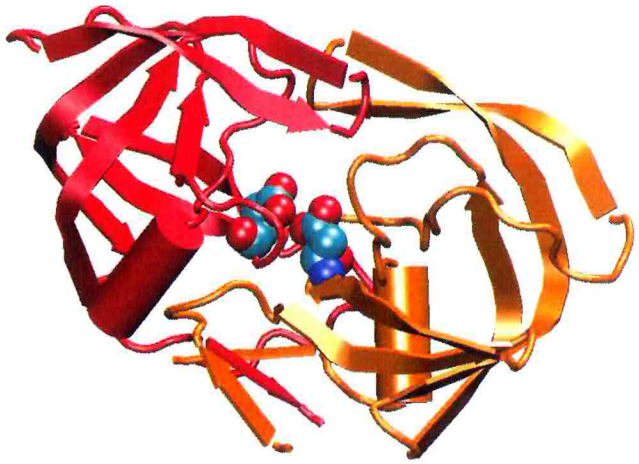


Figure 2.4: 3-D structure of HIV-1 protease. The two symmetric sub-units are colored in orange and red, the catalytic aspartic acids are drawn in balls.

capsid. The out–coming of the structures of the first retropepsin [69] have illuminated the evolutionary relationship between them and pepsins. Whereas the pepsins are single–chain proteins, with an approximate two–fold symmetry, the retro–viral proteases are homo-dimers with two identical subunits related by an exact two-fold axis (see figure 2.4).

Retroviral proteins are usually initially synthesized as polyproteins that have to be cleaved during the maturation process of the virus. This fundamental function for the viral life–cycle is performed by a specific retropepsin. Retropepsins are smaller than pepsins (each domain contains $\approx$ 100-130 aminoacids), whereas the fold of each of their domains resembles that of the N–terminal lobe of pepsins. In particular, each domain has a β-structure, and the external flap is clearly present (Fig. (2.5)). Dimerization of the two subunits occurs by molecular recognition, while the N– and C– terminal residues fold in a 4 strands wide anti-parallel β-sheet, where each chain alternates one strand. Like in pepsins, this structural motif is the only one that cross–links the two subunits.

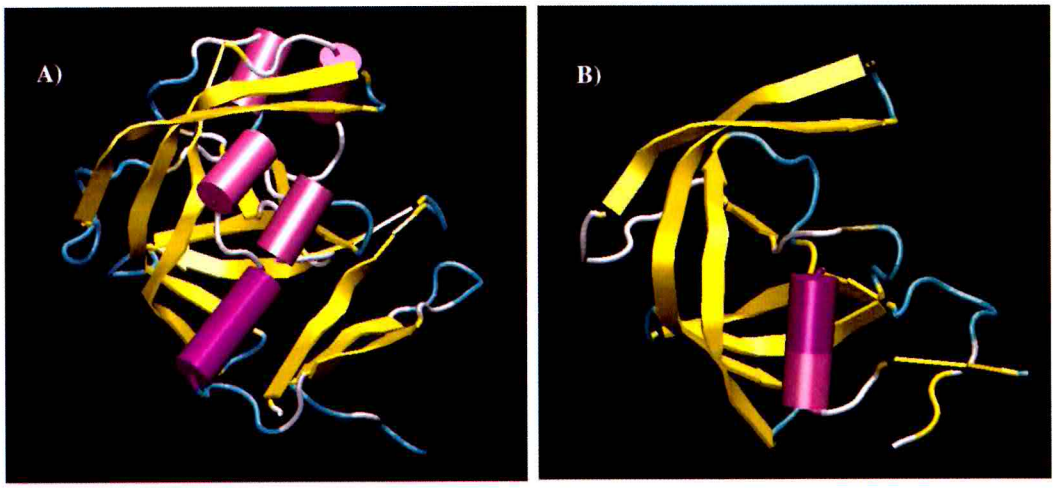


Figure 2.5: Comparison of the N-terminal lobe of a pepsin (A) and a monomeric unit of a retropepsin (B). α-helices are colored in magenta, β-sheets in yellow.

The active site region of retropepsins is very similar to that of pepsins. In particular the already discussed Asp–Thr–Gly–Ala motif (S4–S3–S2–S1)[1] is present in the cleavage loops. The two subunits are separated by a cavity that resembles the groove of pepsins, where the substrate is binding. For effective catalysis, the retroviral APs seem to require at least a seven residues long peptide [70]. These enzymes show unusual specificity in their ability to cleave a peptide bond of polyprotein substrates containing a proline in P_1' [71]. Their action is optimal on oligopeptides containing the Tyr–Pro sequence, although they show some activity against peptides with other aminoacid residues in P_1 and P_1', as, for example, the Met–Met sequence [70].

2.2.2 Mechanism of enzymatic catalysis

The structural aspects of several of AP, as in BACE [26] and in HIV-1 PR [28], confirmed that access to the two active–site aspartate residues from which this family of proteases derives its name, Asp36 (S1) and Asp232 (S1') in BACE

numbering and Asp25 (S1) and Asp25'(S1') in HIV–1 PR, was very restricted due to the limited solvent accessibility of these residues [72, 73].

Since the three–dimensional structures of the catalytic aspartic residues are essentially identical for all aspartic proteases in clan AA, it is generally assumed that this group of enzymes shares a common catalytic mechanism. It is agreed that the aspartic dyad acts as a base as to deprotonate water during the activation process [65, 74, 75]. James *et al.* [75] and Veerapandian *et al.* [74] proposed that the Asp on the C– lobe is the real base that removes one proton from the water molecule, while the Asp on the N–lobe, which is at the beginning in a non–ionized form, donates its proton to the oxygen atom of the carbonyl of the substrate (see Fig. (2.6)). The gem–diol intermediate bounds with both hydroxyl groups to the Aspartic acid of the N–lobe. Then, transfer of the hydrogen atom from the C–lobe Asp to the nitrogen of the scissile bond occurs, forming the two products, and leaving the Asp dyad in the same protonation state of the beginning.

A different mechanism has been proposed by Northrop [65], which postulates the low–barrier hydrogen bond between the two aspartic acids, whose existence has been first predicted in our lab [61], as the key feature of the enzymatic action (see Fig. (2.7)). This mechanism is characterized by a symmetric initial state of the active site, accounts well for the highly symmetric structure of retroviral enzymes, while, in pepsins, structural data would suggest that the N–lobe aspartate is the one that brings the proton.

Same study [76] on the structure of plasmepsin 2 from *Plasmodium falciparum* [77] confirms the possibility that eukariotic APs are indeed able to change their conformation, yet if this might have some influence on their catalytic action is not known. Indeed, theoretical studies by Piana *et al.* [78, 79] and Cascella *et al.* [31] have put into direct correlation, in HIV-1 PR and BACE, such flexibility of the enzymatic scaffold to its catalytic activity. Simulations by McCammon *et al.* [80] have confirmed the flexibility properties HIV-1 PR. Thus, the hypothesis of a general direct correlation between a flexible scaffold and an enzymatic activity in all aspartic proteases seems to be really appealing and reliable. Within this hypothesis, conformational fluctuations of the whole enzyme/substrate complex

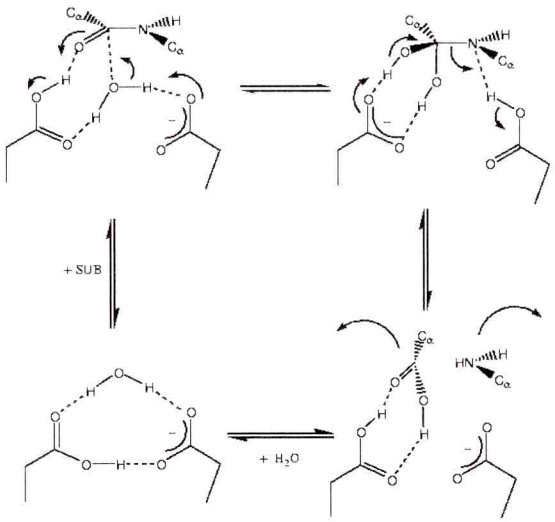


Figure 2.6: General reaction mechanism proposed for catalysis of AP. Starting from the lower–left angle and following the reaction clockwise: Aspartic dyad in the free form; binding of the substrate and nucleophilic attack of water; formation of the tetrahedral *gem*–diol intermediate; protonation of the Nitrogen atom and formation of the products; release of the products and regeneration of the catalyst.

should enhance the catalytic rate by favoring the presence of reactive conformations in the active site along time.

2.3 MM/CG simulations of Cytoplasmatic Aspartic Proteases

2.3.1 Validation of MM/CG model

MM/CG method is tested on the HIV type 1 virus aspartic protease, (HIV-1 PR) [28, 30] and the human β–secretase, (BACE) [26, 31] which, as discussed above, feature the two known folds of cytoplasmatic aspartic proteases. The biological function of these proteins is the catalyzed hydrolysis of peptide chains at specific locations. The dynamical peculiarities of these two proteins make

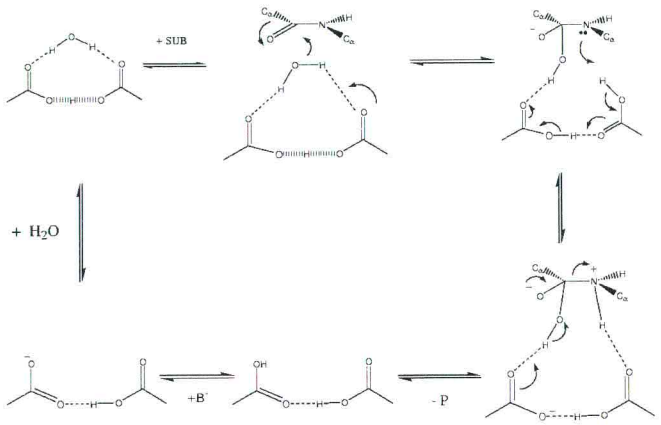


Figure 2.7: Reaction Mechanism as proposed by Nothrop [65]. Several processes occur in the presence of a low–barrier hydrogen bond shared by the Asp dyad.

them a sort of ideal benchmark for our model. In fact, firstly, large–scale motions of these proteins have been well characterized both at an atomistic [30, 31] and a CG level [11, 21]; secondly, these calculations have shown that their large–scale motions are crucially coupled to their enzymatic activity [30, 31]; finally, in spite of the identical cleavage site (a dyad composed of two aspartic residues), these proteins exhibit a large structural diversity: HIV-1 PR is a homodimer containing mostly β–strands, while BACE is a monomer with both α and β secondary structure elements.

HIV-1 PR and BACE MM regions include all residues that play a crucial role in the enzymatic catalysis, namely the aspartic dyad (Asp25 and AspH25' for HIV-1 PR and AspH36 and Asp232 for BACE), the catalytic water molecule and the substrate in the active site (Fig. (2.8)) [30, 31]. The whole systems are composed of 429 and 984 particles for HIV-1 PR and BACE, respectively. The MM/CG simulations are carried out for the timescale of MD simulations: $\sim$ 10 ns and 8 ns for HIV-1 PR and BACE, respectively, required few hours on a Intel xeon

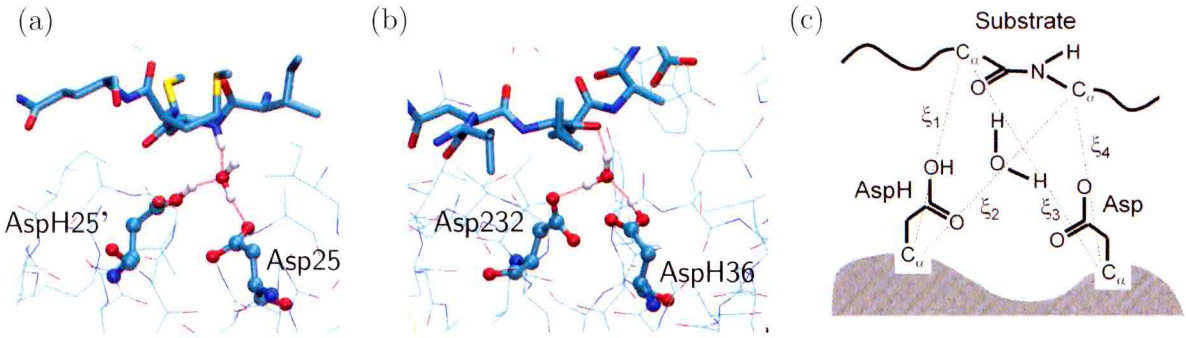


Figure 2.8: HIV-1 PR (a) and BACE (b) active sites. The balls and sticks representation depicts the aspartic dyad with the catalytic water, whereas the licorice representation depicts the substrate (MM region). Red lines represents the H–bond network. The line representation depicts a portion of I region. (c) Schematic representation of the catalytic site for aspartic proteases, ξ_i indicates the distances between the C_α atoms of the aspartic dyad and those of the closest two residues of the substrate.

3.06 GHz PC. In Fig. (2.9) is shown the structure of the two enzymes considered. Different colors depicts the different regions within MM/CG model.

Three parameters were calibrated so as to reproduce the RMSF of HIV-1 PR, calculated with MD and βGM model [11] (Fig. (2.10)a). They are the interaction well depth, V_0 (in Eq. (1.29) and Eq. (1.31)); the cut–off distance within the C_α centroids in CG region interact, r_{cut} and the thickness of the interface I, r_{int}. We have found that the best result are obtained choosing the following values:

$V_0 = 5.3$ kJ mol^{-1};

$r_{cut} = 1.0$ nm;

$r_{int} = 0.6$ nm.

In particular, a very critical task is the choice of the interface thickness. The interface region has to guarantee the geometrical positions and orientation of MM residues, the local electrostatics, and to transmit the modes of vibration of the remainder of the protein, which is mimicked by the CG region. We have then performed MM/CG simulations on BACE and the comparison between our results and those available for BACE and HIV-1 PR (except of course the RMSF of the latter), constitutes an appropriate test for the general validity of our proposed

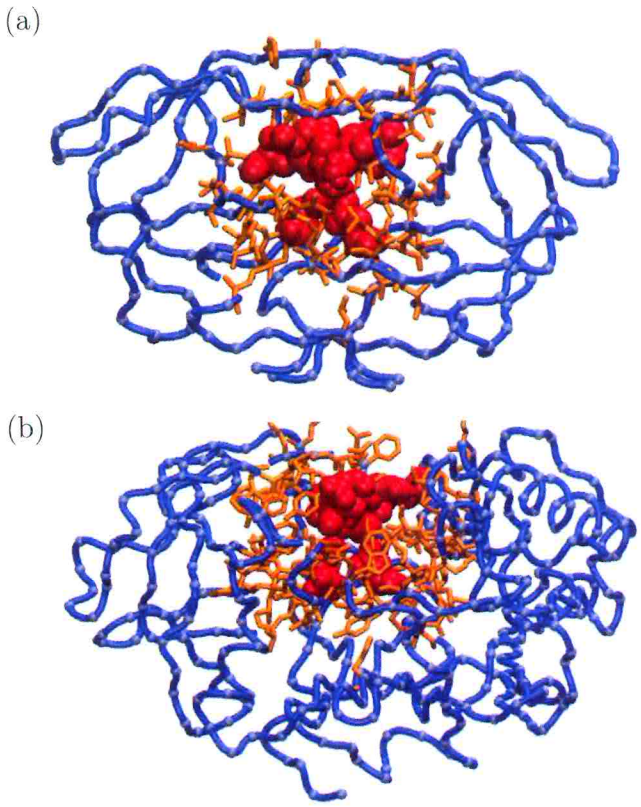


Figure 2.9: HIV-1 PR (a) and BACE (b) structure. The van der Waals red representation depicts the MM residues in the active site; the orange licorice one the residues belong the interface region and the tube representation is the CG region.

MM/CG force field. As a reference CG approach, we chose to adopt the βGM [62], which has been shown to reproduce vibrational modes of both HIV-1 PR and BACE [62, 21]. In addition, we tested whether the MM/CG simulations were able to reproduce local structural features, which can be studied by MD, but not by standard CG models.

Fig. (2.10)b depicted the RMSF of BACE, computed by classical MD, βGM and our MM/CG simulations. As shown in the figure, MM/CG data follow the trend of MD simulations with a straightforward correlation for both proteins. In general, our data agree better than the βGM ones with the MD results. To quantify the accuracy of the vibrational modes of the two proteins obtained by

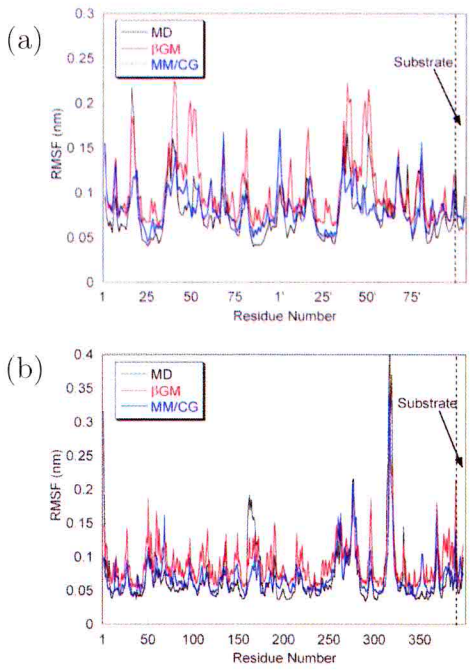


Figure 2.10: RMSF of HIV-1 PR (a) and BACE (b), calculated by means of MD simulations, βGM model calculations and MM/CG simulations.

MM/CG simulations and compare them with βGM calculations, we have represented the meaningful MM/CG (βGM) eigenvectors, $\vec{v}_i$, in terms of the largest N MD eigenvectors, $\vec{v}_j{}^{MD}$, of the corresponding covariance matrices. These matrices, calculated for the C_α atoms, provide information on the degree of correlation between pairs of residues at equal time. $\vec{v}_i$ reads:

$$\vec{v}_i = \sum_{j=1}^{N} c_j^{(i)} \vec{v}_j{}^{MD} \ . \tag{2.2}$$

In particular, we considered the value C_i, defined as the square root of the summation over j of the first N $c_j^{(i)}$:

$$C_i = \sqrt{\sum_{j=1}^{N} \langle \mathbf{v}_j^{MD} | \mathbf{v}_i \rangle^2} \ . \tag{2.3}$$

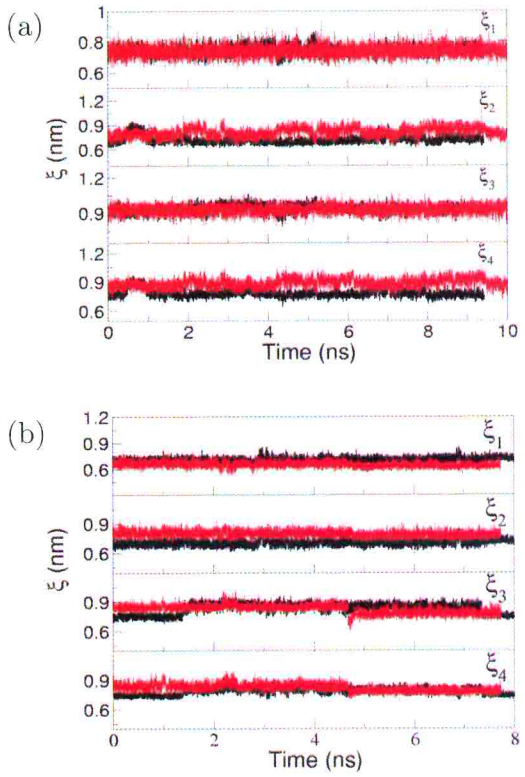


Figure 2.11: Time evolution of the distance between the substrate and the active site for HIV-1 PR (a) and BACE (b). ξ_i are defined as in Fig. (2.8c). Black and red lines refer to MD and MM/CG simulations, respectively.

In this way, it is possible to quantify the ability of the MD eigenvectors to represent the largest MM/CG or βGM ones. The calculated high values of C_i (Table 2.2) indicate that the subspace of the most relevant eigenvectors computed with the MM/CG model and MD almost coincide. For instance, in the case of HIV-1 PR, $C_1 = 0.95$ means that 95% of the first MM/CG eigenvector overlaps with the overall motion of the MD simulation, if the MD trajectory is projected onto the relative first 20 normal modes (*i.e.* $N = 20$). The results are similar to those of βGM, especially for the most important eigenvectors. βGM calculations seem to perform better than MM/CG for HIV-1 PR, but not in the case of BACE.

As a test prediction accuracy of the microscopic dynamical features, *i.e.* the

	HIV-1 PR		BACE	
	MM/CG	βGM	MM/CG	βGM
C_1	0.95	0.97	0.81	0.23
C_2	0.96	0.98	0.86	0.86
C_3	0.92	0.96	0.78	0.87
C_4	0.81	0.92	0.85	0.67
C_5	0.81	0.88	0.87	0.82
C_6	0.85	0.82	0.66	0.84
C_7	0.77	0.88	0.81	0.62
C_8	0.55	0.79	0.75	0.67
C_9	0.55	0.82	0.59	0.64
C_{10}	0.54	0.63	0.65	0.46

Table 2.2: First 10 C_i, calculated as the square root of the sum of the squared principal MD eigenvectors decomposition coefficient, in the case of $N = 20$.

chemical interactions in MM regions, we focused on the motion of the substrate in the active site of the two enzymes, which has been shown to play a functional rule. In fact, the catalytic activity of both HIV-1 PR and BACE is directly related to the distance between the catalytic aspartic dyad and the substrate (ξ_{1-4} in Fig. (2.8)c). These distances are directly modulated by the large–scale motions of the protein scaffold [30, 31] and fluctuate between two characteristic values in a few ns [30, 31]. Fig. (2.11)ab show that MM/CG method is able to reproduce very well these characteristic distances.

Comparing the H–bonds in the active site in BACE (Fig. (2.12)), further establishes the accuracy of our method. The correlation between MD and MM/CG data is very high (0.92).

In summary, our MM/CG simulations are able to reproduce both the mesoscopic (*i.e.* the residue root mean square fluctuations (RMSF) and the principal normal modes) and the local microscopic details (*i.e.* distances between key atoms in the active sites and H–bond patterns) of the two proteins, suggesting that our method can be conveniently applied to those systems for which either the size or the necessity of long–time sampling prevents the application of standard MD techniques. The new computational approach for globular proteins,

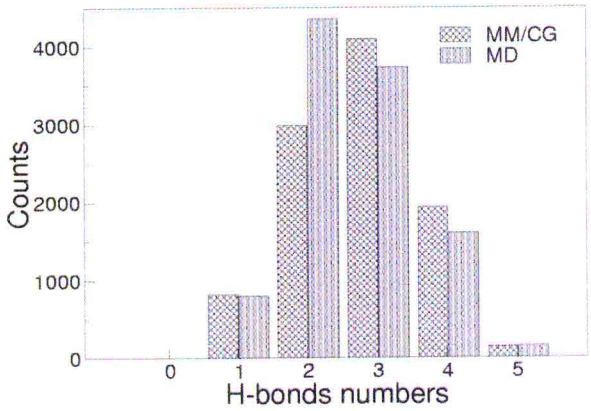


Figure 2.12: H–bond number counts between the aspartic dyad along with the catalytic water and the substrate in BACE. Values were derived by MD and MM/CG simulations.

based on a mixed mesoscopic/ atomistic description may provide a reliable way of overcoming the time/ size–limit bottlenecks that constitute one of the major drawbacks of MD simulations.

Chapter 3

MM/CG simulations of Outer–membrane proteases T

In this Chapter we investigate several aspects regarding a membrane protein, the outer–membrane proteases T (OmpT) [32]. According to the MEROPS classification, this enzyme belongs to aspartic protease (clans AF; family A26; Table 2.1) like HIV-1 PR and BACE, even though it has no sequence homology to other AP.

In the following, we discuss the biological function of OmpT and give the current hypothesis of the enzymatic reaction.

By comparing MD [81] and MM/CG [34] simulation data of OmpT in the free state [32], we establish that the MM/CG model is suitable not only for globular proteins (as HIV-1 PR and BACE) but also for this membrane protein which is immersed in a discontinuous medium. From the analysis of the large scale motions of free OmpT, we further suggest that those motions might play a role for the catalysis, hence sharing some resemblance with those proposed for HIV-1 PR and BACE. MM/CG simulations on the Michaelis complexes of OmpT with the Ala–Arg–Arg–Ala substrate (OmpT/ARRA), further confirm that the substrate motion is strongly correlated with the large scale motions of the entire enzyme, thus meaning that the motion of the peptide inside the cleft is driven by the protein fluctuations [36]. Furthermore, our calculations suggest that the

polarization of the reactants may favor the catalytic reaction [36]. Finally, the MM/CG approach is used to provide structure/ function relationships of protein mutants.

3.1 The Omptin family

3.1.1 Structure and biological function

OmpT is a protease present in the outer membrane of *Escherichia coli*. This enzyme has no sequence homology to other known proteases. The structure of the protein in the free state at 2.6 Å resolution [32] (Fig. (3.1)) highlights an Asp–His together with an Asp–Asp couple in the active site, all which have been shown to be essential for enzymatic activity by mutagenesis [82,83].

OmpT is specific for cleavage between consecutive basic amino acids in protein substrates. Cleavage by OmpT has been described at Arg+Arg [84] Lys+Lys [85], Lys+Arg [86] and Arg+Lys [87].

The omptin gene encodes a protein of 317 amino acids, of which the first 20 are the signal sequence for targeting of the protein into the periplasmic space. The mature protein (297 amino acids, 33 478 Da) folds and inserts into the outer membrane.

The protein is a 10–stranded antiparallel β–barrel. The strands are amphiphatic with the hydrophilic residues pointing inwards into the barrel and the hydrophobic residues exposed to the exterior. The hydrophobic residues form a continuous band of roughly 25 Å in height, which represents the membrane–spanning part of the molecule. On the periplasmic side of the protein short turns and loops connect the strands, whereas at the cell surface displayed part long loops are present. The extracellular part of the molecule contains a large negatively charged groove, which is consistent with the high specificity for positive substrates of the protein [88]. The deep groove is formed by loops L4 and L5 on one side and L1–L3 on the other (Fig. (3.4)a).

For the identification of the active–site residues the role of all conserved serines, histidines and acidic residues have been investigated by mutagenesis [82,83].

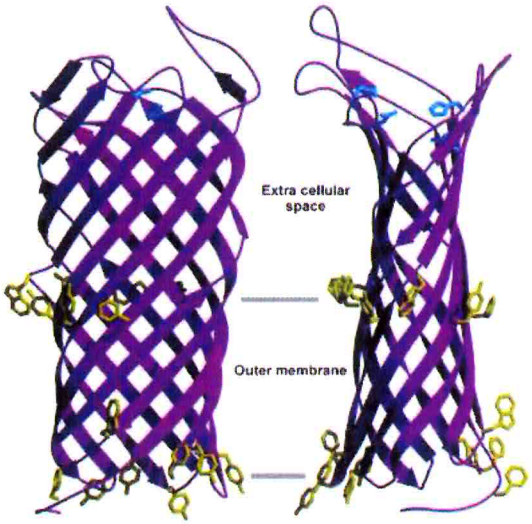


Figure 3.1: Ribbon representation of the structure of OmpT [32] in two orientations at 90°. The extracellular space is located at the top of the figure, and the periplasmatic space is at the bottom. The position of the membrane is delineated by horizontal line between the two orientation. Aromatic residues that are located at the boundary of the hydrophobic and hydrophilic area on the molecular surface and the four catalytic residues located in the active site groove at the top area represented in sticks (in black and light gray, respectively)

His212, Asp83, Asp85 and Asp210 are all essential residues and individual replacement of these residues by alanine one at a time resulted in a 10 000–fold reduction in enzymatic activity. In the crystal structure, a groove is present at the extracellular top of the protein and all these active site residues are located in this region (Fig. (3.1)). Asp210 and His212 are hydrogen bonded to each other, and Asp83–Asp85 are located on the opposite side of the active site groove.

The actual biological function of OmpT remains to be elucidated but several studies suggest that OmpT is involved in urinary tract disease, in DNA excision repair and in the breakdown of antimicrobial peptides [89].

3.1.2 Mechanism of enzymatic catalysis

The presence of a serine and a histidine (Ser99 and His212) in the cleft led to the suggestion that OmpT was a novel–type serine protease [82]. Within this hypothesis, the scissile peptide bond should be attacked by the hydroxyl of the catalytic serine, which in serine protease is usually activated by a histidine residue [90]. Neverthless, an alternative mechanism of action was put forward, based on the following considerations:

(*i*) the two catalytic His and Ser residues are much farther (~ 9 Å) than in actual serine proteases;

(*ii*) several Asp and Glu groups present in the cleft (Fig. (3.4)b) seem to play a key role for catalysis, as shown by molecular biology experiments [82].

In this alternative mechanism, His212 and Asp83 groups activate a water molecule for a nucleophilic attack, while Asp83 and Asp85 contribute to polarize the substrate scissile peptide bond [83,32,91]. Molecular dynamics simulations (MD) on OmpT on 10−ns timescale have further supported this scenario [91,33,92].

In this section, we summarize the currently proposed reaction mechanisms that depends on the protonation state of Asp83 [35].

In these hypothesis, Asp210/His212 represent a catalytic dyad with a water molecule acting as a nucleophile for the hydrolysis reaction [32]. Interestingly, this catalytic dyad has never been observed in proteases, meaning that OmpT would be the first example of a protease using this type mechanism. A peculiar hydrogen bond network orients both the substrate and the nucleophilic water, promoting the cleavage of the peptide bond.

In particular, the Asp83/Asp85 couple plays a crucial role in the catalytic mechanism H–bonding to the nucleophilic water molecule [32,82,83].

If the Asp83 is protonated the transition state of the catalytic reaction is summarized in Fig. (3.2). Within this hypothesis, the proton shared by the carboxyl moieties of Asp83/Asp85 might stabilize the oxyanion intermediate during the reaction [82,83].

If the Asp83 is deprotonated, the latter residue could act like a proton shuttle, deprotonating water at the first step and protonatig the leaving ammine group in the second (see Fig. (3.3)). This proposed mechanism is based on the speculation that we have put forward based on the final configuration at 24 ns of the dynamic simulation performed by Baaden and Sansom [35].

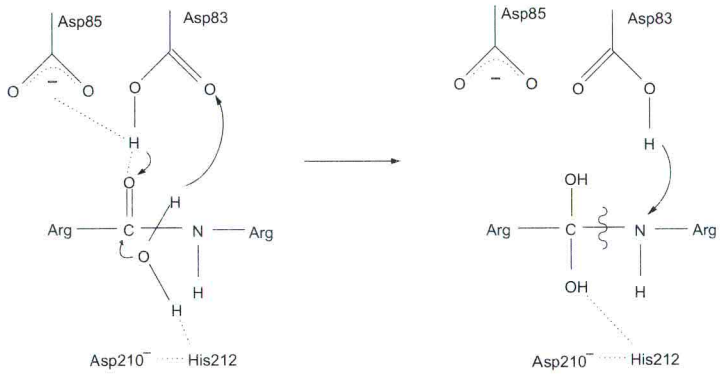


Figure 3.2: Proposed reaction mechanism of OmpT with Asp83 protonated

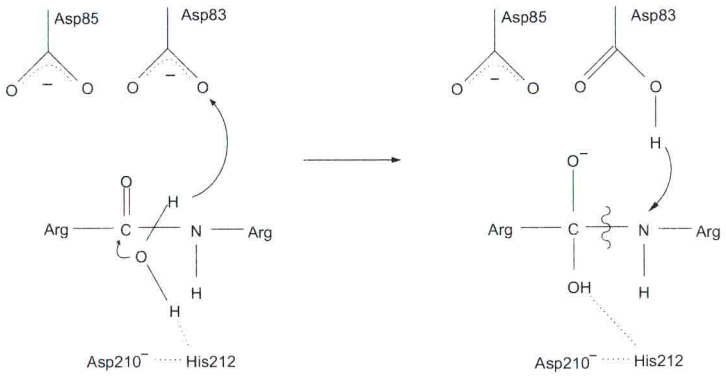


Figure 3.3: Proposed reaction mechanism of OmpT with Asp83 deprotonated

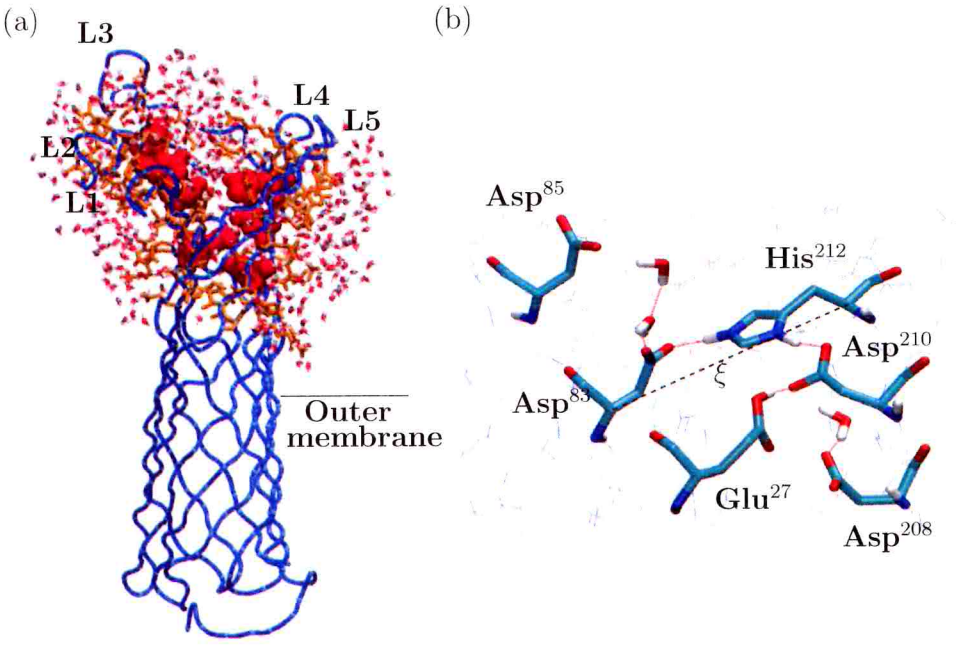


Figure 3.4: a) Structure of OmpT from *Escherichia coli*. The MM, I and CG regions (see figure 2) are represented by red VdW sphere, orange licorice tubes and blue tubes, respectively. Water molecules are also shown. b) OmpT active site: the residues in MM region are depicted with the licorice representation. H-bonds are represented as red lines. The thin lines represent groups included in the I region. The distance between C_α carbon of His^{212} and Asp^{83} (ξ) is represented as a dashed line.

3.2 MM/CG Simulations of OmpT

We investigated the dynamics of OmpT using a MM/CG scheme, as for cytoplasmatic proteases [24]. We provide an all–atom representation of the amino acid residues involved in the ligand binding region of an enzyme (*i.e.* the active site), whereas the remainder of the protein is treated with a widely used coarse–grained (CG) model (Fig. (2)).

As a first step of our investigations, we focus on the enzyme in the free state (Section 3.2.1). Then (Section 3.2.2), we investigate the complex with the sub-

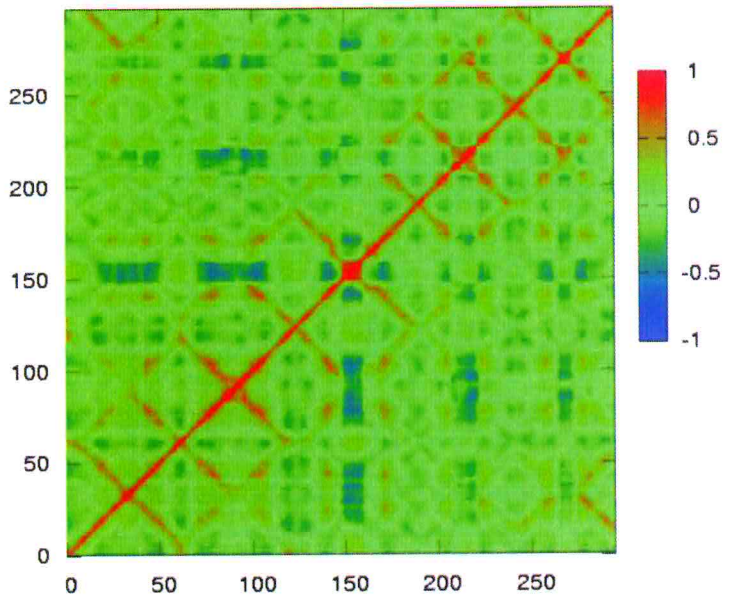


Figure 3.5: Normalized reduced covariance matrix of OmpT C_α atoms.

strate, which provides informations about the catalytic function of the enzyme.

OmpT key residues in the active site (Glu27, Asp83, Asp85, Ser99, Asp97, His101, Asp208, Asp210 and His212) are treated at the atomic detail (MM region). In the case of OmpT complex, also the Ala–Arg–Arg–Ala substrate is in the MM region. A shell of ~ 6 Å constitutes the interface (I) between the MM residues and the CG region, which includes the remainder of the protein. Water molecules are treated explicitely in proximity of the MM and I regions (Fig. (3.4)a and Fig. (3.10)).

3.2.1 Free OmpT

Our initial model is based on the crystal structure of S99A–G216K–K217G OmpT from *E. coli*. These mutations involve residues treated at the all–atom level in our computational scheme. Therefore, we restored the wild–type structure by reversing mutations by using the Swiss–Pdb Viewer program [93]. The pro-

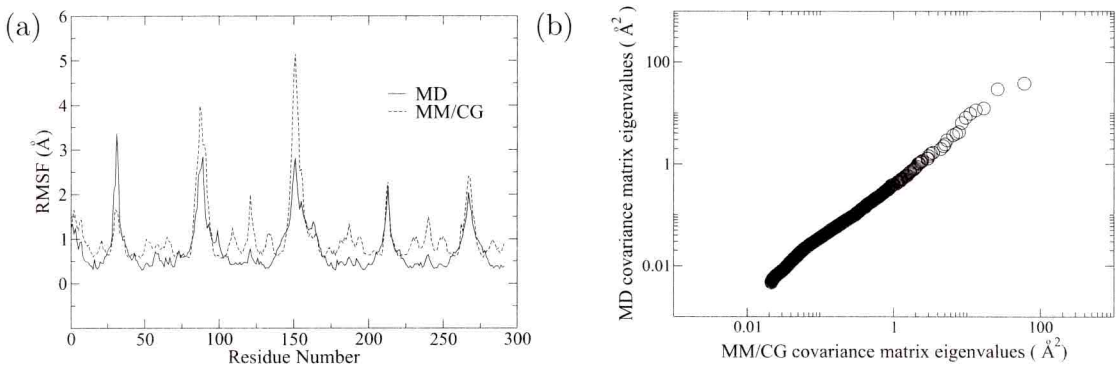


Figure 3.6: a) Root mean square fluctuations of the various residues in OmpT obtained from MD and MM/CG simulations. b) Scatter plot of equally ranking eigenvalues of the MD and MM/CG covariance matrices.

tonation state of the residues present in the catalytic cleft was set as in ref. [91]: Glu^{27} was deprotonated, His^{212} was protonated in ϵ, whilst all the Asp residues were taken as ionized (Fig. (3.4)b). A water layer of 12 Å, centered on the MM region, was added, corresponding to 63 SPC water molecules [94]. As a result, the total system was composed of 2062 atoms.

The simulations were performed using a modified version of the Gromacs 3.2.1 program [54]. The leap–frog stochastic dynamics algorithm was used to integrate the equations of motion with a time step $\Delta t = 2$ fs and a friction coefficient $\gamma_i = m_i/\tau$, where $\tau = 0.5$ ps is the time constant for the coupling and m_i is the mass of ith particle. Temperature was maintained at 300 K using a Berendsen thermostat [95]. The SHAKE algorithm [40] was used to keep fixed the distance of bonds containing hydrogens. No cut–off was used for the non–bonded interactions.

The system was first relaxed with a 100−ps run with positional restraints on solute positions and then simulated for 70 ns without positional restraints.

The property on which this work is focused on are the large–scale conformational fluctuations. Specifically, we calculate concerted motions, which are related to the degree of correlation of pairs of residues at equal times. This information is summarized in the covariance matrix, $\mathcal{C}$, whose elements are given in Eq. (1.35).

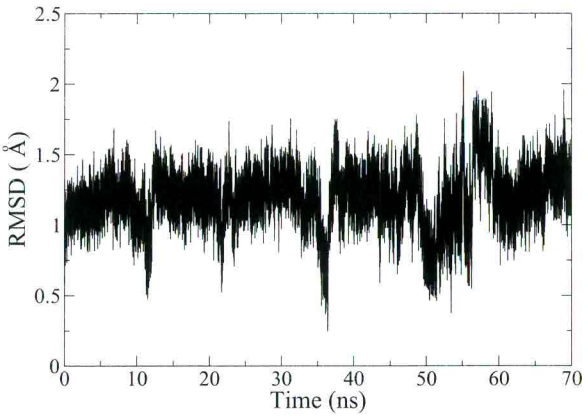


Figure 3.7: RMSD computed on the C_α atoms of the MM residues.

Results

In this Section, we compared our results with those of all–atom molecular dynamics simulation. Then, we investigated the motion of the protein frame during the MM/CG simulation. Finally, we analyzed the structural properties of the important residues for the protein function.

Comparison with previous MD calculations

The covariance matrix, $\mathcal{C}$, (Fig. (3.5)) describes the degree of correlation of pairs of residues at the same time. The accuracy of our method for reproducing the all–atom MD on the protein frame has been established by a comparison with $\mathcal{C}$–derived properties obtained with a 10–ns MD simulation on the same system [33]:

(i) The root mean square fluctuations (RMSF in Eq. (1.33)), given by the square root of the trace of the covariance matrix in Eq. (1.35) (see Section 1.3.1), were well reproduced, with a correlation coefficient between the two sets of data of 0.79 (Fig.(3.6)a). We also compared the RMSF given by the crystallographic structure with those obtained by using the MM/CG simulation, finding a correlation coefficient of ~ 0.66.

(ii) The projections of the MM/CG eigenvectors associated with the five

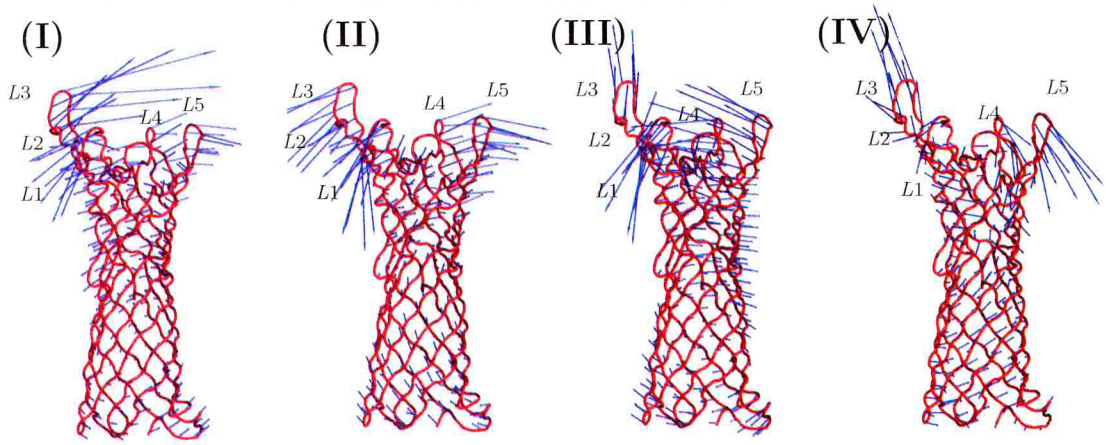


Figure 3.8: The eigenvectors relatives to the largest four eigenvalues are represented (for each of the atomic components) as arrows of width and length proportional to the modulus. The protein is shown in trace representation.

largest eigenvalues of $\mathcal{C}$ (which represent the large–scale motion of the system) onto the the correspondent eigenvectors calculated with MD, overlapped by 75%.

(iii) Finally, the eigenvalues of $\mathcal{C}$ calculated with the two potentials were fitted with a exponent of 0.97 (Fig.(3.6)b), showing that practically the two entries almost coincide. These values are proportional to the characteristic timescale of the collective vibrational excitations. Therefore, the rate of molecular events in MM/CG simulations practically corresponds to that of MD simulations.

In addition, the stability of the protein with respect to the X–ray structure [32] has been shown by the low value of the RMSD of the C_α carbons during the dynamics (Fig. (3.7)).

Large–scale motions of the protein

Based on the calculation of $\mathcal{C}$ (Fig. (3.5)), we suggested that the motion of the groups ranging from 150 to 160 (constituting the loop L3 surrounding the active site, see figure 3.4a) is anticorrelated with that of the scaffold of the protein, in particular with groups 20–50 (belonging to the β–sheets β1 and β2 and the loop L1 bridging them), 70–110 (β3, L4 and β4), 170–180 (β6), 200–210 (β7), 220–230

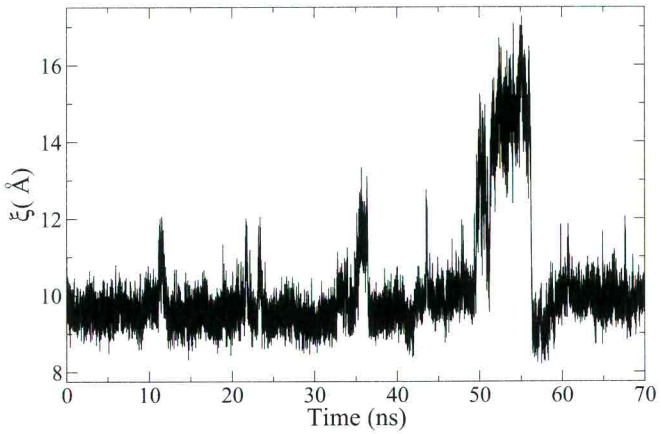


Figure 3.9: Distance between the C_alpha carbons of His^{212} and Asp^{83}, ξ, as a function of time.

(β8), 250–265 (β9), 272–285 (β10). In addition, the β–sheets from β1 to β10 are strongly correlated among them. As a result, L3 moves in counter direction with respect to the compact and rigid β–barrel core (and the bridging loops L1, L2, L4 and L5). This is indicated more clearly in Fig. (3.8), showing the four most significant eigenvectors of the covariance matrix, that account for $\sim$ 50% of the overall residue mobility. These motions might be essential for driving the residues in the active site to a specific configuration and, therefore, could be important for OmpT catalytic activity.

H–bond network in the active site

All the residues putatively involved in the enzymatic hydrolysis [83] interact with each other. During the dynamics, Asp^{210} keeps its original position by H–bonding to His^{212} and by forming a direct or water–mediated H–bond with Glu^{27}. At the same time, Glu^{27} interacts through a water bridge or a direct H–bond with Asp^{208}. Asp^{83} and Asp^{85} interact through (generally) two or three bridging waters. Table (3.1) shows that most of these structural features reproduce X–ray

data.

An important feature of the dynamics in the active site concerns the distance between the key residues His^{212} and Asp^{83} (ξ), which also represents the width of the cleft (Fig. (3.9)). During most of the dynamics, ξ oscillates around an average value of 9.7 Å corresponding to Asp^{83} and His^{212} forming a direct H–bonding interaction. However, at times, ξ increases for a few ns, allowing a water molecule to form a bridge between the two residues. In some cases (*e.g.* at $\sim$ 50 ns) more than one water molecule (two or three) is inserted between Asp^{83} and His^{212} for a short time.

	C_α–C_α MM/CG distance (Å)	C_α–C_α crystallographic distance (Å)
His^{212}–Asp^{83}	10.2 ± 1.5	12.5
His^{212}–Asp^{210}	6.2 ± 0.2	6.5
Asp^{83}–Asp^{85}	6.7 ± 0.4	6.9
Glu^{27}–Asp^{208}	8.4 ± 0.7	8.8
Glu^{27}–Asp^{210}	9.1 ± 0.9	9.2

Table 3.1: C_α–C_α distances of pairs of residues in the active site calculated from the MM/CG simulation and the crystallographic structure. MM/CG values were estimated by averaging the distances over the whole simulation, whereas the errors were estimated by the corresponding standard deviations.

Discussion and Concluding Remark

Our MM/CG calculations on OmpT were able to reproduce both local and mesoscopic properties obtained with MD simulations. As the latter are reproduced by means of a CG representation, this finding points to the suitability of CG approaches not only for globular proteins, which are immersed in a homogeneous medium (*i.e.* water) [24] but also for membrane proteins, whose environment features discontinuities at the water–polar head and apolar chain–polar head interfaces.

Our calculations suggest that the distance between the putative catalytic dyad His^{212} and Asp^{83} [83] periodically (with a 12−ns period) increases from ~ 10 Å to 15 Å. This opening motion origins from the large–scale fluctuations of L2, which are correlated to those of loop L4 (Fig. (3.8))). In fact, His^{212} is located in $\beta 8$, close to the loop L4, whereas Asp^{83} is located in $\beta 3$, close to the loop L2, at the other side of the catalytic cleft. As a result, a water molecule can be accommodated in a suitable position for the enzymatic hydrolysis. At the speculative level, this motion might therefore play a role for catalysis. This mechanism shares some resemblance with those proposed for other proteases (*i.e.* the aspartyl proteases HIV-1 PR [28] and BACE [26]), where the concerted motion of the loops is crucial to drive the substrate in an active conformation [30, 31, 80]. Simulations of the complex with OmpT substrate are required to assess this issue.

3.2.2 OmpT in complex with Ala–Arg–Arg–Ala substrate

In this Section, we expose the MM/CG structural features of the Michaelis complex of OmpT [32, 36].

Test MM/CG calculations on the enzyme in the free state against all–atom MD simulations have shown that the approach is well suited also for a membrane protein [34], see Section 3.2.1. Here, we extend our investigation to the Michaelis complex of the protein with a model substrate (Ala–Arg–Arg–Ala). We carry out simulations on four OmpT/ARRA Michaelis complexes (**A**–**D**), which differ for: (i) the protonation state of the putative catalytic residue D83; (ii) for N– and C– terminal tails, which are considered in the Zwitterionic form (charged state) or capped with acetyl and N–methyl groups (neutral state), respectively, analogously to previous MD calculations [35].

The aim of our MM/CG simulations is twofold. First, we investigate the role of conformational fluctuations of the substrate in the active site in μs time–scale. We find that large–scale motions of the protein and the electrostatic field, evaluated on such time–scale, may impact on the function of the enzyme. Second, we use the MM/CG approach, as a efficient tool to investigate the effects on

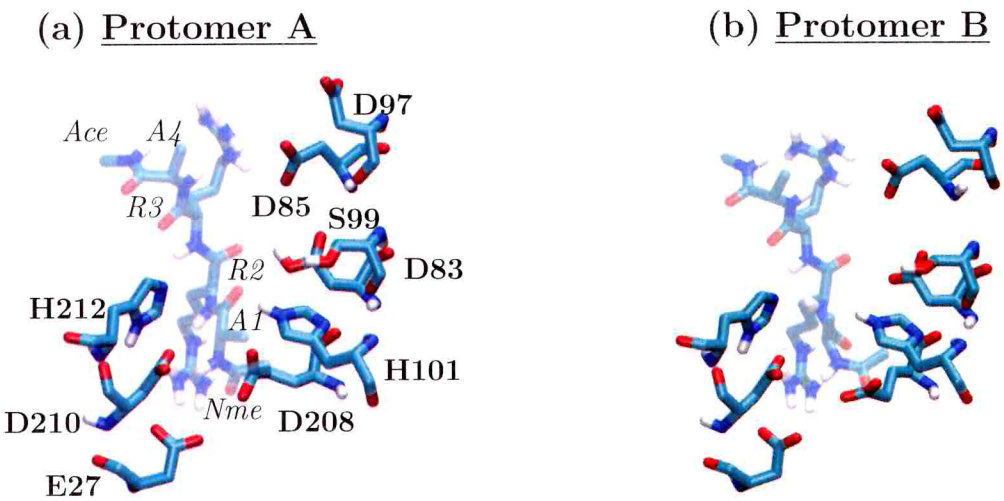


Figure 3.10: (a) and (b): geometry of the active site of **A** and **B**, respectively, after 24 ns of MM simulations [35]. The residues in licorice representation constitute the residues in the MM region. The substrate is depicted with a transparent effect. The catalytic water, which is suggested to perform the hydrolysis reaction, lies between H212 and D83 and it is not shown here for the sake of clarity.

mutations of the enzymatic function. We focus on protein mutants S99A and H212A, which are experimentally known to be much less efficient than the wild–type (wt) (residual activity within 0% and 4% relative to the wt [83]). Whilst H212A clearly affects the chemistry of the cleavage site, H212 residue being in the active site, it is intriguing that also S99, which is not located in the cleavage site (Fig. (3.10)ab), is a very important residue for catalysis. Our calculations, on the ~ 0.1 μs time–scale, provide a structural basis for the dramatic decrease in catalytic activity. Because of its extremely cheap computational cost (two order of magnitude faster than standard all–atom MD), the methodology emerges as a powerful tool to investigate structure/ function relationships of high–throughput site–directed mutagenesis data.

Methods

Hybrid MM/CG simulations were performed using a modified version of the Gromacs 3.2.1 program [24]. In this approach, the enzymatic active site is treated at atomic details with the GROMOS96 43a1 force field [55] (MM region). ~ 400 SPC water molecules [94] are added around the MM region. This corresponds to a water layer of ~ 15 Å (as in Fig. (3.4)a). The rest of the system studied is treated as a modified Go–model (CG region) by including solely the C_α's, formally *in vacuo*. I is the interface region located between the MM and CG regions (Fig. (2)). This approach has been shown to describe accurately structure and dynamics at the active site of aspartyl proteases in complex with their substrate [24] and of OmpT in the free state [34] (discussed in Section 2.3 and 3.2.1, respectively). The computational cost is about two orders of magnitude smaller than that of the corresponding all–atom MD [24]. The entire systems were composed of $\sim 4,400$ particles.

The analysis of the trajectories of protomers **A** and **B** were performed over the last 0.95 μs, that is after the equilibration phase (Fig. (B.1)).

The leap–frog stochastic dynamics algorithm was used to integrate the equations of motion with a time step $\Delta t = 2$ fs and a friction coefficient $\gamma_i = m_i/\tau$, where $\tau = 0.5$ ps is the time constant for the coupling and m_i is the mass of ith particle. A cut–off distance of 14 Å was used for the electrostatics [1]. A cut–off of 14 Å was also used for the van der Waals interactions. The pair list was updated every 10 steps. The SHAKE algorithm [40] was used to keep fixed the distance of bonds containing hydrogens.

The systems were first relaxed by 1 ns MM/CG with positional restraints on OmpT/ARRA complexes to minimize the energy of the solvent. Then further 1 ns with positional restraints on OmpT, were performed for OmpT/ARRA complex to allow the ligand to accommodate itself inside the binding pocket with the MM/CG potentials.

[1] This very crude assumption in the treatment of the electrostatics appears to be justified by the simplicity of the model used. Careful checks were made on energy conservation. In addition, test calculations with a longer cut–off for electrostatics on a test system showed very similar results to those with the cut–off.

Complex id	Simulated Time (μs)
A: protonated D83, neutral N– and C–termini (AceARRANme)	1.00
B: deprotonated D83, neutral N– and C–termini (AceARRANme)	1.00
C: protonated D83, charged N– and C–termini (NH_3^+ARRACOO^-)	0.05
D: deprotonated D83, charged N– and C–termini (NH_3^+ARRACOO^-)	0.05
S99A–**A**	0.16
S99A–**B**	0.16
H212A–**A**	0.15
H212A–**B**	0.15

Table 3.2: Protomers and mutants of OmpT/ARRA complexes investigated in this work.

Finally, we performed an atomic–force field based MD simulation of a reference system. This is diglycine (AceGGNme) in a periodic box of 779 water molecules. The GROMOS96 43a1 [55] and SPC [94] force fields for the dipeptide and water were used, respectively. Room conditions (T=300 K, P=1 bar) were achieved by coupling the system with Berendsen thermostat [95] with τ=1.0 ps and a Berendsen barostat [95] with compressibility of 4.5 10^{-10}bar^{-1} in all three dimensions. The time step of the integration was 2 fs. Electrostatic interactions were calculated using a cut–off of 18 Å, with van der Waals interactions truncated at 14 Å. 0.015 μs of trajectory were collected.

The following properties were calculated: (i) Large scale motions were investigated by calculating the eigenvectors of the covariance matrix $\mathcal{C}$ [56], defined in Section 1.3.1. (ii) Electric field in the MM region, using the GROMOS96 43a1 force field [55].

Results

Systems investigated

We focus on the four protomers of the wt OmpT in complex with its substrate ARRA obtained after 24 ns of all–atom MD simulations [35]. The protomers differ

in the protonation of D83 at the active site and in the N– and C– termini of the peptide (see Table 3.2). The catalytic histidine residue, H212 was assumed to be protonated in δ as in [91, 35] (Fig. (3.10)ab). S99A and H212A mutants of the two protomers **A** and **B** were built by simply replacing these residues with A residue.

The protomers underwent MM/CG simulations for the timescales summarized in Table 3.2.

MM/CG of protomers A and B

The structure of protomers **A** and **B** (Fig. (3.10)ab) is well maintained for the time–scale investigated (1 μs): the C_α root mean square deviation (RMSD) shows that **A** and **B** reached their final configurations after $\sim$0.05μs and then fluctuated around them for the remainder of the simulations (see Fig. (B.1)ab). These appear to be productive Michaelis complexes: a water molecule bridges H212 and and D83 thus pointing towards the substrate carbonyl carbon. This water molecule is located in the right position to perform a nucleophilic attack on the carbonyl group of the substrate. At times a second water molecule bridges H212 and D83. This happens more often in **B** because in this protomer a longitudinal fluctuation of the β–barrel causes an opening of the cleft (δ distance), as we discuss later. Periodically, the H212–water(s)–D83 interaction is replaced by a direct H–bond between D83 and S99.

The enzyme–substrate interactions along with interactions within the active site are maintained during the dynamics of both protomers (Table 3.3, Fig. (3.10)ab): (i) Hδ@H212 H–bonds to Oδ@D210 or to O@D210; (ii) at time Nϵ@H212 H–bonds with H@R3 of the substrate. The lower activity of H212A [83] underlines the key role of this interaction; (iii) R2 of the substrate forms salt bridges with E27 and D208, R3 with D97 and D85; (iv) Hϵ@H101 H–bonds either to S99 or to D83; (v) at times, S99 H–bonds to D85 and D97; (vi) S99 backbone H–bonds to D83 backbone; (vi) D85 backbone H–bonds to D97 backbone.

Next, we investigate the polarization of the reactants (substrate carbonyl

	A	**B**	H212A–**A**	H212A–**B**	S99A–**A**	S99A–**B**
Hδ@H212–Oδ@D210	2.6(±0.8)	2.5(±0.7)	//	//	7.0(±3.5)	2.8(±0.7)
Hδ@H212–O@D210	5.2(±0.8)	4.3(±1.2)	//	//	7.0(±2.1)	5.0(±0.8)
Nϵ@H212–H@R2	4.7(±0.7)	5.0(±1.6)	//	//	6.4(±2.2)	5.0(±0.7)
Nϵ@H212–C@R2	4.4(±0.5)	6.2(±1.4)	//	//	6.5(±1.9)	4.4(±0.7)
Nϵ@H212–Oδ@D83	5.8(±1.0)	8.0(±1.5)	//	//	11.3(±1.8)	7.9(±1.4)
Cζ@R2–Cδ@E27	4.8(±0.5)	5.0(±0.8)	6.7(±2.4)	5.0(±0.7)	6.2(±1.7)	6.3(±1.4)
Cζ@R2–Cγ@D208	5.0(±0.7)	5.2(±0.9)	9.0(±2.3)	8.4(±0.3)	7.6(±2.1)	6.6(±0.7)
Cζ@R2–Cγ@D210	8.5(±0.7)	8.0(±0.9)	10.1(±3.0)	4.0(±0.4)	9.5(±1.6)	6.8(±1.2)
Cζ@R3–Cγ@D85	6.0(±1.0)	6.5(±1.3)	9.7(±1.6)	7.0(±1.7)	9.4(±1.4)	6.7(±1.2)
Cζ@R3–Cγ@D97	7.0(±1.5)	8.0(±1.8)	11.6(±2.1)	9.0(±1.6)	11.8(±3.4)	15.0(±3.1)
Hϵ@H101–Oγ@S99	3.1(±0.9)	4.2(±1.2)	4.8(±1.0)	3.0(±0.7)	//	//
Oδ@D97–Hγ@S99	5.5(±1.3)	6.6(±2.1)	3.5(±2.3)	8.2(±2.6)	//	//
Hδ@D83–O@R2	4.5(±1.1)	//	5.5(±1.8)	//	7.5(±1.4)	//
Hδ@D83–Oδ@D97	8.8(±1.2)	//	7.9(±1.3)	//	4.0(±1.2)	//
Hϵ@H101–Oδ@D83	3.7(±1.0)	3.5(±1.1)	4.6(±2.0)	3.1(±0.9)	5.0(±1.1)	3.5(±0.9)
H@D83–O@S(A)99	2.0(±0.2)	2.0(±0.3)	2.0(±0.2)	2.0(±0.2)	2.2(±0.4)	4.0(±0.6)
O@D83–H@S(A)99	2.0 (±0.3)	2.4(±0.5)	2.5(±0.6)	2.0(±0.2)	2.6(±0.7)	2.5(±0.9)
H@D85–O@D97	2.1(±0.4)	2.3(±0.3)	2.3(±0.7)	2.4(±0.7)	3.0(±1.2)	5.0(±1.8)
O@D85–H@D97	2.0(±0.4)	2.6(±1.0)	2.1(±0.4)	2.7(±0.7)	3.5(±1.6)	5.3(±1.7)

Table 3.3: MM/CG simulations of **A** and **B**, H212A–**A**, H212A–**B**, S99A–**A** and S99A–**B**: selected MD–averaged distances (Å) within the active site of the protein. Standard deviations (STD) are reported in parenthesis.

group C@R2 of the substrate and catalytic water, Fig. (3.11)ab), described here in terms of active site electric field. We will consider the direction of the field along the C=O bond of the substrate (Fig. (3.11)a) and the water C_2 axis (Fig (3.11)b). As for the substrate, the field of **A** does not show any preferential direction with respect to the C=O bond: the angle α between the field and the vector identified by the C=O bond (Fig. (3.11)a and Table in Fig. (3.11)) is spread within 0° and 200°, with a very high standard deviation value around its average value ($\langle\alpha\rangle = 125°(44°)$). A large spread is also found in a reference system here investigated, the glycine dipeptide in water solution, for which $\langle\alpha\rangle = 100°(59°)$. Instead, in **B** the field spreads to a much lower extent ($\langle\alpha\rangle = 140°$ (10°)) and it is partially aligned with the C=O bond (Fig. (3.11)a and Table in Fig. (3.11)). We conclude that the carbonyl carbon is expected to be more electrophilic than in water solution and in **A**. As for the catalytic water, the field in **A** is much less spread than that of the reference system, which practically does not show any preferential direction ($\langle\beta\rangle \sim 104°(14°)$ and 93°(60°), respectively, Fig. (3.11)b and Table in Fig. (3.11)). However, its direction is almost orthogonal relative to the

	α	STD α	β	STD β
Diglycine	100°	59°	93°	60°
Protomer A	125°	44°	104°	14°
Protomer B	140°	10°	β_1=92° β_2=15°	STDβ_1=12° STDβ_2=8°

Figure 3.11: In the top of the picture: electrostatic field acting on carbonyl carbon (α angle): (a), and on the dipole axes of water, C_2 (β angle): (b). The table shows the average values and the correspondent STD of α and β angle computed for dyglicine, protomer **A** and protomer **B**.

C_2 axis (Fig. (3.11)b and Table in Fig. (3.11)): therefore, we expect that the field will not affect significantly the nucleophilic power of the water. In **B**, the angle is also not spread and it shows a bimodal distribution, either orthogonal to C_2 axis ($\langle\beta\rangle \sim 90°(12°)$, Table in Fig. (3.11)), or almost aligned to it ($\langle\beta\rangle \sim 15°(8°)$, Table in Fig. (3.11)). Thus, the field in **B** might render water more nucleophilic than in water. In conclusion, the field in **B** (in contrast to **A** and to the reference system) significantly polarizes the C=O bond and the catalytic water molecule. Such polarization effect could be used by the enzyme in order to enhance the electrophilicity of the carbonyl carbon and the nucleophilicity of the water oxygen.

Now we turn our attention to the structural fluctuations of the substrate in the binding cavity, which have been shown to play a functional role for other proteases such as HIV-1 protease [30] and BACE [31]. For these enzymes it was found that the distance between the catalytic dyad and the substrate fluctuates around characteristic values corresponding to different mutual positions of the catalytic water relative to the substrate carbonyl carbon: only conformations in which the distance between the enzyme and the substrate is at a minimum turned out to

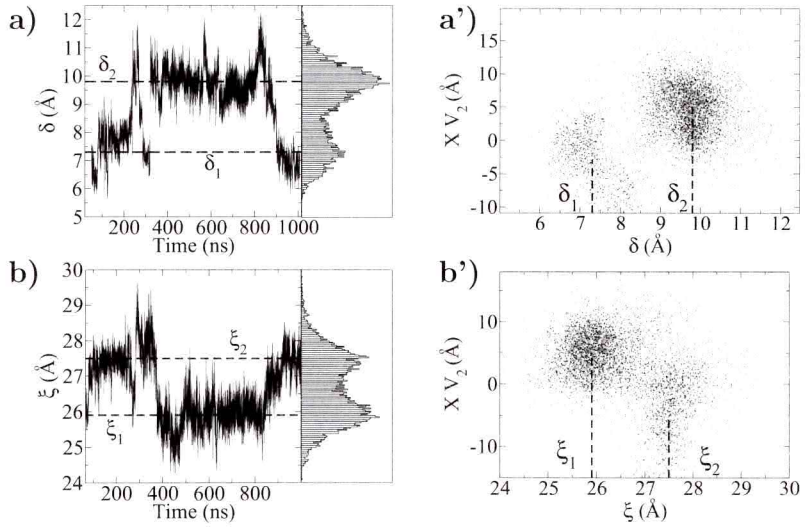


Figure 3.12: Left: time evolution of δ: (a), and ξ: (b) parameters (defined in the text), respectively. Right: projection of $\mathbf{V}_2$ (defined in the text, and representing one of the largest scale motion of the protein) versus δ distance: (a'), and versus ξ distance: (b').

be catalytically efficient [30,31]. Because those motions are correlated to the the large–scale motions ("essential eigenvectors" [56]) of the proteins, the enzyme might play a role for the reaction by steering the substrate into its appropriate reactive conformation.

In order to ascertain, whether this is the case also for this protease, we monitor here the distance of the center of mass of ARRA peptide from the D83–D85 dyad (δ) and from the center of mass of the β–barrel (ξ), which geometrically represents the axis of the protein β–barrel (see Fig. (3.4)a).

δ is affected by the width of the cleft being proportional to the distance between the putative catalytic dyad D83–H212 and therefore its fluctuations might modulate the position of the water inside the catalytic cleft. ξ is affected by the distance between the substrate and the catalytic water. Thus, δ and ξ can be used as suitable descriptors of the enzyme "active" conformations [96].

In **A**, $\langle\delta\rangle \sim 7.8(0.5)$ Å and $\langle\xi\rangle \sim 24.1(0.4)$ Å, showing a sharp Gaussian–like distribution (see Fig. (B.5)ab). In **B**, both quantities feature bimodal distribu-

tions: δ fluctuates from $\delta_1 = 7.3$ Å to $\delta_2 = 9.8$ Å (Fig. (3.12)a) and ξ from $\xi_1 = 25.9$ Å to ξ_2 27.5 Å (Fig. (3.12)b). Interestingly, the transitions from ξ_1 to ξ_2 and from δ_1 to δ_2 occur at the same time, suggesting that the oscillation of the cleft is correlated with the oscillation of the substrate along the axis of the β–barrel (Fig. (3.12)ab). An analysis of the essential eigenvectors [56] confirms that this is the case. In order to ascertain the presence of relevant correlations, we considered the projection of $\mathbf{V}_2$ of the covariance matrix $\mathcal{C}$, which is one of the largest eigenvectors, onto the trajectory $\mathbf{X}$ ($\mathbf{X} \cdot \mathbf{V}_2$). This quantity gives information about the motion of the protein along the eigenvector direction showing that $\mathbf{V}_2$ mostly affect the solvent–exposed loops embracing the active site. The plot of the projection *versus* the ξ and δ shows a clear correlation between the two sets. From the pictures in Fig. (3.12)a'b', indeed, two mostly populated regions corresponding to δ_1 and δ_2 and to ξ_1 and ξ_2 are visible. Thus, $\mathbf{V}_2$ induces relevant variations in the relative distances of the active site. This finding strongly suggests that the strategy employed for the catalytic reaction in **B** might be very similar to that found in several proteases [96], in particular in HIV-1 PR [30] and BACE [31]. In contrast, no relevant correlations are observed in **A** (data not shown).

MM/CG of protomers C and D

The Michaelis complex in **C** and **D** disrupt already after 10 ns: The NH_3^+ terminal groups, which formed salt bridges with D210 and E27 at the beginning of the dynamics, rotate and form a salt bridge with D83 and an H–bond with Nϵ@H212 (Fig. (B.2)ab). Consequently, the carbonyl carbon, which undergoes the nucleophilic attack, moves away from the putative catalytic dyad composed by H212 and D83 (see Fig. (B.2)ab), whilst the oxygen of such a group flips, pointing toward the D210 and H212 dyad. Such non–productive Michaelis complex, in which there is no putative nucleophilic agent in the close proximity of the carbonyl carbon, is maintained for further $\sim 0.04 \mu$s, after which we decided to stop the simulation. We conclude that the presence of the charged termini in

addition to the two positive arginines affect the structure of the Michaelis complexes in both **C** and **D**.

MM/CG of H212A and S99A

H212A and S99A mutations show a residual activity ranging within 0% and 4% [82].

Here we carry out MM/CG simulations to investigate the structure of their Michaelis complexes and further compare them to the wt. Of course, we consider **A** and **B** protomers. The time–scale investigated is shorter (0.15 μs) as we are solely interested in constructing structural models.

In H212A, the mutation disrupts the H–bonds with Nϵ@H212 and the amide group of the substrate, which contributes to maintain fixed the position of the substrate inside the enzymatic cleft. In the wt, R2 and R3 of the substrate form salt bridges with E27 and D208 and with D85 and D97, respectively. These two residues rotate in both **A** and **B** and their side chains face the solvent (Table 3.3 and Fig. (B.6)ab). At the end of the simulation, the carbonyl carbon, which is cleaved by the wt enzyme, moves away from the the putative catalytic dyad H212–D83 and the cleft is filled by water.

In S99A, the replacement of S99 with alanine disrupts the S99–Hϵ@His101 H–bond (Fig. (B.7)). The H atom of the same histidine H–bonds to D83 in **A** and **B**.

In **A** the breaking of S99–H101 H–bond causes a rearrangement of H101 and consequently D83, which is protonated, rearranges and H–bonds to D97 (Table 3.3 and Fig. (B.7)a). Consequently, the interaction between the proton of D83 and the substrate is lost, and the latter fluctuates allowing a rotation of the side chain of R2 of the substrate. As a result, R2 H–bonds with D210 and the catalytic residue H212 is permitted to move further apart from the active site (Table 3.3). This causes a drastic change of the substrate configuration and its partial detachment (Fig. (B.7)a).

In **B**, D83, which is ionized, forms a stable H–bond with Nϵ@H101. However,

because of the lack of S99–D97 H-bonding, D97 moves away in turn causing the loss of H–bond interactions between the backbone of D85 and D97 (Fig. (B.7)b). As a result, the salt bridge between the D85, D97 and R3 is lost, allowing the side chain of the latter residue to rotate and get solvated. In this case, there is also found a partial detachment of the substrate (Table 3.3, Fig. (B.7)b).

Discussion and Concluding Remark

We have presented an MM/CG study of wt and mutant OmpT in complex with its substrate ARRA. We focused on four protomer (**A**–**D**, which differ for the protonation state of D83 and the charge of the substrate: $_{Ace}$ARRA$_{Nme}$ or $_{NH3^+}$ARRA$_{COO^-}$).

During the simulation of wt, the complexes with substrates with charged tails evolved to a non–productive Michaelis complex, because the charged tail groups interact with the residues among the catalytic cleft causing a distortion of the ARRA peptide and its detachment from the binding pocket. Obviously, our investigation does not rule out the presence of longer substrates with charged tail groups, as those substrates may have their termini located outside the catalytic cleft.

In contrast, complexes with the substrates with neutral tails provided a productive Michaelis complexes and both remained stable over 1–μs dynamic simulations. The two complexes (**A** and **B** in Fig. (3.10)ab) exhibit different protonation states of D83: in **A** it is protonated while in **B** it is ionized. The protomers are characterized by significant differences in the electrostatic polarization of the reactants. In **B**, the active site polarizes both the catalytic water and the carbonyl carbon, rendering the first more nucleophilic and the second more electrophilic relative to a reference system in water (diglycine dipeptide). In contrast, in **A** no significant polarization is observed and the electric field acting on the carbonyl carbon bond and the catalytic water is very similar to that calculated for the reference system diglycine in water solution.

Protomer **B** which shows a well defined electrostatic field direction, is also characterized by large conformational fluctuations of the substrate triggered by

global large–scale motions (Fig. (3.12)a'b') and populates significantly different conformations (see Fig. (3.12)ab and Fig. (B.3)b) in contrast to **A** in which the substrate fluctuates around a well defined conformation (see Fig. (B.3)a and Fig. (B.5)ab). These findings are highly suggestive of a functional role of large–scale fluctuations of complex **B**. Such conclusions about the electric field acting on the reactants and about the motion of the substrate inside the catalytic cleft can not to be drawn if one is limited to the typical time of MD (0.01–0.1 μs).

We have next used the MM/CG approach on H212A and S99A mutants (for both protomers **A** and **B**) to provide the structural basis for the much lower activity of these mutants relative to wt. Whilst this is rather straightforward in H212A, as it is part of the catalytic dyad, this result is intriguing for S99A. Our simulations suggest that, in H212A, the ARRA peptide detached spontaneously in both protomers due to the lost of the H–bond interaction between Nϵ@H212 and the amide group of the substrate, present in the wt. S99A, indeed, allowed a similar detachment of the peptide due to the disruption of the geometry of the active site in both protomers, in spite of the fact that this residue is not located at the active site. In **A** the loss of the H–bond interaction between S99 and D83 allows the rotation of the latter residue causing the breaking of interaction between D83 and the substrate. In contrast, in **B** the detachment of the peptide is due to the loss of the salt bridge between R3 and D85–D97 after the braking of S99–D97 interaction. We conclude that not only first–shell H–bond interactions (such those formed by H212), but also second shell H–bonding (such as that of S99) play an important role for the stability of the geometry of the active site. Removing any of those may cause a high instability of the active site and therefore a reduced activity.

In conclusion, our MM/CG approach emerges as a powerful tool to investigate μs (potentially function–related) dynamics of enzymes, which may impact on function as well as an efficient and fast tool for computational structural biology. MM/CG, by allowing running more numerous and longer simulations, is expected, on the one hand to improve confidence of the results, and on the other one it may strengthen the interaction between molecular biology experiments and

simulations.

Summary, Conclusions and Perspectives

This thesis is devoted to the construction, the validation and the application of a novel hybrid Molecular mechanics/ Coarse Grained model we discussed so far: the MM/CG model [24].

The idea of implementing the MM/CG approach arises from the necessity of overcoming the drawbacks coming from all–atoms molecular dynamic simulations and from the coarse grained approaches. MD is a very powerful tool to predict structural, dynamical and thermodynamical properties of biological molecules [2], but the current computational power constrains this analysis to time scales of $\sim$100 ns, too short to follow several important biological processes. CG models are useful to understand large scale phenomena, but they can not describe the exquisite molecular recognition events between enzymes and their substrates.

Our MM/CG approach overcomes some of these limitations by taking the advantages of both MD and CG methods: the amino acid residues involved in the ligand binding are treated with atomic details (MM region) with a molecular mechanics force field, whereas the rest of the protein is treated at the CG level. In this way the necessary details associated with the biological activity and the "large" time scale over which it occurs are preserved.

We tested our method on two proteins of great pharmacological relevance, belonging to the aspartic protease class: the HIV type 1 virus aspartic protease, (HIV-1 PR) [28, 30] and the human β-secretase, (BACE) [26, 31]. The first is a major target for anti–AIDS therapy [29]; the second plays a role in the progression

of Alzheimer's disease [27]. In spite of the identical cleavage site (a dyad composed of two aspartic residues), these proteins exhibit a large structural diversity: HIV-1 PR is a homodimer containing mostly β-strands, while BACE is a monomer with both α and β secondary structure elements. Our MM/CG simulations were able to reproduce both the mesoscopic (*i.e.* the residue root mean square fluctuations (RMSF) and the principal normal modes) and the local microscopic details (*i.e.* distances between key atoms in the active sites and H-bond patterns) of the two proteins. Comparison is also made with a pure coarse–grained model, the β–Gaussian model [97, 24, 62].

We next tested our model on a membrane protein, the outer membrane protease T (OmpT) [32]. Comparing MD [81] and MM/CG [34] simulation data of OmpT in the free state we ascertain that our model is suitable not only for globular proteins (as HIV-1 PR and BACE) but also for membrane proteins which are immersed in a discontinuous medium. In addition, we analyze the large scale motions that affect the enzyme, finding that probably those fluctuations might play a role for the catalysis hence sharing some resemblance with those proposed for HIV-1 PR [30] and BACE [31]. In order to elucidate the importance of the large scale fluctuations of OmpT and to give a complete picture of the dynamical evolution of this enzyme, we performed the MM/CG simulations on the Michaelis complexes of OmpT [36] (OmpT/ARRA). The analysis of conformational fluctuations of the substrate in the active site of OmpT in μs time–scale shows that large–scale motions of the protein and the electrostatic field may impact on the function of the enzyme [36], as in several aspartic proteases [96]. We find that such conclusions can not to be drawn if one covers to the typical time of MD (0.01–0.1 μs).

Our long MM/CG simulations have also permitted to study the effect of mutations on the OmpT complex [36]. Two OmpT mutants, which are experimentally known to be much less efficient than the wild–type (wt), are investigated: H212A, which involves a residue at the active site, and S99A, which instead is not located in the cleavage site. The simulations (over 0.1 μs) show that both mutations cause the partial detachment of the substrate from the active site, consistently with the

reduced activity of the mutants. Because of its extremely cheap computational cost (two order of magnitude faster than standard all–atom MD), the methodology emerges as a powerful tool to investigate structure/ function relationships of high–throughput site–directed mutagenesis data.

In conclusion, our MM/CG approach emerges as a powerful tool to investigate μs (potentially function–related) dynamics of enzymes, which may impact on function, as well as an efficient and fast tool for computational structural biology. MM/CG, by allowing running more numerous and longer simulations, is expected, on the one hand to improve confidence of the results, and on the other one to strengthen the interaction between molecular biology experiments and simulations.

As future perspectives of the work, we plan to introduce long–range electrostatic in the CG region, following ref. [98]. This will allow to a realistic description of protein polar effects. In addition, we plan to implement a novel efficient schemes aimed to prevention of solvent diffusion out of the classical MM region.

Appendix A

The βGaussian model: βGM

Here, I present the results of my first computational study on the large scale fluctuations acting on proteins investigated with a CG approach. The theory of the βGaussian model [11] is described. Then, we carried out the characterization of the the low energy conformational distortions that the biopolymer can sustain (large scale structural changes) for BACE, also investigated by MM/CG model in Chapter 2, which give considerable insight into the functional activity of enzymes.

This Appendix is based on the work in ref. [21].

A.1 theory

Since the main objective is the modeling of the large scale fluctuations in a protein, it is convenient to reduce the spatial degrees of freedom of the biopolymer through a CG approach where a two–particle representation is used for each amino acid. Besides the C_α atom, an effective C_β centroid is employed to capture, in the simplest possible way, the sidechain orientation in a given amino acid (except for Gly for which only the C_α atom is retained). This reduced structural representation affects the form of the effective Hamiltonian associated with the coarse–grained structural representation. Arguably, the simplest energy function for the system can be constructed by assuming that all centroids in the protein (C_α and/or C_β) whose separation is smaller than a given interaction distance, R,

interact through the same pairwise potential, V [99]. The information of which centroids are in interaction in the native state is aptly summarized in the native contact matrix Δ_{ij}^{XY} which takes on the values of 1 (0) if the native separation of the particles of type X and Y, belonging respectively to residues i and j, is below (above) R. The system energy function evaluated on a trial structure, Γ, can then be written as:

$$\mathcal{H}(\Gamma) = \mathcal{H}_{BB}(\Gamma) + \mathcal{H}_{\alpha\alpha}(\Gamma) + \mathcal{H}_{\alpha\beta}(\Gamma) + \mathcal{H}_{\beta\beta}(\Gamma) \,, \tag{A.1}$$

where:

$$\begin{aligned} \mathcal{H}_{BB}(\Gamma) &= K \sum_i V(d_{i,i+1}^{C_\alpha - C_\alpha}) \,, \\ \mathcal{H}_{\alpha\alpha}(\Gamma) &= \sum_{i<j} \Delta_{ij}^{C_\alpha - C_\alpha} V(d_{i,j}^{C_\alpha - C_\alpha}) \,, \\ \mathcal{H}_{\alpha\beta}(\Gamma) &= \sum_{i<j} \Delta_{ij}^{C_\alpha - C_\beta} V(d_{i,j}^{C_\alpha - C_\beta}) \,, \\ \mathcal{H}_{\beta\beta}(\Gamma) &= \sum_{i<j} \Delta_{ij}^{C_\beta - C_\beta} V(d_{i,j}^{C_\beta - C_\beta}) \,. \end{aligned} \tag{A.2}$$

In Eq. (A.2), d_{ij}^{XY} indicates the actual separation of the particles in the trial structure, Γ . The indices i and j run over all integer values ranging from 1 up to the protein length, N. To account for the protein chain connectivity, in Eq. (A.2) it is introduced a backbone energy term which mimics the higher strength of the bond for consecutive amino acids with respect to non–covalent contact interactions. Consistently with previous studies [62], R is taken equal to 7.5 $\mathring{\text{A}}$ and K = 1.

The general pairwise Hamiltonian in Eq. (A.1) is subject to an important requirement since it must guarantee that the assigned reference conformation is at the global energy minimum. The customary way to accomplish this [9,100,101] is to assume that the potentials V^{X-Y} appearing in Eq. (A.2) attain their global minimum in correspondence of the native separation of the centroids. For small fluctuations around the native structure, the interaction energy of two centroids,

i and j , can then be expanded in terms of the deviations from the native distance vector, $\vec{r}_{ij}$. If one indicates the deviation vector as $\vec{x}_{ij}$, so that the total distance vector is $\vec{d}_{ij} = \vec{r}_{ij} + \vec{x}_{ij}$, the pairwise interaction can be written as:

$$V(d_{ij}) \approx V(r_{ij}) + \frac{k}{2} \sum_{\mu\nu} \frac{r_{ij}^{\mu} r_{ij}^{\nu}}{r_{ij}^{2}} x_{ij}^{\mu} x_{ij}^{\nu} \ , \tag{A.3}$$

where μ and ν denote the Cartesian components, x, y and z, and k is the second derivative of V at its minimum. The expansion of Eq. (A.3) brings about a dramatic simplification of Hamiltonian in Eq. (A.1) which, in fact, acquires a quadratic dependence in terms of the deviation vectors, $\vec{x}$. However, a further simplification of the energy function can be achieved by exploiting the fact that the coordinates of the C_α centroids encode, in an almost unique way, the positions of all other atoms in the protein, and hence also of the C_β atoms. The construction scheme adopted in the βGM defines the location of the ith C_β through a co–planar version of the Park and Levitt construction rule [102]:

$$\vec{r}_{C_\beta}(i) = \vec{r}_{C_\alpha}(i) + l \frac{2\vec{r}_{C_\alpha}(i) - \vec{r}_{C_\alpha}(i+1) - \vec{r}_{C_\alpha}(i-1)}{|2\vec{r}_{C_\alpha}(i) - \vec{r}_{C_\alpha}(i+1) - \vec{r}_{C_\alpha}(i-1)|} \ , \tag{A.4}$$

where $l = 3$ Å. Thus, the degrees of freedom can be reduced to only those of the C_αs. The linear relationship between the C_α and C_β coordinates recast the Hamiltonian as follows:

$$\mathcal{H} = \frac{1}{2} k \sum_{ij,\mu\nu} x_{i,\mu} \mathcal{M}^{-1}_{ij,\mu\nu} x_{j,\nu} \ , \tag{A.5}$$

where $x_{i,\mu}$ is the deviation of ith C_α along the μ axis and $\mathcal{M}^{-1}$ is a $3N \times 3N$ symmetric matrix. The elastic response of the system is uniquely dictated by the eigenvalues and eigenvectors of $\mathcal{M}^{-1}$. In the next subsection it is shown a complete derivation of Hamiltonian in Eq. (A.5).

In thermal equilibrium, each amino acid moves under the action of the quadratic Hamiltonian in Eq. (A.5) subject to a viscous friction originating from interactions with the surrounding solvent as well as with the rest of the protein [103].

The viscous hindrance of the motion is so important that the protein dynamics becomes severely overdamped [104, 103]. In this case it seems appropriate to describe the dynamics of the amino acids within a Langevin framework [105, 53]:

$$\gamma_i \dot{x}_{i,\mu}(t) = -k \sum_{j,\nu} \mathcal{M}^{-1}_{ij,\mu\nu} x_{j,\nu}(t) + \eta_{i,\mu}(t) \ , \tag{A.6}$$

where γ_i is the viscous friction coefficient and $\eta_{i,\mu}(t)$ is a stochastic noise satisfying the relations [105, 53]:

$$\langle \eta_{i,\mu}(t) \rangle = 0 \ , \qquad \langle \eta_{i,\mu}(t) \eta_{j,\nu}(t') \rangle = \delta_{ij} \delta_{\mu\nu} \delta(t - t') 2 K_B T \ \gamma_i \ . \tag{A.7}$$

Within this dynamical framework it is possible to calculate exactly how correlations among the displacements of various pairs of residues decay as a function of time. We start by considering the case where the various viscous coefficients in Eq. (A.6) take on the same value, γ. In this case one has [106]:

$$\langle x_{i,\mu}(t) x_{j,\nu}(t + \Delta t) \rangle = \frac{K_B T}{k} \sum_l v_i^l v_j^l \lambda_l \ e^{-\frac{1}{\lambda_l} \frac{k}{\gamma} \Delta t} \ , \tag{A.8}$$

where the $\langle \cdot \rangle$ in Eq. (A.8) denotes the usual canonical thermodynamics average. The vector $\vec{v}^{\,l}$ and the scalar λ_l are, respectively, the lth eigenvector and the lth eigenvalue of the matrix $\mathcal{M}$ ordered in such a way that $\lambda_l > \lambda_{l+1}$. It should be noted that since the Hamiltonian is invariant under rotations and translations, it will always possess (at least) six eigenvectors that are obviously excluded from the sum in Eq. (A.8) [62, 107]. The eigenvectors $\{\vec{v}^{\ l}\}$ represent therefore the independent modes of structural relaxation in the protein, while the associated decay times are given by:

$$\tau_l = \lambda_l \frac{\gamma}{k} \ . \tag{A.9}$$

Of particular interest, for the purpose of identifying the concerted large scale structural fluctuations occurring in a protein, is the degree of correlation of pairs of residues at equal times. This information is summarized in the covariance

matrix, $\mathcal{C}$, whose elements are obtained by $\Delta t = 0$ in Eq. (A.8):

$$\mathcal{C}_{ij,\mu\nu} \equiv \langle x_{i,\mu} x_{j,\nu} \rangle = \frac{K_B T}{k} \sum_l \lambda_l \, v_i^l v_j^l = \frac{K_B T}{k} \mathcal{M}_{ij,\mu\nu} \, . \tag{A.10}$$

Due to the fact that this matrix contains the full three-dimensional information about pair correlations its linear size is $3N$. Typically it is important only to quantify the relative degree of correlation of any two residues; in this case one can consider the normalized reduced covariance matrix (of linear size N) which is defined as:

$$\mathcal{C}_{ij} = \frac{\langle \vec{x}_i \cdot \vec{x}_j \rangle}{\sqrt{\langle |\vec{x}_i|^2 \rangle \langle |\vec{x}_j|^2 \rangle}} = \frac{\mathcal{C}_{ij,\nu\nu}}{\sqrt{\mathcal{C}_{ij,\nu\nu} \mathcal{C}_{jj,\mu\mu}}} \, , \tag{A.11}$$

where summation over repeated indices is implied. It is important to stress that the elements of the covariance matrix are thermodynamic averages reflecting equilibrium properties of the system and, hence, are independent of the friction coefficients. Through Eq. (A.10) one sees that the same set of vectors $\{\vec{v}^{\,l}\}$ as describes the modes of relaxation also describes the independent modes of structural fluctuations in thermal equilibrium. The contribution of each mode to the pair correlation is proportional to λ_l, thus establishing the intuitive result that the modes associated with the longest relaxation times are those responsible for the largest structural fluctuations. The equivalence of the vectors describing the modes of relaxation and structural fluctuations is a consequence of having used the same viscous coefficient for all amino acids in the protein. When this simplifying assumption is not made one has that the modes and times of relaxation are determined by diagonalizing instead the symmetric matrix:

$$\widetilde{\mathcal{M}}_{ij,\mu\nu} = \sqrt{\gamma_i \gamma_j} \mathcal{M}_{ij,\mu\nu} \, . \tag{A.12}$$

On the other hand, the modes of structural fluctuation are the same as before, as can be ascertained by calculating directly the thermodynamic average $\langle x_{i,\mu} x_{j,\nu} \rangle$ with the canonical weight associated with Hamiltonian in Eq. (A.5).

A.1.1 βGM Hamiltonian

It is supposed that the interaction between the C_α centroids is a two body interaction. Thus, we define $\vec{r}_{ij} = \vec{r}_i - \vec{r}_j$, where i and j are a couple of aminoacids and $\vec{r}_i$ stands for the native position of ith centroid. We also define $\vec{x}_i$ and $\vec{x}_j$ that are the deviation from the native position of the ith and jth aminoacids, respectively. For little deviation from the native structure, the energy potential $V(d)$, with $d = |\vec{r}_{ij} + \vec{x}_{ij}|$, can be approximated at second order of Taylor series expansion:

$$V(d) = V(|\vec{r}_{ij}|) + \frac{1}{2}\sum_{\mu\nu} \frac{\partial^2 V(d)}{\partial d^2}\frac{\partial d}{\partial x^\mu}\frac{\partial d}{\partial x^\nu}|_{\vec{x}_{ij}=0} x^\mu_{ij} x^\nu_{ij} \ , \tag{A.13}$$

where μ and ν stand for the Cartesian coordinates.

Choosing the energy scale in such way that $\frac{\partial^2 V(d)}{\partial d^2} = 0$ if $\vec{x} = 0$, the Hamiltonian of the system around the native state reads:

$$\beta\mathcal{H} = V_{chain} + \frac{\beta}{2}\sum_{ij,i\neq j} \Delta_{ij} \sum_{\mu\nu} \frac{r^\mu_{ij} r^\nu_{ij}}{r^2_{ij}} x^\mu_{ij} x^\nu_{ij} \ , \tag{A.14}$$

where $\beta = (K_B\ T)^{-1}$ and Δ_{ij} is the native contact matrix, defined as:

$$\Delta_{ij} = \begin{cases} 1 & \text{if } d_{ij} < R \\ 0 & \text{if } d_{ij} > R \end{cases} \ , \tag{A.15}$$

where R is setted to be 7.5 $\overset{\circ}{\text{A}}$.

The first term in Eq. (A.14), V_{chain}, is an isotropic term. It can be written as:

$$V_{chain} = K\sum_{i=1}^{N-1} (\vec{x}_{i,i+1})^2 \ , \tag{A.16}$$

where K is the strength between consecutive aminoacids. The second term in Eq. (A.14) is a two body potential with an anisotropic chain term.

The Hamiltonian $\mathcal{H}$ in Eq. (A.14) can be rewritten in the following form:

$$\begin{aligned}\beta\mathcal{H} &= \sum_{\mu\nu,ij,i\neq j}\left[\frac{K}{2}(\delta_{i,j-1}+\delta_{i,j+1})\delta_{\mu,\nu}x_{ij}^{\mu}x_{ij}^{\nu}+\beta\Delta_{ij}\frac{r_{ij}^{\mu}r_{ij}^{\nu}}{r_{ij}^{2}}x_{ij}^{\mu}x_{ij}^{\nu}\right]\\ &= \sum_{\mu\nu,ij}\left[\frac{K}{2}(\delta_{i,j-1}+\delta_{i,j+1})\delta_{\mu,\nu}+\beta(1-\delta_{ij})\Delta_{ij}\frac{r_{ij}^{\mu}r_{ij}^{\nu}}{r_{ij}^{2}}\right]x_{ij}^{\mu}x_{ij}^{\nu}\\ &\equiv \sum_{\mu,\nu,i,j}\frac{A_{ij,\mu\nu}}{2}x_{ij}^{\mu}x_{ij}^{\nu}\,, \end{aligned} \tag{A.17}$$

where $\delta_{i,j}$ is a function defined as:

$$\delta_{i,j}=\begin{cases}1 & \text{if } i=j\\ 0 & \text{if } i\neq j\,.\end{cases} \tag{A.18}$$

Being $\vec{x}_{ij}=\vec{x}_i-\vec{x}_j$, the Hamiltonian in Eq. (A.17) becomes:

$$\begin{aligned}\beta\mathcal{H} &= \frac{1}{2}\sum_{ij,\mu\nu}A_{ij,\mu\nu}[\,x_i^{\mu}x_i^{\nu}-x_i^{\mu}x_j^{\nu}-x_j^{\mu}x_i^{\nu}+x_j^{\mu}x_j^{\nu}\,]\\ &\equiv \frac{1}{2}\sum_{ij}\sum_{\mu\nu}x_i^{\mu}\mathcal{M}^{-1}_{ij,\mu\nu}x_j^{\nu}\,, \end{aligned} \tag{A.19}$$

that is the shape of Hamiltonian in Eq. (A.5).

Considering the fact that the matrix $\mathcal{M}^{-1}$ is symmetric changing i with j and that $\beta\to\infty$ when $T\to 0$, one has:

$$\mathcal{M}^{-1}_{ij,\mu\nu}=\begin{cases}2\sum_{l\neq i}\Delta_{il}\frac{r_{il}^{\mu}r_{il}^{\nu}}{r_{il}^{2}} & \text{if } i=j\\ -2\Delta_{ij}\frac{r_{ij}^{\mu}r_{ij}^{\nu}}{r_{ij}^{2}} & \text{if } i\neq j\,.\end{cases} \tag{A.20}$$

A.2 βGM functional motion of BACE

In the past few decades several experimental and theoretical studies have pointed out that, although proteins possess an atomic density comparable to that of crystalline solids, they are much more flexible that the latter [108, 109, 110]. This

unusual elasticity, whose origin putatively resides in the neat secondary and tertiary protein organization [111], is a prerequisite for biological functionality. In fact, in order to carry out their biological tasks, proteins and enzymes need to sustain conformational distortions where groups of several amino acids are significantly displaced from the reference native states. The timescales associated with such configurational changes, which may occur over several nanoseconds, are also long compared to those of atomic motions [112].

From a computational point of view, the most direct way of observing these rearrangements would be by recourse to a molecular dynamics simulation [2]. Present implementations of this scheme allow one to follow the dynamical evolution of a large protein (of a few hundred residues) in its surrounding solvent for about ten nanosecond. This timescale is large enough for gaining considerable insight into several dynamical aspects of large proteins but may be inadequate for characterizing accurately the functional movements mentioned before [113].

Several studies have attempted to bridge the gap between the timescales of feasible MD simulations and those of biologically relevant protein movements by resorting to a mesoscopic rather than a microscopic approach [99]. A key contribution in this framework has been the observation that the overdamped dynamics of a protein in its solvent can be described as occurring in an effective quadratic potential [114, 7, 104, 10]. This observation was further supported by Tirion who pointed out that, in a normal mode analysis of protein vibrations, the complicated classical force field could be replaced by harmonic couplings with the same spring constants [99]. These results stimulated a variety of studies where the elastic properties of proteins have been described by means of coarse–grained models where amino acids are replaced by effective centroids (corresponding to the C_α and/or C_β atoms) and the energy function is reduced to harmonic couplings between pairs of spatially close centroids. These approaches have been found to be in accord with both experimental and MD characterization of the overall protein elastic behaviour [9, 115, 116, 100]. In particular, the recently introduced βGaussian model has been shown to be quite effective in identifying, with a modest computational effort, the most important conformational changes that a

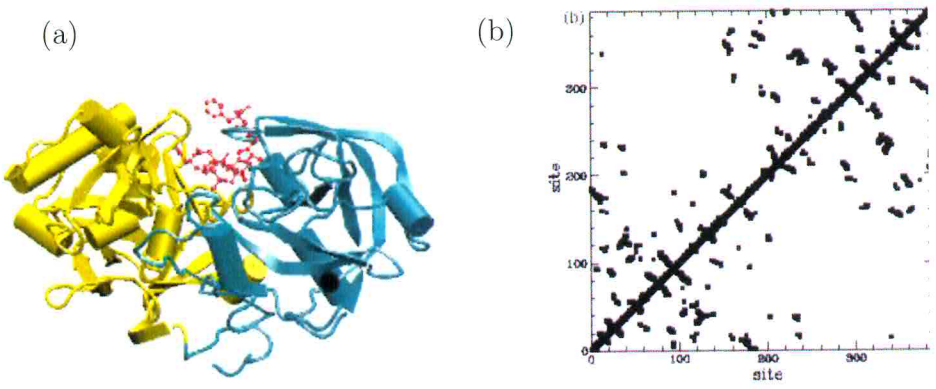


Figure A.1: (a) The structure of the human β–secretase dimer (PDB file 1fkn) complexed with an inhibiting peptide (represented with a wire frame). (b) The contact matrix, , associated with the native (crystallographic) structure of the enzyme.

protein can undergo in thermal equilibrium [62].

These vibrational modes are here investigated for a protein, human β–secretase (denoted as BACE hereafter) which is an important representative of the pepsins, a family of enzymes which are capable of cleaving a peptidic substrate through a chemical reaction (hydrolysis) involving an aspartic dyad, as discussed in detail in Section 2.2.2. Recently, the BACE, a membrane–anchored extracellular protein, has been the subject of several experimental investigations since it constitutes a key target for drugs used in treatments of Alzheimers disease (AD) [117, 118]. This reasons have motivated a large body of scientific investigations aimed at clarifying the cleavage mechanisms of the BACE and its differences in functionality with respect to typical members of the pepsin family. Interestingly, from a structural point of view, see Fig. (A.1), the core region of the BACE presents only minor differences from typical pepsin members. Most of the structural changes are instead located at the surface of the protein, in the form of six loop insertions, and at the C–terminus, by a 35–residue long extension [26].

As is visible in Fig. (A.1)a, where the BACE is shown together with a bound inhibiting peptide, a long cleft of 35 $\mathring{\text{A}}$ is present at the interface of the two lobes (highlighted with different colours). Interestingly, a very small number of residue pairs are found in contact at the lobe interface and they are mostly

constituted by the aspartic acids involved in the catalytic activity. From analogy with other enzymes [30, 79], we have found that also the BACE catalytic action depends not only on the favourable chemical interaction of the substrate and the aspartic dyads but also on the possibility that the conformational fluctuations of the enzyme itself may modulate the activity. We therefore have undertaken the task of characterizing in detail the elastic response of the BACE by resorting to the βGM. An independent term of comparison for the robustness of these findings was provided by an all–atom molecular dynamics simulation. It is found that the two methods, which are very different in spirit, provide a consistent picture for the largest protein rearrangements occurring in thermal equilibrium. Finally, we have characterized the extent to which the motion of a substrate bound to the enzyme is correlated with the latter. By these means we have identified a limited number of amino acids in the enzyme which have a strong mechanical bearing of the conformational fluctuations of the enzyme despite the lack of spatial proximity [97].

A.2.1 Conformational fluctuations of the BACE

We have applied the βGM to the native structure of the BACE shown in Fig. (A.1)a. The associated normalized reduced covariance matrix (Eq. (A.11)), computed through the βGM, is shown in Fig. (A.2). The comparison of the covariance matrix with the contact map of Fig. (A.1)b shows that the highest positive correlations are observed, as expected, in the correspondence of contacting residues. More interesting is the presence of negative correlations which signal important mechanical couplings between regions that are not in spatial proximity [78, 79].

Inspection of Fig. (A.2) reveals a significant degree of anti–correlation among residues on one lobe (especially 40–55 and 100–114) and those on the other one (especially 260–272 and 305–320). The fact that the conformational fluctuations of the two lobes are, on average, directed in opposite directions suggests that the enzymatic functional modes may be based on an opening/closing mechanism of the lobes.

This insight can be considerably refined by examining the individual con-

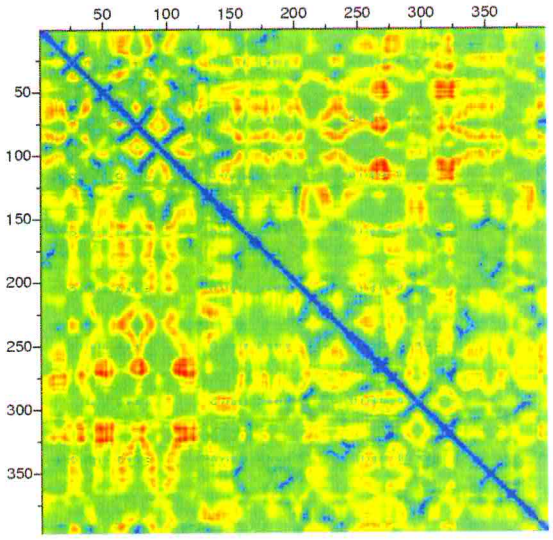


Figure A.2: Reduced covariance matrices for BACE, as obtained from βGM computations. Entries with values close to 1 (strong positive correlation) are shown in blue, while the red patches indicate anti–correlated regions.

tribution of the various eigenvectors of $\mathcal{M}$ (defined in Eq. (A.5)) to the pair correlations. It is apparent from the decomposition of Eq. (A.10) that the weight of the eigenvectors is proportional to the corresponding eigenvalues. For this reason we shall now describe the structural deformations encoded by the first three eigenvectors which, alone, are responsible for a good fraction of the overall residue mobility. By superimposing on the native structure the distortion associated with the first mode one can ascertain that it involves the movement of the lobes along opposite directions with respect to the plane identified by the cleft of the enzyme (each lobe moving almost rigidly while the connecting regions between the lobes are almost static, as shown in Fig. (A.3)). In the second mode the anticorrelated motion of the lobes is still visible but this time themotion occurs mostly parallel to the cleft plane, resulting in a shear deformation, as depicted in Fig. (A.3), while in the third mode, the two lobes appear to rotate in opposite directions. In all three modes the high mobility of the exposed loop ranging from Val309 to Asp317 is very noticeable.

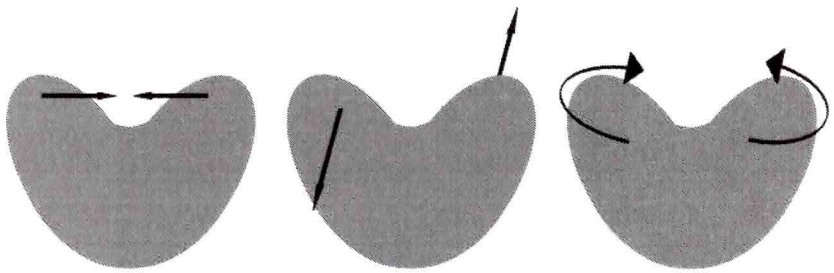


Figure A.3: Pictorial representations of the first three modes (from left to right) which describe the largest conformational distortions in the BACE.

The results obtained with the βGM relies on several simplifying assumptions (including the use of the same friction coefficient for all residues, a limitation that will be removed later). It is therefore important to verify the validity of the conclusions reached here against independent terms of reference. Ideally, one would like to compare the model results against direct experimental determination of the quantities of interest here, such as the correlation of residue motion. However, this is not currently feasible in a direct way and hence such detailed information can only be obtained from dynamical simulations with all–atom interaction potentials. MD simulation of the BACE in explicit solvent was carried out [31]. The simulated system was constituted by the protein immersed in a water box of size 75 $\mathring{\text{A}}$ $\times$87 $\mathring{\text{A}}$ $\times$90 $\mathring{\text{A}}$ to which nine sodium counter–ions were added to ensure the overall charge neutrality. The whole system, composed of about 48 000 atoms, underwent 20 ns of MD simulations (after relaxation) carried out at 300K with theGROMACS program [119].

We carried out the comparison of the βGM and MD through a series of steps which include the comparison of the overall mobility of the various amino acids, of corresponding entries of the covariance matrices and of the largest eigenvectors of the covariance matrix. These comparisons are here exploited not only to check the consistency of the two approaches, but also as a means to assign the βGM parameter k in Eq. (A.5) from the MD comparison.

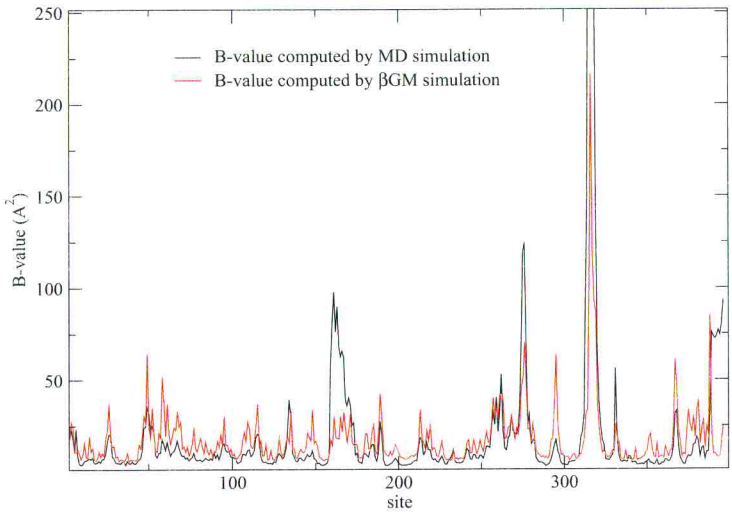


Figure A.4: B–value (defined in Eq. (1.34)) of the various residues in the BACE obtained from MD and calculated from the βGM after an optimal choice of the harmonic coupling parameter, k.

The first quantity that we shall compare is the mean square fluctuation which summarizes the overall mobility of any given residue. Within the βGM this quantity is given by the diagonal elements of $\mathcal{M}$:

$$\langle \vec{x}_i \cdot \vec{x}_i \rangle = \sum_{\alpha} \mathcal{M}_{ii,\alpha\alpha} \ , \tag{A.21}$$

while, in the context of molecular dynamics, the thermodynamic average of Eq. (A.21) is aptly replaced by the time average over the simulation run. It should also be noted that, usually, it is not easy to ascertain whether the simulated trajectory is sufficiently long that thermodynamic averages can be legitimately replaced with dynamical ones. A practical way of checking a posteriori the validity of this ergodicity assumption is to check the consistency of the essential eigenspaces pertaining to different parts of the simulated trajectories. Indeed, this type of analysis shows that when the simulated time span greatly exceeds the typical timescale associated with the slow dynamical modes then the essential subspaces are robustly identified [120]. Besides this strategy, a recent theoretical

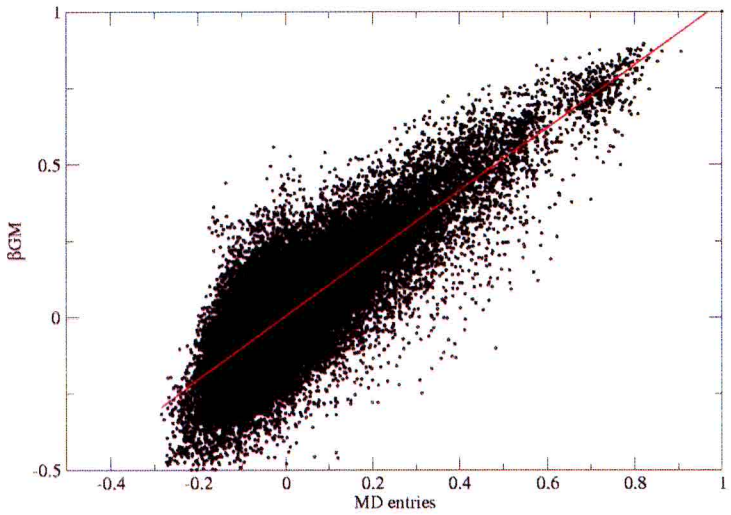


Figure A.5: A scatter plot of corresponding entries of the reduced covariance matrices obtained by the βGaussian model and MD simulations of BACE. The interpolating line, determined by linear regression, is also shown.

study has introduced some valuable quantitative criteria by which it is possible to decide whether a given MD trajectory is too short to meet the ergodicity requirements [113]. These criteria are, again, based on the analysis of the essential dynamical spaces. When the dynamical sampling of phase space is insufficient, the eigenvectors of the essential spaces have distinctive cosine–like shapes. Conversely, it has been shown that when the cosine content of the relevant dynamical eigenspaces is negligible, the ergodicity assumption appears justified [113]. In the present study we have adopted both these criteria: the total simulation time span of 20 ns is about thirty times bigger than the longest autocorrelation time found in the system; in addition the cosine content of the top five essential eigenvectors was, on average, below 10%.

It must be noted that, within the βGM, the mean square displacements are proportional to $K_B T/k$. The comparison of the MD and model mean square displacements offers, therefore, the opportunity to set the effective value of k by, *e.g.*, matching the average mean square displacements in the two cases. This

criterion yields, for this particular protein, the value $k^{-1} = 0.7$ Å^2 $/K_BT$. The profiles for the residues mobility according to MD and the βGM (having fixed k to the value mentioned above) are shown in Fig. (A.4).

As mentioned before, Fig. (A.4) reveals the high mobility of the exposed loop spanning residues 309–317. Aside from this, the two profiles have a good degree of correlation since the linear correlation coefficient is 0.78.

Having ascertained the consistency of the overall residue mobility of MD and βGM we analysed the agreement of the covariance matrices which convey information on pair correlations. We carried out the comparison in two stages: first by comparing corresponding entries of the matrices and subsequently by measuring the overlap between the significant subspaces of the matrices.

The degree of accord of the normalized reduced covariance matrices of the MD and βGM is conveniently summarized through a scatter plot of corresponding entries of the two matrices. The results are summarized in Fig. (A.5) where we have deliberately omitted the diagonal entries of the matrix which are equal to 1 in both cases. Since the length of the BACE approaches 400 residues, the data set shown in Fig. (A.5) is constituted by more than 7.5×104 distinct data points. The linear correlation coefficient of the MD and βGM data is r = 0.83. This represents a strikingly high value given the huge number of degrees of freedom. We conclude the discussion of the scatter plot of Fig. (A.5) by mentioning that, in the case of perfect correlation of two normalized (adimensional) covariance matrices, the data would align along the diagonal of the graph in Fig. (A.5). Interestingly, the best–fit line lies very close to the diagonal having a slope of s = 1.04, a fact which further testifies to the overall viability of the Gaussian scheme.

As a final test, we wish to measure the consistency of the most important eigenspaces of the two matrices, that is the eigenvectors associated with the largest eigenvalues of the covariance matrix (for βGM this is equivalent to considering the largest eigenvalues of $\mathcal{M}$, see Eq. (A.10)). In fact, these eigenspaces are the ones that describe the most significant modes of distortion of the molecule in the three–dimensional space. Various strategies have been proposed for evaluating the accord of the significant eigenspaces of the covariance matri-

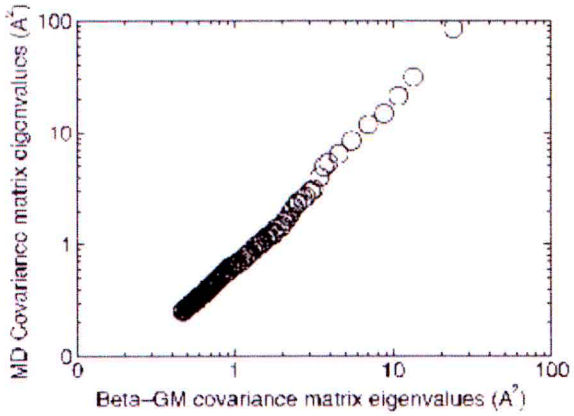


Figure A.6: A scatter plot of equally ranking eigenvalues of the MD and βGM covariance matrices.

ces [10, 56, 113]. Here we have adopted the simple strategy of considering the top 10 eigenvectors obtained from MD and the βGM and calculating the modulus of the scalar product of any pair of vectors coming from the two sets. It is found that the overlap between the two essential eigenspaces is considerable since about 75% of the norm of the first four eigenvectors of the βGM is projected onto the first four eigenvectors of MD.

It is also interesting to evaluate the agreement of corresponding (*i.e.* equally ranking) eigenvalues of the covariance matrices. This analysis is carried out in the scatter plot of Fig. (A.6) where, for the βGM, the ith largest ranking eigenvalue of the covariance matrix $\mathcal{M}$ was taken multiplied by the previously found coefficient K_BT/k. Since the eigenvalues span a few orders of magnitude, the plot is presented in a log–log format and reveals a linear relationship.

After having characterized the accord of the conformational fluctuations occurring in thermal equilibrium within the all–atom MD and the βGM we focus on the characterization of the mechanical coupling between the BACE enzyme and the substrate (peptide) on which it acts. This analysis is accomplished by applying the Gaussian model not to the isolated enzyme but to the enzyme/substrate complex shown in Fig. (A.1)a. The crucial questions to formulate in this context

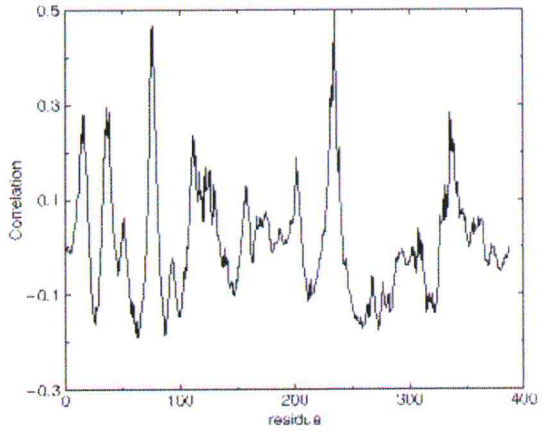


Figure A.7: The curve indicates the degree of correlation, calculated with βGM, of a residue in the middle of the inhibiting peptide and the 387 residues that constitute the BACE.

relate to the possible existence of non–trivial mechanical couplings between the substrate and residues in the enzyme. For other enzymes, in particular the HIV-1 PR, it has been shown that several residues of HIV-1 PR despite being far away from the substrate have an important mechanical bearing on the latter [62,30]. In fact, as the explicit MD calculation has shown, the high degree of such coupling between such sites and the cleavage is such that the detailed chemical identity of the former strongly influences the substrate binding affinity of the latter [60]. Therefore, mechanical couplings can be so important for enzymatic catalysis that mutations at a small number of key enzymatic sites (even if distant from the active site) can dramatically alter the enzyme reactivity. We have therefore undertaken a similar analysis here and calculated the degree of correlation of the displacement of residues in the substrate and all residues in the BACE. The correlation profiles are nearly identical for all residues in the substrate and a representative profile is shown in Fig. (A.7).

From the profile a strong positive correlations is discernible in correspondence with residues 72 and 230 (numbering according to the PDB file). This reflects the close spatial proximity of the substrate and these contacting residues. The

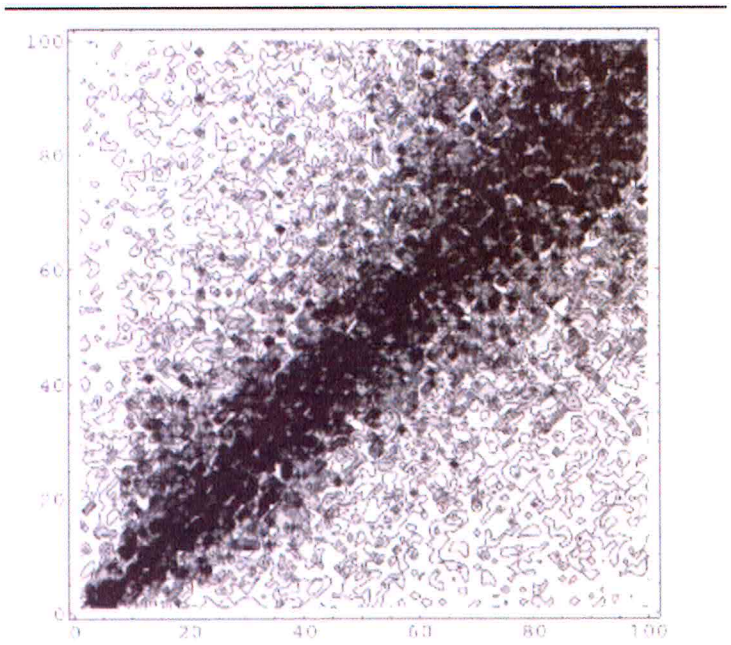


Figure A.8: Density plot for the modulus of the scalar product of the eigenvectors of matrix $\tilde{\mathcal{M}}$ (Eq. (A.12)) (rank index given on the x axis) against those of matrix (rank index given on the y axis). Entries with values close to 1 [0] are shown in black [white].

crucial feature is however contained in the peaks of negative correlations which occur in correspondence with residues 23, 61, 83, 255, 269, 317. These residues are very distant from the substrate (25–30 $\mathring{\text{A}}$) and yet are able to influence the conformational fluctuations of the substrate bound to the active site. These sites are located at the outward face of each of the two lobes, in close analogy to the key mechanical sites found for HIV–1 PR. This fact provides additional evidence in support of the steering of the substrate towards a reactive conformation being achieved not only through finely tuned chemical couplings with the enzymes but also through subtle mechanical influence of sites far apart from the active region.

We conclude our analysis by discussing the possibility of gaining insight through the βGM into the time–scales involved in the most important conformational fluctuations. To do so it is necessary to have a quantitative estimate for the effective

friction coefficients appearing in Eq. (A.6). It has been pointed out before [2, 10] that the effective friction coefficients experienced by the various amino acids in a protein depend not only on the interaction with the solvent but are severely affected by the protein density itself, in their neighbourhood. In particular, Hinsen *et al* [103] have provided a phenomenological characterization of the viscous coefficient which they found to have an approximately linear relationship in terms of the local atomic density. We have translated this phenomenological relationship to one in terms of the number of contacts, n_i^c , in which residue i takes part (calculated with a cut–off distance of 7.5 $\overset{\circ}{\text{A}}$) and thus arrived at the phenomenological relationship:

$$\gamma_i = B n_i^c \,, \tag{A.22}$$

where $B = 1.53 \times 10^3$ amu ps^{-1}. We have therefore calculated the various modes of relaxation in this new context where the friction coefficient is not the same for all residues. Strikingly we have found that there is a good correspondence between the modes of relaxation (which depend on γ_i and reflect dynamical properties) and the eigenvectors which describe conformational fluctuations in thermal equilibrium. This correspondence is obviously exact in the case when γ_i is the same for all sites but is not otherwise expected in the case of heterogeneity of the friction coefficient. The degree of accord of the two sets of eigenvectors is shown in Fig. (A.8) and indeed the thick cloud around the diagonal highlights their good correspondence. The timescales which control the decay of conformational fluctuations in a dynamical trajectory are given by the eigenvalues of $\widetilde{\mathcal{M}}$ in Eq. (A.12) divided by k. Therefore, we obtain that the slowest relaxation time calculated within the Gaussian model has a decay time of 1.2 ns. The analysis of the decay of autocorrelation in the whole MD trajectory indicates 0.7 ns as the slowest relaxation time of the system. It appears, therefore, that the simple Gaussian approach can correctly identify the timescales of the system autocorrelation within a factor of two. Therefore, despite the several approximations which are at the basis of the βGM, the latter appears to be useful also for estimating, with a modest computational expenditure, the correct order of magnitude of the system autocorrelation time. This may be exploited to obtain a preliminary indication

of the lapse of time that needs to be covered in an all–atom MD simulation to ensure that the system has sufficient time to explore significantly different regions of the phase space.

Appendix B

Supporting information for the OmpT complexes simulations

Wild–type Enzyme

The root mean square deviations (RMSD) of the C_α of **A** and **B** is stable after $0.05\mu s$ and fluctuates around an average value of $\sim$3 Å, as shown in Fig (B.1)ab [1].

The RMSD values for the C_α atoms of the substrate relative to the MD snapshot at 0.05 μs of the wt as a reference structure (Fig. (B.3)ab), are either characterized by small fluctuations ($\sim$ 0.4 Å) around an average value of $\sim$ 2.3 Å (protomer **A** , see Fig. (B.3)a or they show many peaks featuring fluctuations of larger amplitude of $\sim$ 1.7 Å (protomer **B** , see Fig. (B.3)b). In addition, based on a calculation of the root mean square fluctuations (RMSF) of each residue, we conclude that loops L2, L3 and L4 (Fig. (3.4)a) as well as the substrate in **B** are more mobile than in **A**. Thus, in **B** the substrate and loops L2–L4 explores many conformational sub–states during the simulation (Fig (B.4)a). This is consistent with the analysis of the large scale motions of the protomers, which suggest that the motions of the substrate in the catalytic cleft are concerted with that of the loops, the correlation being larger for **B** (Fig. (B.4)b).

[1] RMSD of **C** and **D** are not reported, as after only 10 ns the geometries of the active site become very distorted (see Chapter 3.2.2 and Fig. (B.2)ab).

The δ and ξ distances, defined in Chapter 3.2.2, characterize the motion of the substrate inside the catalytic cleft. In **A**, as shown in Fig. (B.5)ab, δ and ξ oscillate around an average value ($\langle\delta\rangle \sim 7.8(0.5)$ Å and $\langle\xi\rangle \sim 24.1(0.4)$ Å) with a sharp Gaussian–like distribution. In contrast, in **B** the distances feature bimodal distributions, as discussed in Chapter 3.2.2.

Mutants

In H212A, the side chains of R2 and R3 of the substrate, which form a salt bridge with E27–D208 and D85–D97, respectively, rotate and face the solvent at the end of the simulation (0.05 μs, Fig. (B.6)ab).

In S99A (protomer **A**) R2 side chains interacts at 0.05 μs with D210 breaking the D210–H212 H–bond. This causes a motion of H212 away from the catalytic site (see Fig. (B.7)a). In S99A (protomer **B**), the salt bridge between R3 and D85–D97 is disrupted after 0.02 μs. This causes, as for H212A and S99A of **A** mutants, a motion of the substrate carbonyl carbon away from the catalytic dyad D83–H212 (see Fig. (B.7)b).

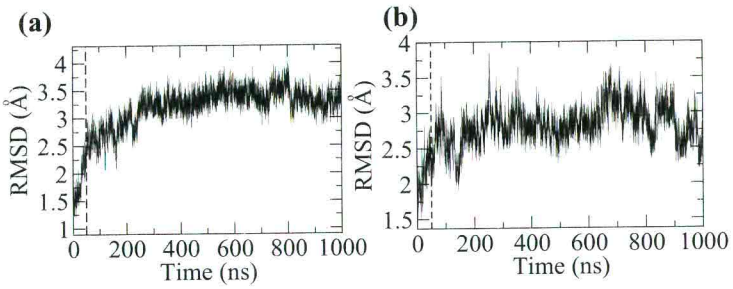


Figure B.1: Time evolution of RMSD computed on **A**: (a), **B**: (b).

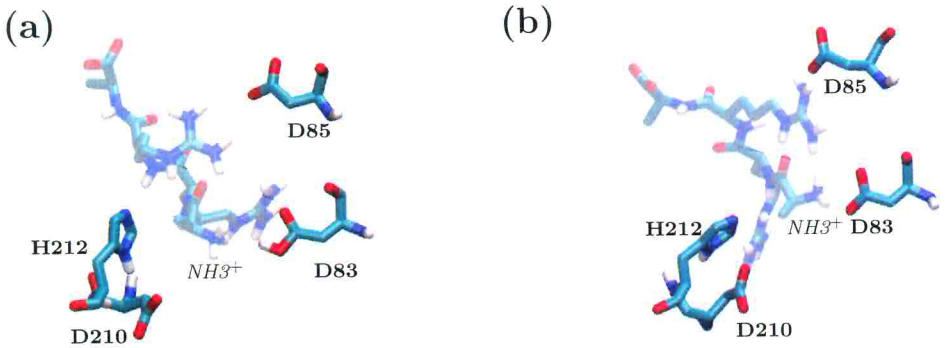


Figure B.2: A snapshot of the active site at 10 ns of MM/CG simulation for protomer **C**: (a) and for protomer **D**: (b). The substrate is depicted with a transparent effect.

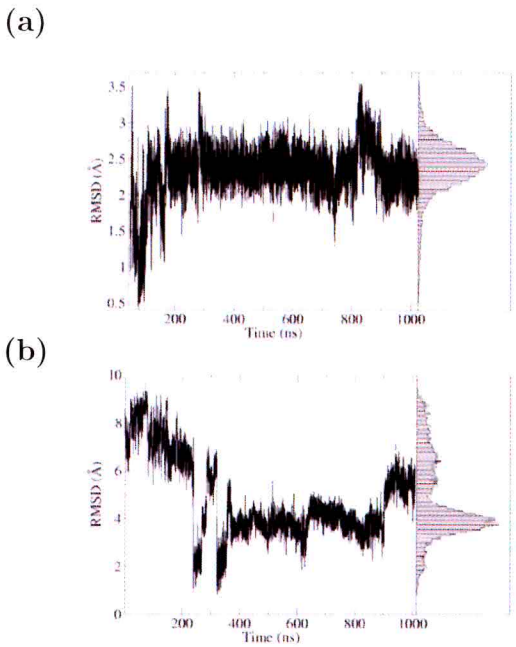


Figure B.3: RMSD time evolution of the C_α atoms of the substrate of **A**: (a), and **B**: (b). On the right of each graph there is the correspondent distribution of RMSD values.

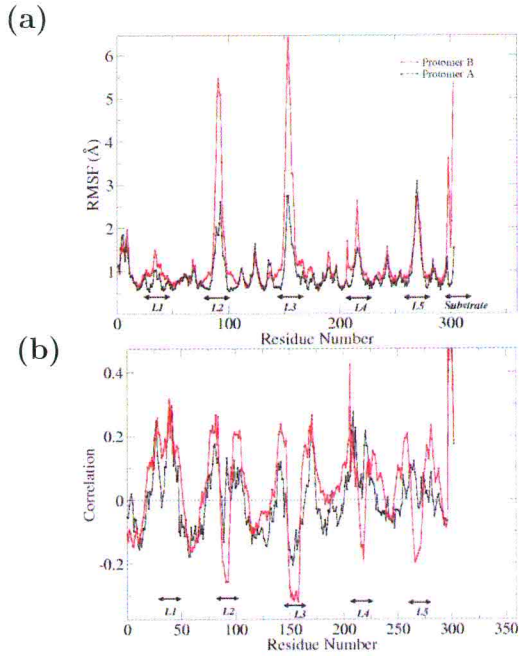


Figure B.4: (a) Root mean square fluctuations related to Protomer A and B (black and red, respectively). (b) The curve indicates the degree of correlation, calculated with the normalized reduced covariance matrix, of the C_*alpha*a atom of residue R300.

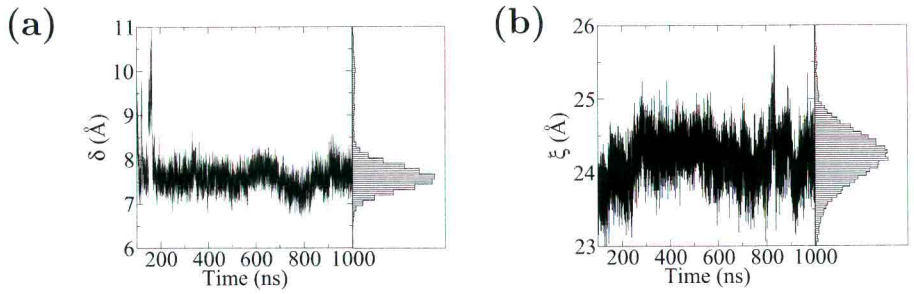


Figure B.5: The two pictures depict the time evolution of the distance between the center of mass of the C_α atoms of the substrate and the center of mass of the dyad D83, D85: (a), and the center of mass of the β–Barrel: (b).

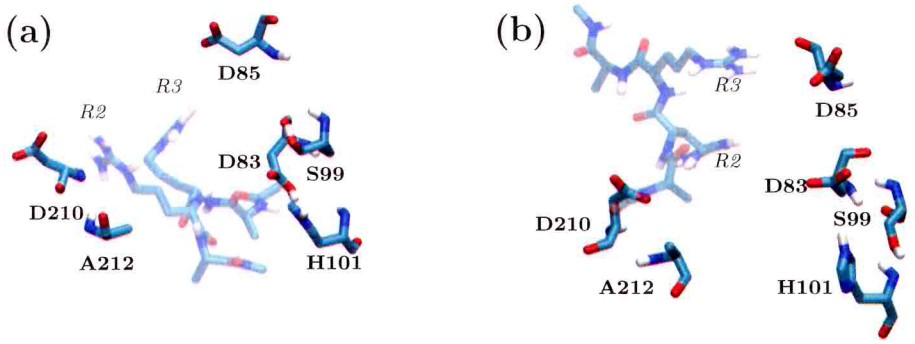


Figure B.6: Geometry of the active site after 150 ns of the H212A mutant on protomer **A**: (a) and on protomer **B**: (b). The substrate is depicted with a transparent effect.

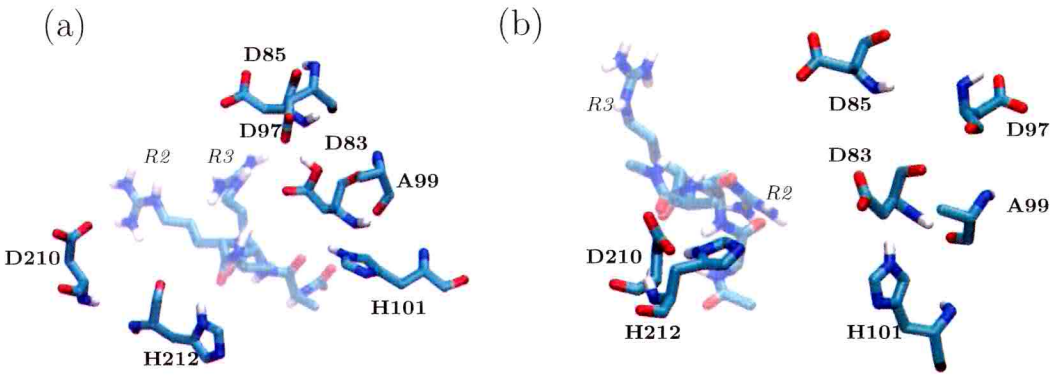


Figure B.7: Geometry of the active site after 150 ns of the S99A mutant on protomer **A**: (a) and on protomer **B**: (b).

Bibliography

[1] S. A.Adcock and J. A. McCammon. Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.*, 106:1589–1615, 2006.

[2] M. Karplus. Molecular dynamics simulations of biomolecules. *Acc. Chem. Res.*, 35:321–323, 2002.

[3] C. L. Brooks, M. Karplus, and B. M. Pettitt. *Proteins: a theoretical perspective of dynamics, structure, and thermodynamics.* Wiley, New York, 1988.

[4] M. Levitt and A. Warshel. Computer simulation of protein folding. *Nature*, 253:694–698, 1975.

[5] A. Kolinski, A. Godzik, and J. Skolnick. A general method for the prediction of the three dimensional structure and folding pathway of globular proteins: Application to designed helical proteins. *J. Chem. Phys.*, 98:7420–7433, 1993.

[6] C. Micheletti, F. Seno, J. R. Banavar, and A. Maritan. Learning effective amino acid interactions through iterative stochastic techniques. *Proteins*, 42:422–431, 2001.

[7] T. Noguti and N. Go. Collective variable description of small-amplitude conformational fluctuations in a globular protein. *Nature*, 296:776–778, 1982.

[8] S. Swaminathan, T. Ichiye, W. van Gusteren, and M. Karplus. Time dependence of atomic fluctuations in proteins: analysis of local and collective motions in bovine pancreatic trypsin inhibitor. *Biochemistry*, 21:5230–5241, 1982.

[9] I. Bahar, A. R. Atilgan, and B. Erman. Direct evaluation of thermal fluctuations in proteins using a single parameter harmonic potential. *Fold. Des.*, 2:173–181, 1997.

[10] K. Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33:417–429, 1998.

[11] C. Micheletti, P. Carloni, and A. Maritan. Accurate and efficient description of protein vibrational dymanics: Comparing molecular dynamics ans gaussian models. *Proteins*, 55:635–645, 2004.

[12] Y. Zhou and M. Karplus. Interpreting the folding kinetics of helical proteins. *Nature*, 401:400–403, 1999.

[13] T. Schlick and J. Olson. Trefoil knotting revealed by molecular dynamics simulations of supercoiled dna. *Science*, 257:1110–1115, 1992.

[14] J. Trylska, V. Tozzini, and J. A. McCammon. Exploring global motions and correlations in the ribosome. *Biophys J.*, 89:1455–1463, 2005.

[15] S. O. Nielsen, C. F. Lopez, G. Srinivas, and M. L. Klein. Coarse grain models and the computer simulation of soft materials. *J. Phys.-Condens. Matter*, 16:R481–R512, 2004.

[16] P. J. Bond and M. S. Sansom. Insertion and assembly of membrane proteins via simulation. *J Am Chem Soc.*, 128:2697–2704, 2006.

[17] S. Takada. Protein folding simulation with solvent-induced force field: folding pathway ensemble of three-helix-bundle proteins. *Proteins*, 42:85–98, 2001.

[18] Y. Fujitsuka, S. Takada, Z. A. Luthey-Schulten, and P. G. Wolynes. Optimizing physical energy functions for protein folding. *Proteins*, 54:88–103, 2004.

[19] M. Khalili, A. Liwo, F. Rakowski, P. Grochowski, and H. A. Scheraga. Molecular dynamics with the united-residue model of polypeptide chains. Langevin and berendsen-bath dynamics and tests on model alpha-helical systems. *J Phys Chem B Condens Matter Mater Surf Interfaces Biophys*, 109:13785–13810, 2005.

[20] I. A. Hubner, E. J. Deeds, and E. I. Shakhnovich. High-resolution protein folding with a transferable potential. *Proc Natl Acad Sci U S A*, 102:18914–18919, 2005.

[21] M. Neri, M. Cascella, and C. Micheletti. Influence of conformational fluctuations on enzymatic activity: modelling the functional motion of $\beta-$secretase. *J. Phys.-Condens. Matter*, 17:S1581–S1593, 2005.

[22] H. Abe and N. Go. Noninteracting local-structure model of folding and unfolding transition in globular proteins. Application to two-dimensional lattice proteins. *Biopolymers*, 20:1013–1031, 1981.

[23] C. Micheletti, J. R. Banavar, A. Maritan, and F. Seno. Protein structures and optimal folding from a geometrical variational principle. *Phys. Rev. Lett.*, 82:3372–3375, 1999.

[24] M. Neri, C. Anselmi, M. Cascella, A. Maritan, and P. Carloni. Coarse-grained model of proteins incorporating atomistic detail of the active site. *Phys. Rev. Lett.*, 95:DOI:218102–1, 2005.

[25] A. J. Barrett, N. D. Rawlings, and J. F. Wössner. *Handbook of Proteolytic Enzymes*. Academic Press, London, 1998.

[26] L. Hong, G. Koelsch, X. Lin, S. Wu, S. Terzyan, A. K. Ghosh, X. C. Zhang, and J. Tang. Structure of the protease domain of memapsin 2 (beta-secretase) complexed with inhibitor. *Science*, 290:150–153, 2000.

[27] S. J. Pollack and H. Lewis. Secretase inhibitors for alzheimer's disease: challenges of a promiscuous protease. *Curr. Opin. Investig. Drugs*, 6:35–47, 2005.

[28] M. Miller, J. Schneider, B. K. Sathyanarayana, M. V. Toth, G. R. Marshall, L. Clawson, L. M. Selk, S. B. H. Kent, and A. Wlodawer. Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 a resolution. *Science*, 246:1149–1152, 1989.

[29] M. Boffito *et al.* Current status and future prospects of therapeutic drug monitoring and applied clinical pharmacology in antiretroviral therapy. *Antivir. Ther.*, 10:375–392, 2005.

[30] S. Piana, P. Carloni, and M. Parrinello. Role of conformational fluctuations in the enzymatic reaction of HIV-1 protease. *J. Mol. Biol.*, 319:567–583, 2002.

[31] M. Cascella, C.Micheletti, U. Rothlisberger, and P. Carloni. Evolutionarily conserved functional mechanics across pepsin-like and retroviral aspartic proteases. *J. Am. Chem. Soc.*, 127:3734–3742, 2005.

[32] L. Vandeputte-Rutten, R. A. Kramer, J. Kroon, N. Dekker, M. R. Egmond, and P. Gros. Crystal structure of the outer membrane protease ompt from escherichia coli suggests a novel catalytic site. *EMBO J.*, 20:5033–5039, 2001.

[33] J. D. Faraldo-Gomez, L. R. Forrest, M. Baaden, P. J. Bond, C. Domene, G. Patargias, J. Cuthbertson, and M. S. Sansom. Conformational sampling and dynamics of membrane proteins from 10-nanosecond computer simulations. *Proteins*, 57:783–791, 2004.

[34] M. Neri, C. Anselmi, V. Carnevale, A. V. Vargiu, and P. Carloni. Molecular dynamics simulations of outer-membrane protease t from *E. coli* based on a hybrid coarse-grained/atomistic potential. *J. Phys.-Condens. Matter*, 18:S347–S355, 2006.

[35] Marc Baaden and Mark Sansom (private comunication) kindly provided us the final configuration of the four protomers (**A**–**D**) of OmpT colplexes after ∼24 ns.

[36] M. Neri, V. Carnevale, C. Anselmi, A. Maritan, and P. Carloni. Microseconds dynamics simulation of the outer membrane proteases t. *Submitted*, 2007.

[37] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman. A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges–the resp model. *J. Phys. Chem.*, 94:10269–10280, 1993.

[38] W. D. Cornell, P. Cieplak, C. I. Bayly, and P. A. Kollman. Application of resp charges to calculate conformational energies, hydrogen–bond energies, and free–energies of solvation. *J. Am. Chem. Soc.*, 115:9620–9631, 1993.

[39] W. F. van Gunsteren and H. J. Berendsen. A leap-frog algorithm for stochastic dynamics. *Mol. Sim.*, 1:173–185, 1988.

[40] J. P. Ryckaert, G. Ciccotti, and H. J. C. Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n–alkanes. *J. Comput. Phys.*, 23:327–341, 1977.

[41] H. J. C. Berendsen, J. P. M. Postma, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[42] S. Nosé and M. L. Klein. Constant pressure molecular dynamics for molecular systems. *Mol. Phys.*, 50:1055–1076, 1983.

[43] S. Nosé. A molecular dynamics method for simulations in the canonical ensemble. *Mol. Phys.*, 52:255–268, 1984.

[44] S. Takada. Go-ing for the prediction of protein folding mechanisms. *Proc. Natl. Acad. Sci. U.S.A.*, 96:11698–11700, 1999.

[45] E. Alm and D. Baker. Prediction of protein folding mechanisms from free energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. USA*, 96:11305–11310, 1999.

[46] N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich. Discrete molecular dynamics studies of the folding of a protein-like model. *Folding Des*, 3:577–587, 1998.

[47] C. Hardin, Z. Luthey-Schulten, and P. G. Wolynes. Backbone dynamics, fast folding, and secondary structure formation in helical proteins and peptides. *Proteins: Struct. Funct. Genet.*, 34:281–294, 1999.

[48] C. Clementi, H. Nymeyer, and J. N. Onuchic. Topological and energetic factors: what determine the structural details of the transition state ensemble and n-route intermediates for protein folding? an investigation for small globular proteins,. *J. Mol. Biol.*, 298:937–953, 2000.

[49] M. S. Li and M. Cieplak. Folding in two-dimensional off-lattice models of proteins. *Phys. Rev. E*, 59:970–976, 1999.

[50] G. Iori, E. Marinari, and G. Parisi. Random self-interacting chains: a mechanism for protein folding. *J. Phys. A*, 24:5249–5362, 1991.

[51] C. Clementi, A. Maritan, and J. R. Banavar. Folding, design and determination of interaction potentials using off-lattice dynamics of model heteropolymers. *Phys. Rev. Lett.*, 81:3287–3290, 1998.

[52] M. Cieplak and T. X. Hoang. Scaling of folding properties in go models of proteins. *J. Biol. Phys.*, 26:273–294, 2000.

[53] M. Doi. *Introduction To Polymer Physics.* Oxford science publications, Great Britain, first edition, 1996.

[54] D. van der Spoel, E. Lindahl, B. Hess, A. R. van Buuren, E. Apol, P. J. Meulenhoff, D. P. Tieleman, A. L. T. M. Sijbers, K. A. Feenstra, R. van Drunen, and H. J. C. Berendsen. *Gromacs User Manual version 3.2.* 2004.

[55] W. F. van Gunsteren, S. R. Billeter, A. A. Eising, P. H. Hünenberg, P. Krüger, A. E.Mark, W. R. P Scott, and I. G. Tironi. *Biomolecular Simulation: The GROMOS96 manual and user guide.* Hochschulverlag AG an der ETH Zurich, Zurich, Switzerland, 1996.

[56] A. Amadei, A. B. M. Linssen, and H. J. C. Berendsen. Essential dynamics of proteins. *Proteins*, 17:412–425, 1993.

[57] B. Hess. Similarities between principal components of protein dynamics and random diffusion. *Phys. Rev. E*, 62:8438–8448, 2000.

[58] B. Hess. Convergence of sampling in protein simulations. *Phys. Rev. E*, 65:031910, 2002.

[59] S. Piana, P. Carloni, and U. Rothlisberger. Drug resistance in HIV-1 protease: Flexibility-assisted mechanism of compensatory mutations. *Protein Sci.*, 11:2393–2402, 2002.

[60] S. Piana, P. Carloni, and U. Rothlisberger. Drug resistance in HIV-1 protease: flexibility-assisted mechanism of compensatory mutations. *Protein Science*, 11:2393–2402, 2002.

[61] S. Piana and P. Carloni. Conformational flexibility of the catalytic asp dyad in HIV-1 protease: an ab initio study of the free enzyme. *Proteins*, 39:26–36, 2000.

[62] C. Micheletti, P. Carloni, and A. Maritan. Accurate and efficient description of protein vibrational dynamics: Comparing molecular dynamics and gaussian models. *Proteins*, 55:635–645, 2004.

[63] L. Vandeputte-Rutten and P. Gros. Novel proteases: common themes and surprising features. *Curr. Op. Struc. Biol.*, 12:704–708, 2002.

[64] J. Tang, P. Sepulveda, J. Marciniszyn, K. C. S. Chen, W. Y. Huang, N. Tao, D. Liu, and J. P. Lanier. Amino-acid sequence of porcine pepsin. *Proc. Natl. Acad. Sci. USA*, 70:3437–3439, 1973.

[65] D. B. Northrop. Follow the protons: A low-barrier hydrogen bond unifies the mechanisms of aspartic proteases. *Acc. Chem. Res.*, 34:790–797, 2001.

[66] A. R. Sielecki, M. Fujinaga, R. J. Read, and M. N. G. James. Refined structure of porcine pepsinogen at 1.8 a resolution. *J. Mol. Biol.*, 219:671, 1991.

[67] T. Hofmann, R. S. Hodges, and M. N. G. James. Effect of ph on the activities of penicillopepsin and rhizopus pepsin and a proposal for the productive substrate binding mode in penicillopepsin. *Biochemistry*, 23:635–643, 1984.

[68] X. Lin, G. Koelsch, S. Wu, D. Downs, A. Dashti, and J. Tang. Human aspartic protease memapsin 2 cleaves the beta-secretase site of beta-amyloid precursor protein. *Proc. Natl. Acad. Sci. USA*, 97:1456–1460, 2000.

[69] M. Miller, M. Jaskolski, R. Mohana, J. Leis, and A. Wlodawer. Crystal-structure of a retroviral protease proves relationship to aspartic protease family. *Nature*, 337:576–579, 1989.

[70] P. L. Darke, R. F. Nutt, S. F. Brady, V. .M Garsky, T. M. Ciccarone, C. T. Leu, P. K. Lumma, R. M. Freidinger, D. F. Veber, and I. S. Sigal. HIV-1 protease specificity of peptide cleavage is sufficient for processing of gag and pol polyproteins. *Biochem. Biophys. Res. Comm.*, 156:297–303, 1988.

[71] L. H. Pearl and W. R. Taylor. A structural model for the retroviral proteases. *Nature*, 329:351–354, 1987.

[72] M. N. G James, I. N. Hsu, and L. T. J. Delbaere. Penicillopepsin from penicillium janthinellum: Crystal structure at 2.8 Å and sequence homology with porcine pepsin. *Nature*, 266:140–145, 1977.

[73] I. N. Hsu, L. T. J. Delbaere, M. N. G. James, and T. Hofmann. Mechanism of acid protease catalysis based on the crystal structure of penicillopepsin. *Nature*, 267:808–813, 1977.

[74] B. Veerapandian, J. B. Cooper, A. Sali, T. L. Blundell, R. L. Rosati, B. W. Dominy, D. B. Damon, and D. J. Hoover. Direct observation by x-ray-analysis of the tetrahedral intermediate of aspartic proteinases. *Prot. Sci.*, pages 322–328, 1992.

[75] M. N. G. James, A. R. Sielecki, K. Hayakawa, and M. H. Gelb. Crystallographic analysis of transition-state mimics bound to penicillopepsin - difluorostatine-containing and difluorostatone-containing peptides. *Biochemistry*, 31:3872–3886, 1992.

[76] J. Marcinkeviciene, L. M Kopcho, T. Yang, R. A. Copeland, B. M. Glass, A. P. Combs, N. Falahatpisheh, and L. Thompson. Novel inhibition of porcine pepsin by a substituted piperidine - preference for one of the enzyme conformers. *J. Biol. Chem.*, 32:28677–28682, 2002.

[77] A. M. Silva, S. V. Gulnik, P. Majer, J. Collins, T. N. Baht, P. J. Collins, R. E. Cachau, K. E. Luker, I. Y. Gluzman, S. E. Francis, A. Oksman, D. E. Goldberg, and J. W. Erickson. Structure and inhibition of plasmepsin ii, a hemoglobin-degrading enzyme from plasmodium falciparum. *Proc. Natl. Acad. Sci. USA*, 93:10034–10039, 1996.

[78] S. Piana, P. Carloni, and M. Parrinello. Role of conformational fluctuations in the enzymatic reaction of HIV–1 protease. *J. Mol. Biol.*, 319:567–583, 2002.

[79] S. Piana, P. Carloni, and U. Rothlisberger. Drug resistance in HIV-1 protease: Flexibility–assisted mechanism of compensatory mutations. *Prot. Sci.*, 11:2393–2402, 2002.

[80] A. L. Perryman, J. H. Lin, and J. A. McCammon. HIV-1 protease molecular dynamics of a wild–type and of the V82F/I84V mutant: Possible contributions to drug resistance and a potential new target site for drugs. *Protein Sci.*, 13:1108–1123, 2004.

[81] P. J. Bond, J. D. Faraldo-Gomez, and M. S. Sansom. Ompa: a pore or not a pore? simulation and modeling studies. *Biophys. J.*, 83:763–775, 2002.

[82] R. A. Kramer, N. Dekker, and M. R. Egmond. Identification of active site serine and histidine residues in escherichia coli outer membrane protease ompt. *FEBS Lett.*, 468:220–224, 2000.

[83] R. A. Kramer, L. Vandeputte-Rutten, G. J. de Roon, P. Gros, N. Dekker, and M. R. Egmond. Identification of essential acidic residues of outer membrane protease ompt supports a novel active site. *FEBS Lett.*, 505:426–430, 2001.

[84] B. Hammarberg, T. Moks, M. Tally, A. Elmblad, E. Holmgren, M. Murby, B. Nilsson, S. Josephson, and M. Uhlen. Differential stability of recombinant human insulin–like growth factor ii in escherichia coli and staphylococcus aureus. *J Biotechnol.*, 14:423–437., 1990.

[85] K. Sugimura and T. Nishihara. Purification, characterization, and primary structure of escherichia coli protease vii with specificity for paired basic residues: identity of protease vii and ompt. *J Bacteriol.*, 170:5625–5632, 1988.

[86] K. Sugimura and N. Higashi. A novel outer-membrane-associated protease in escherichia coli. *J Bacteriol.*, 170:3650–3654, 1988.

[87] C. Hanke J. Hess, G. Schumacher, and W. Goebel. Processing by ompt of fusion proteins carrying the hlya transport signal during secretion by the escherichia coli hemolysin transport system. *Mol Gen Genet.*, 233:42–48, 1992.

[88] N. Dekker, R. C. Cox, R. A. Kramer, and M. R. Egmond. Substrate specificity of the integral membrane protease ompt determined by spatially addressed peptide libraries. *Biochemistry*, 40:1694–1701, 2001.

[89] S. Stumpe, R. Schmid, D. L. Stephens, G. Georgiou, and E. P. Bakker. Identification of ompt as the protease that hydrolyzes the antimicrobial peptide protamine before it enters growing cells of escherichia coli. *J. Bacteriol.*, 180:4002–4006, 1998.

[90] A. J. Barret, N. D. Rawlings, and J. F. Woessner. *Handbook of Proteolytic Enzymes.* Academic Press, San Diego, CA, USA, 1998.

[91] M. Baaden and M. S. P. Sansom. Ompt: molecular dynamics simulations of an outer membrane enzyme. *Biophys. J.*, 87:2942–2953, 2004.

[92] P. J. Bond and M. S. Sansom. The simulation approach to bacterial outer membrane proteins. *Mol. Membr. Biol.*, 21:151–161, 2004.

[93] N. Guex and M. C. Peitsch. Swiss-model and the swiss-pdb viewer: an environment for comparative protein modeling. *Electrophoresis*, 18:2714–2723, 1997.

[94] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, and J. Hermans. Interaction models for water in relation to protein hydration. In B. Pullman, editor, *Intermolecular Forces*, pages 331–342. Reidel, Dordrecht, 1981.

[95] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. Molecular dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[96] V. Carnevale, S. Raugei, C. Micheletti, and P. Carloni. Convergent dynamics in the protease enzymatic superfamily. *J. Am. Chem. Soc. In press*, 128:9766–9772, 2006.

[97] M. Neri, M. Cascella, and C. Micheletti. Influence of conformational fluctuations on enzymatic activity: modelling the functional motion of b-secretase. *J. Phys.-Condens. Matter*, 17:S1581–S1593, 2005.

[98] Y. Levy and J. N. Onuvhic. Mechanisms of protein assembly: Lessons from minimalist models. *Acc. Chem. Res*, 39:135–142, 2006.

[99] M. M. Tirion. Large amplitude elastic motions in proteins from a single–parameter, atomic analysis. *Phys. Rev. Lett.*, 77:1905–1908, 1996.

[100] P. Doruker, A.R. Atilgan, and I. Bahar. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to alpha-amylase inhibitor. *Proteins*, 40:512–524, 2000.

[101] A. R. Atilgan, S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys. J.*, 80:505–515, 2001.

[102] B. Park and M. Levitt. Energy functions that discriminate x-ray and near-native folds from well-constructed decoys. *Proteins*, 258:367–392, 1996.

[103] K. Hinsen, A. Petrescu, S. Dellerue, M. Bellissent-Funel, and G. R. Kneller. Harmonicity in slow protein dynamics. *Chem. Phys.*, 261:25–37, 2000.

[104] B. Brooks and M. Karplus. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc. Natl. Acad. Sci. U.S.A.*, 82:4995–4999, 1985.

[105] J. Howard. *Mechanics of motor proteins and the cytoskeleton.* Sinauer Associates, Sunderland, MA, 2001.

[106] S. Chandrasekhar. Stochastic problems in physics and astronomy. *Rev. Mod. Phys.*, 15:1–89, 1943.

[107] C. Micheletti, F. Cecconi, A. Flammini, and A. Maritan. Crucial stages of protein folding through a solvable model: Predicting target sites for enzyme-inhibiting drugs. *Protein Science*, 11:1878–1887, 2002.

[108] M. Levitt, C. Sander, and P. S. Stern. Protein normal-mode dynamics: trypsin inhibitor, crambin, ribonuclease and lysozyme. *J. Mol. Biol.*, 181:423–447, 1985.

[109] T. Horiuchi and N. Go. Projection of monte carlo and molecular dynamics trajectories onto the normal mode axes: human lysozyme. *Proteins*, 10:106–116, 1991.

[110] D. Ben Avraham. Vibrational normal–mode spectrum of globular proteins. *Physical Review B*, 47:14559–14560, 1993.

[111] C. Micheletti, G.L. Lattanzi, and A. Maritan. Elastic properties of proteins: Insight on the folding process and evolutionary selection of native structures. *J. Mol. Biol.*, 321:909–921, 2002.

[112] M. Garcia-Viloca *et al.* Quantum dynamics of hydride transfer catalyzed by bimetallic electrophilic catalysis: Synchronous motion of mg2+ and h- in xylose isomerase. *J. Am. Chem. Soc.*, 124:7268–7269, 2002.

[113] B. Hess. Convergence of sampling in protein simulations. *Phys. Rev. E*, 65:031910, 2002.

[114] J. A. McCammon, B. R. Gelin, M. Karplus, and P. G. Wolynes. The hinge-bending mode in lysozyme. *Nature*, 262:325–326, 1976.

[115] I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman. Vibrational dynamics of folded proteins: significance of slow and fast motions in relation to function and stability. *Physical Review Letters*, 80:2733–2736, 1998.

[116] R. L. Jernigan, M. C. Demirel, and I. Bahar. Relating structure to function through the dominant modes of motion of dna topoisomerase ii. *International Journal of Quantum Chemistry*, 75:301–312, 1999.

[117] S. Sinha and I. Lieberburg. Cellular mechanisms of β–amyloid production and secretion. *Proc. Natl. Acad. Sci. U.S.A.*, 96:11049–52, 1999.

[118] X. Lin, G. Koelsch, S. Wu, D. Downs, A. Dashti, and J. Tang. Human aspartic protease memapsin 2 cleaves the β–secretase site of β–amyloid precursor protein. *Proc. Natl. Acad. Sci. U.S.A.*, 97:1456–60, 2000.

[119] E. Lindahl, B. Hess, and D. van der Spoel. Gromacs 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.*, 7:306–317, 2001.

[120] A. Amadei, M. A. Ceruso, and A. Di Nola. On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins' molecular dynamcis simulations. *Proteins*, 36:419–424, 1999.