Evolutionarily-conserved functional Mechanics in Pepsins and Retropepsins

Thesis submitted for the degree of

Doctor Philosophiæ

Candidate:
Michele Cascella

Supervisor:
Prof. Paolo Carloni

October, 2004

Contents

Introduction

Aspartic proteases are a ubiquitous class of enzymes, which use aspartic acids into the active site to hydrolyze peptide bonds. The members of this enzymatic class are involved in a variety of metabolitic processes [1], and are important targets in different pharmacological therapies. The first aspartic protease to be sequenced was porcine pepsin [2], by Tang *et al.* in 1973, while the first 3D structure (penicillopepsin) appeared ten years later, by James and Sielecki [3]. Up to date, two families have been structurally determined: pepsins and retropepsins [1, 4].

Recently, our lab has shown that in the far most characterized retropepsin, aspartic protease from HIV-1, conformational fluctuations play a major role for the enzymatic catalysis. In fact, the protein acts as a sophisticated machinery capable of steering the substrate toward the catalytic Asp dyad to favor a reactive conformation [5, 6, 7]. This result has pointed out that the flexibility of the protein scaffold may play a crucial role in enzymatic activity, and it has been uncovered by analogous results obtained for oxido-reductase enzmyes, such as flavin reductase [8] or dihydrofolate reductase [9]. Since,in spite of the structural differences and the low sequence identity [4], it is believed that the folds of pepsins and retropepsins are evolutionarily related, the hypothesis of functional mechanical fluctuations to be a general characteristic of pepsins and retropepsins sounded like an appealling task for computational investigations.

This issue has been addressed in this thesis by a series of computational investigations: first, the reaction of hydroysis of a peptide bond in water has been inferred, in order to get some ideas of where the catalytic action of the enzyme could play a major role. In fact, although the general features of the catalytic action are understood [10], details of the enzymatic mechanism are still under debate. The study on the reference reaction has been carried on coupling Car-Parrinello [11] to Multiple-Steering Molecular Dynamics, an efficient technique originating from Jarzynski's equality [12, 13] that allows reconstruction of the free energy profiles from non-equilibrium simulations. During this thesis the reliability of this relatively new methodology has been first tested in a classical MD framwork, focusing on the single water molecule ex-

change reaction in the solvation shell of alkali ions in water [14].

Then, one specific member of the pepsin family (human β-secretase) has been studied by means of classical and hybrid QM/MM simulations [15]. The choice of β-secretase as been dictated by the intrinsic interest of this enzyme, being itself a key-target in research and devlopment of possible pharmaceutical therapies against Alzheimer's Disease [16].

The choice of DFT-MD based approaches, such as Car Parrinello or hybrid QM/MM molecular dynamics, is validated by the fact that such techniques have emerged in these last years as powerful tools for description of a variety of molecular systems of biological interest [17, 18], from enzymatic reactions [19, 7] to drug-DNA interactions [20, 21].

Finally, results found in β-secretase have been matched and compared to data from Coarse-grained computations, following a scheme developed here at SISSA [22], made on a set of different AP structures and to familial characterizations based on multiple sequence alignment.

Our results suggest the existence of an extremely refined mechanism of enzymatic activity modulation through mechanical fluctuations common to all pepsins and retropepsins, in spite of the differences in their folds.

Chapter 1

Families of Aspartic Proteases and Biology of β-Secretase

1.1 The aspartic protease enzymatic class

Hydrolysis of peptide bonds is a key step in a various number of biological processes [23]. This biological task is typically regulated by enzymatic activity. Up to date, five classes of proteases (also called peptidases) are known, namely, serine and threonine proteases, cysteine proteases, aspartic proteases and metalloproteases [24]. Examples of processes in which protease activity is important include cell growth, cell death, blood clotting, immune defence and secretion [25]. Moreover, pathogenic viruses and bacteria use proteases for their life cycle and for infection of host cells.

In past decades, the understanding of the structure and function of the aspartic proteases (APs) has greatly increased[1].

APs are directly dependent on aspartic acid residues for catalytic activity, and make up a widely distributed class of enzymes. They are ubiquitous in life beings, from vertebrates to fungi, plants and to retroviruses (see tab. 1.1). APs are characterized by optimal acid pH[2], inhibited by pepstasin, and show specificity for extended peptide substrates [1]. Up to now, only three APs families are known: pepsins, retro-

[1]Interestingly, human societies have been unwillingly getting practical benefits from enzymatic activity of APs long before knowing them: chymosin has been used for millennia in the making of cheese, and APs are known to be involved in making soy sauce, which reportedly originated during the Zhou dynasty (770-221 B.C.).

[2]in 1909, Sörensen noted that if activities of pepsin were plotted against hydrogen ion concentrations, the results were similar. To solve his scaling problem, he employed a logarithmic abscissa - and invented pH [26]

pepsins, and AP of the Cauliflower mosaic virus type [4].

Structural data are only available for two of these families: pepsins and retropepsins, showing that they are at least distantly related in evolution [10]. A total of 370 structures were available in the Protein Data Bank in 2002, as reported by Dunn [27], and the number is constantly increasing. Among them, more than 50% are of HIV-1 protease, which is one of the major targets in HIV-1 pharmaceutical therapy [28], and arguably, one of the proteins for which one the largest number of structures has been solved. Little is known about the third family, whose archetypal member has been detected in the Cauliflower Mosaic Virus [29], although it might be related as well to retropepsins [4].

1.1.1 Pepsin family

All members of the pepsin family have been found in eukaryotes. The family includes animal enzymes from the digestive tract, such as pepsin and chymosin, lysosomal enzymes, such as cathepsin D, and enzymes involved in post-translational processing, such as renin and yeast aspartic protease 3. Members from protozoa, such as Plasmodium, fungi and plants are also known. Most of the pepsin members have a molecular weight of $\approx$ 35 kd. The enzymes are approximately 330 amino acids long, with only $\approx$ 5% of sequence identity between all members of the family. The zymogens feature a N-terminal pro-peptide of up to 50 amino acids long, which is cleaved upon activation.

Pepsins are bi-lobed monomeric proteins, with the active-site cleft located between the lobes (see fig. 1.1). 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 [10, 30]. 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 fact, it has been proposed that the ancestral enzyme evolved by gene duplication followed by gene fusion [31]. Experimental evidence that dimers of the N-terminal lobe of pepsinogen can express some catalytic activity [32] confirm this hypothesis.

A large cleft, about 35 Å long, runs entirely across the molecule, and separates the two domains. Six segments make up the based pleated sheet that cross-links the two lobes; each domain contributes three strands, the central two being longer (9 amino acids) than the external one (5 amino acids). This is the only structured motif shared by the two lobes.

class	name	source	PDB code
mammalian	pepsin	human	1QRP
		pig	4PEP
(fish!)		Atlantic cod	1AM5
	gastricin	human	1HTR
	renin	human	1HRN
		mouse	1SMB
	cathepsin D	human	1LYA
	β-secretase	human	1FKN
	uropepsin	human	1FLH
	chymosin	calf	3CMS
fungal	endothiapepsin	*E.parasitica*	1EPM
	penicillopepsin	*P. janthinellum*	1BXO
	rhizopuspepsin	*R. chinesis*	3APR
	SAPs	*C. albicans*	1EAG
		C. tropicalis	1J71
	proteinase A	*S. cerevisiæ*	1DPJ
	mucorpepsin	*M. pussillus*	1MMP
		R. miehei	2ASI
	aspergillopepsin	*A. phœnicis*	1IBQ
protozoan	plasmepsins	*P. falciparum*	1PFZ
		P. vivax	1QS8
		C. chabaudi	model
plants	phytepsin	*H. vulgare*	1QDM
	cardosin	*C. cardunculus*	1B5F
viral	HIV-1		1DAZ
	HIV-2		1IDA
	SIV		1AZ5
	FIV		4FIV
	RSV		2RSP
	EIAV		2FMB

Table 1.1: Aspartic proteases of known structure. For each protein the PDB structure at best resolution has been reported. Abbreviations: SAPs= secreted aspartic peptidases of *Candida*; SIV= simian immunodeficiency virus; FIV= feline immunodeficiency virus; RSV= rous sarcoma virus; EIAV= equine infectious anaemia virus.

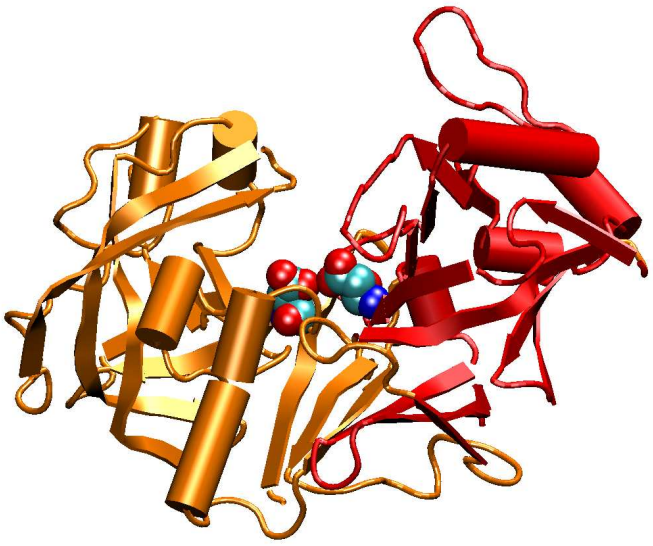


Figure 1.1: 3-D structure of porcine pepsin [2]. The N- and C-lobes are coloured in orange and red, respectively. The catalytic aspartic acids are drawn in spheres.

In almost all the members of the pepsin family, the catalytic Asp are contained in an Asp-Thr-Gly-Xaa motif, where Xaa is a serine or a threonine, the side-chain of which H-bonds directly to the Asp (Fig. 1.2).

This motif in the active site is in common with retropepsins, although in that case the Xaa residue is an Ala. This mutation is present in the C-terminal lobe of human renin. This particular pepsin, like retropepsins, is characterized by an higher pH optimum for its enzymatic activity. Site-directed mutagenesis experiments [33] support the idea that the lack of a hydrogen bond on the catalytic aspartate may influence its acidity.

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. 1.3). The majority of the members of the family show specificity for the cleavage of bonds in peptides of at least six residues with hydrophobic amino acids in both the P1 and P1' positions [34]. The specificity sub-sites are formed by hydrophobic residues surrounding the catalytic Asp dyad, and by the residues in the flap-turn. Mutations in these regions correspond to the necessity of binding different kinds of substrates.

Some fungal aspartic proteases are able to activate trypsinogen by cleavage of a

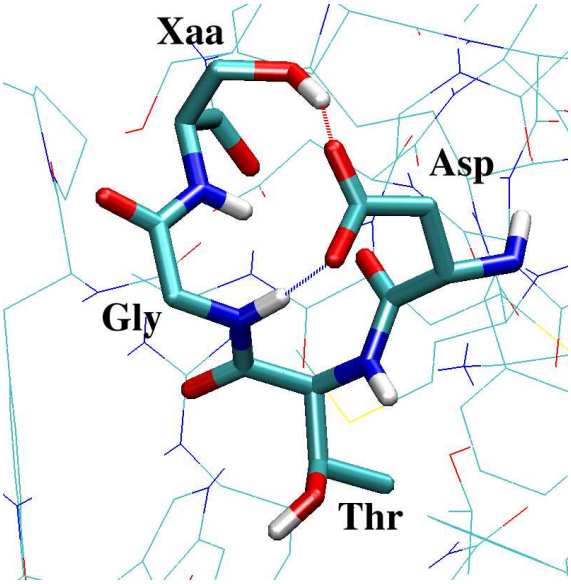


Figure 1.2: conserved active site loop in Aspartic proteases. The Xaa residue, in this case is a Ser, typical of the N-terminal lobe of pepsins.

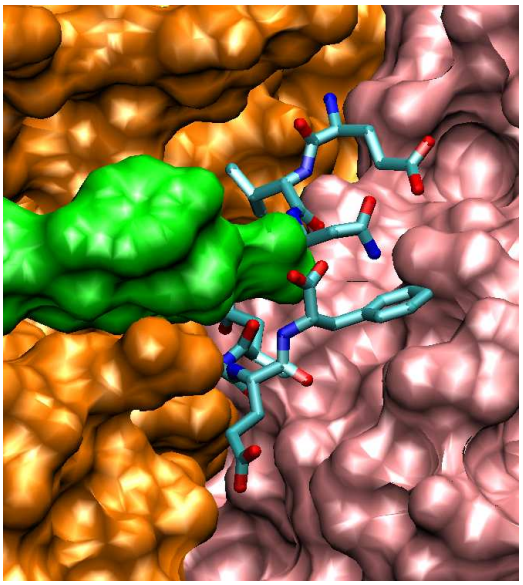


Figure 1.3: Example of ligand binding cleft of pepsins (human β-secretase, in particular). The surfaces of the two lobes are coloured in orange and rose, respectively, while the flap is coloured in green. The ligand is represented in cylinders.

Lys+Ile bond, showing as affinity for the cationic Lys side-chain. In fact, all family members able to activate trypsinogen show a conserved aspartate in the flap turn [35]. Yeast aspartic proteinase 3, instead, requires basic amino acids in both the S1 and S1' subsites.

Conserved Cysteine-Cysteine disulfide bridges are reported in the literature [4]. In fact, the increase of available sequences/structures has shown that a relatively large variability in their position and number is allowed. Three S-S bridges characterize the archetypal pepsin: one on the N-terminal lobe, following the catalytic Asp, and two on the C-terminal lobe, one preceding the second Asp, and the other before the C-terminus. The first loop is conserved in most of the family members, except for several fungal enzymes and human β-secretase. This loop is generally composed by five residues but larger loops in the same position are present in the database (eg. candidapepsin). The second loop is commonly 5-6 residues long, and it is present only in animal proteins. The third, and the largest, is conserved in all the enzymes but polyporopepsin, which does not contain any cystein residue. Other relevant exceptions are cathepsin D, which contains an additional cystein-cystein bond in the N-terminal lobe, and human β-secretase, which has no disulfide bridges in the N-terminal lobe, while it is characterized by three cysteine-cysteine loops in the C-terminal one (that is, one more than usual).

The fact that the two domains of pepsins have different disulfide bonding patterns indicates that loops have been being introduced after the internal duplication of the ancestral genome.

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 in lysosomes and endosomes. Renin catalyzes the conversion of angiotensinogen to angiotensin, a clinically important step in hemostasis. cathepsin D has been implicated in the metastasis of breast cancer and in Alzheimer's disease. Very recently, two new human aspartic proteases, memapsin 1 and 2, have been cloned [36]. Most importantly, memapsin 2 has been identified as β-secretase (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.

1.1.2 Retropepsin family

The retropepsin family is typically expressed in retroviruses. The out-coming of the structures of the first retropepins [37] have illuminated the evolutionary relationship

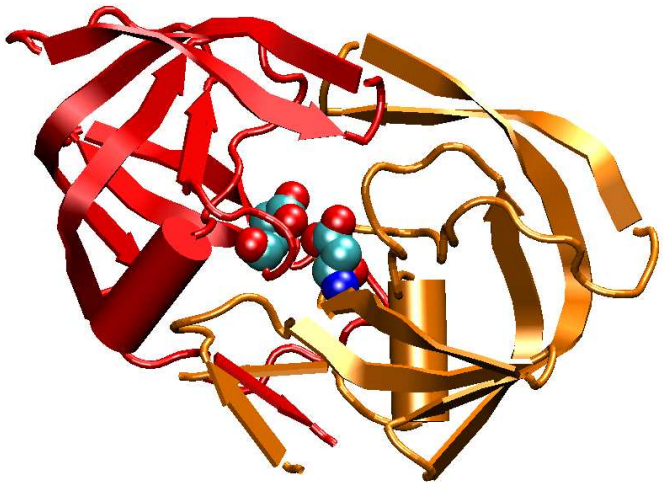


Figure 1.4: 3-D structure of HIV-1 protease. The two symmetric sub-units are coloured in orange and red, the catalyitc aspartic acids are drawn in balls.

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 fig. 1.4). The hypothesis of the existence of such a fold for an AP family by Tang *et. al.* [31] has been one of the greatest examples of the predictive power of considerations based on evolution in structural biology.

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. 1.5). 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.

The active site region of retropepsins is very similar to that of pepsins. In particular the already discussed Asp-Thr-Gly-Ala motif is present in the cleavage loops. The

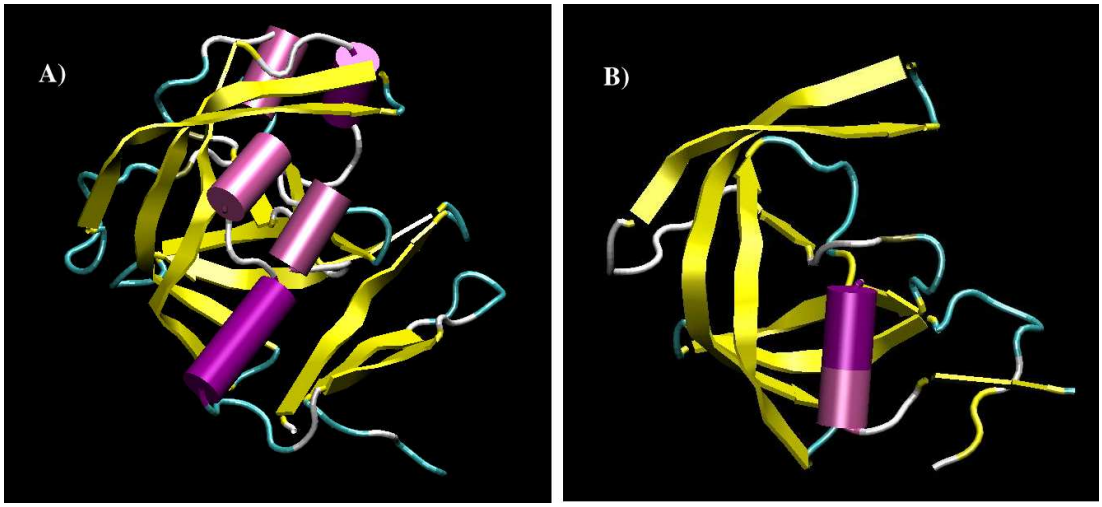


Figure 1.5: Comparison of the N-terminal lobe of a pepsin (A) and a monomeric unit of a retropepsin (B). α-helices are coloured in magenta, β-sheets in yellow.

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 [38]. These show unusual specificity in their ability to cleave a peptide bond of polyprotein substrates containing a proline in P_1' [39]. 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'; among all, the Met-Met sequence has been found particularly effective [38].

1.1.3 Mechanisms of enzymatic catalysis

Aspartic proteases catalyze hydroysis of peptide bonds by activating the nuclophilic attack of a water molecule to the carbonyl carbon [10]. Taking the different 3D structures of pepsins complexed with pepstatin as models for the tetrahedral intermediate, different mechanisms have been proposed [27]. In fact, the observation by Bott *et al.* [40] and James *et al.* [41] on the structures of the active sites of rhizopuspepsin and penicillopepsin bound to pepstatin showed that the hydroxyl group of the inhibitor was bound to the aspartic dyad, replacing the position of the water molecule seen in the native enzyme. However, the real structure of the transition state should be different, as another hydroxyl group is present in the diol intermediate.

It is agreed that the aspartic dyad acts as a base as to deprotonate water during

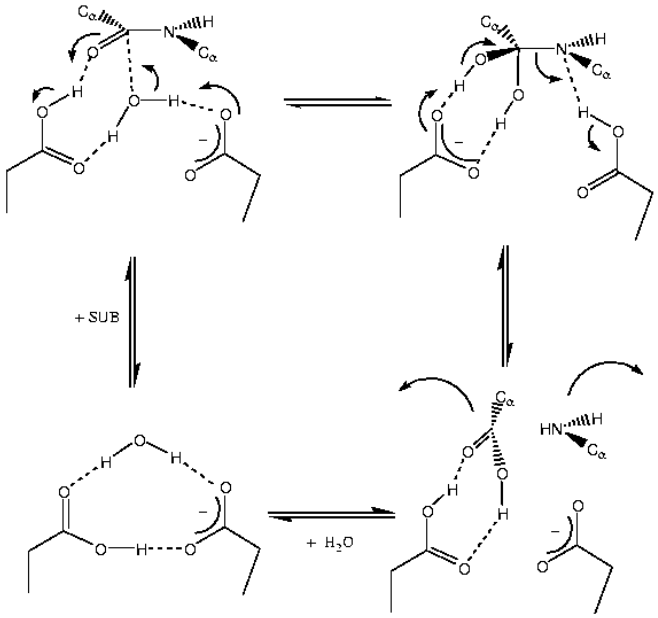


Figure 1.6: General reaction mechanism for catalysis of AP. Starting from the lower-left angle and following the reaction clockwise: Aspartic dyad in the free form; binding of the substrate and nucleophilic attack of water; formation of the tetrahedral *gem*-diol intermediate; protonation of the Nitrogen atom and formation of the products; release of the products and regeneration of the catalyst.

the activation process [10, 42, 43]. The microscopic details of this mechanism, on the contrary, are still under debate. James *et al.* [43] and Veerapandian *et al.* [42] 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. 1.6). 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 [10], which postulates the low-barrier hydrogen bond between the two aspartic acids, whose existence has been first predicted in our lab [44], as the key feature of the enzymatic action. Within this mechanism, the catalyst is not reformed at the product release state, and therefore an isomerization step is required (see fig. 1.7). The suggestion of an isomechanism in

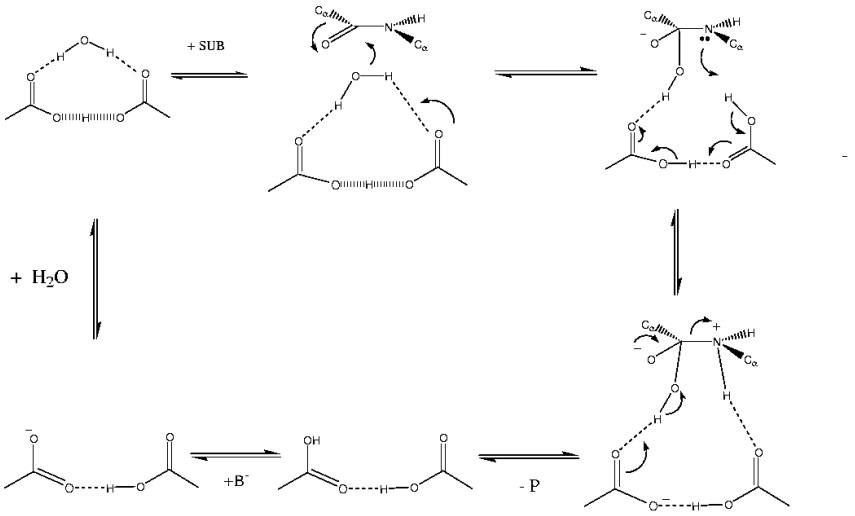


Figure 1.7: Reaction Mechanism as proposed by Nothrop. All the process occurs in the presence of a low-barrier hydrogen bond shared by the Asp dyad.

which the enzyme finishes one catalytic cycle in the wrong ionic state and must lose a proton does help to explain some kinetic and isotope data in the retroviral enzymes, as those reported by Meek *et al.* [45] for HIV-1 protease. The mechanism proposed by Northrop [10], which 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.

A role in the stabilzation of the asymmetry in the protonation state of the two aspartates might be played by a buried water molecule, which is H-bonded to the Ser of the N-terminal lobe catalytic loop, as proposed by Adreeva and Rumsh [46].

The study by Marcinkeviciene *et al.* [47] on inhibitor binding to porcine pepsin provides support for the concept of an isomechanism where different conformational isomers or prototropic forms are in equilibrium. The structure of plasmepsin 2 from *Plasmodium falciparum* [48] has revealed a significant rotation of the C-terminal domain with respect to the N-terminal one (5.2°). This finding 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.* [5, 6] have put into direct correlation, in HIV-1 protease, such flexibility of the enzymatic scaffold to its catalytic activity, at least for HIV-1 protease. Recent simulations by McCammon *et al.* [49] have confirmed the flexibility properties of such enzyme. 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 should enhance the catalytic rate by favouring the presence of reactive conformations in the active site along time. In this thesis, some theoretical investigations that lead to promising results on the subject will be presented.

1.2 Biological function of β-secretase and Alzheimer's Disease development

1.2.1 Historical perspective

Few subjects in biomedicine have aroused the interest of the scientific community as has had Alzheimer's disease (AD). Neurodegenerative disorders have become a common problem in modern societies, where life expectation, in the last century, has been constantly rising up. Among them Alzheimer's disease has emerged as the most prevalent form of late-life mental failure in humans [16].

Although neuropathological studies led quite soon to recognition of the commonness of the syndrome, it was only after the '60s that the first descriptions of the lesions in the brain tissues, namely, senile (neuritic) plaques and neurofibrillary tangles, related to AD came out.

In the late '70s and early '80s it became increasingly clear that AD, unlike Parkinson's disease, did not involve degeneration of a single transmitter class of neurons, but was highly heterogeneous. Advances in biochemical pathology, which lead to definition of the composition of plaques and tangles, and advances in molecular genetics of AD, which defined the critical role of the subunit proteins in the fundamental mechanisms of AD, have lead to a growing consensus about how at least the familial forms of the disorder may begin. The result of this continuing work is that a rough temporal outline of the disease cascade has begun to emerge.

In particular, one of the key biochemical steps that lead to formation of neuritic plaques involve the enzymatic action of β-secretase, a member of the pepsin AP class.

As inhibition of this enzyme might have a relevant positive fall-out in a pharmacological therapy against at least some forms of AD, it is therefore interesting to concentrate some attention on its features.

1.2.2 Neuritic plaques and the amyloid β-protein

Composition of the Neuritic plaques

Neuritic plaques are microscopic foci of extra-cellular amyloid β-protein (Aβ) deposition, generally found in large numbers in the limbic and association cortices [50]. Aβ aggregates principally in a filamentous form. The neurites are marked by structural abnormalities, like enlarged lysosomes or increment in the number of micotondria. The time that it takes to develop such a neuritic plaques is not known, but these lesions probably evolve gradually over a substantial period of time, perhaps many months or years.

Origin of the amyloid β-protein: APP

Aβ is derived from its large precursor protein (APP) by sequential proteolytic cleavages. APP comprises a heterogeneous group of ubiquitously expressed polypeptides ranging from 110 to 140 kDa. APP is a single transmembrane polypeptide that is translocated into the endoplasmic reticulum via its signal peptide, and then post-translationally matured through the secretory pathway [16]. Both during and after its trafficking through the secretory pathway, APP can undergo different proteolytic cleavages to release secreted derivatives into vesicle lumens and the extracellular space (see fig. 1.8). The main proteolytic cleavage process occurs 12 aminacids before the single trans-membrane domain of APP (α-cleavage site). This processing results in the secretion of a large fragment, called α-APP$_s$, and in retention of an 83-aa long residue C-terminal fragment in the membrane (C83). Alternatively, some APP molecules are cleaved by β-secretase, which cuts the APP chain 16 residues before the α-cleavage site, and thus, generating a slightly smaller ectodomain derivative (β-APP$_s$), and retaining a 99-aa long residue in the membrane (C99). After both α- and β- processes, the fragment retained in the membrane is cleaved 40 or 42 amino acids after the β-cleavage site by another enzyme, called γ-secretase. This ulterior cleavage leads to secretion of a 24/26 aa long peptide fragment, called p3, or to A$\beta_{40/42}$, depending on whether the initial cut was performed by α- or β-secretase, respectively[3].

[3]Until 1992, it was assumed that Aβ generation was a pathological event, because the cleavage of the C99 fragment from γ-secretase activity appeared to occur in the middle of the trans-membrane

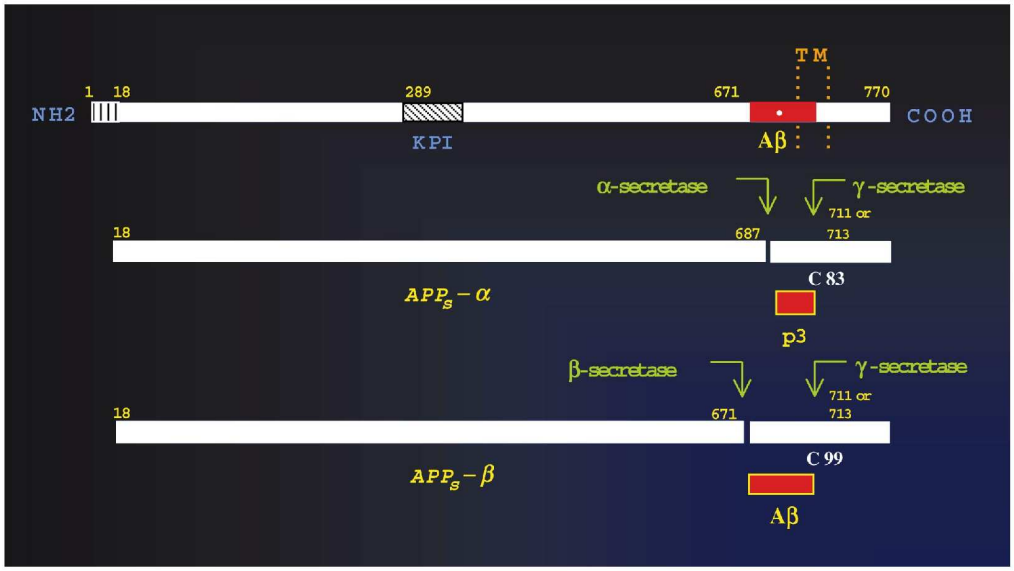


Figure 1.8: Schematic diagrams of the β-amyloid precursor protein and its principal metabolic derivatives. Top diagram depicts the 770 splice form of APP. The first 17 N-terminal amino acids work as a signal peptide. KPI indicates a serine protease inhibitor domain. TM indicates the trans-membrane domain at amino acids 700-723. The amyloid protein fragment is represented by the red band, the white dot is the cleavage site of α-secretase. The following diagrams represent the proteolyitc derivatives according to the α- or β- secretory paths.

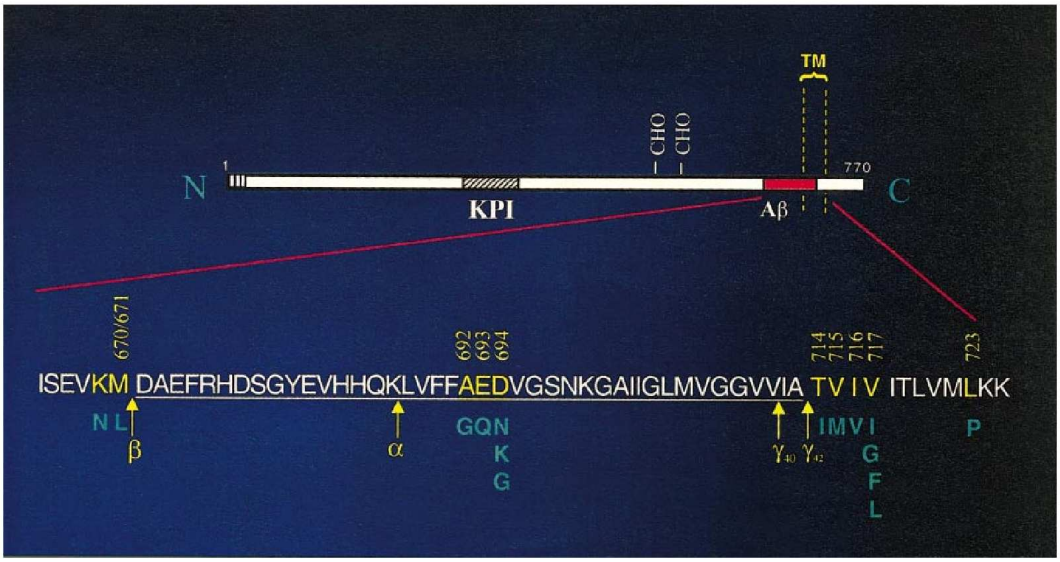


Figure 1.9: Mutations of APP identified with development of different familial AD. α-, β- and γ-secretase cleavage sites are explicitly indicated by the corresponding greek letters.

Familial Alzheimer's disease

It has been known for several decades that clinically typical AD can cluster in families and can specifically be inherited in an autosomal dominant fashion. Determining how frequently genetic factors underlie the disease is difficult, as AD was not specifically diagnosed and recorded before the last two decades; some investigators believe that in the fullness of time, a large majority of AD cases will be shown to have genetic determinants.

Missense mutations in APP Although mutations in APP that lead to AD have only been confirmed in some two dozen or so families worldwide, their locations at the cleavage sites of the three $\alpha-$, $\beta-$ and γ-secretases have provided critical insights into the mechanism of development AD, although in sporadic cases of AD other external factors, such as anti-nerve growth factor (NGF) antibodies [55, 56], may play

domain. It was assumed that this would require the release of C99 from the membrane, e.g., as a result of some pre-existing membrane injury that allowed access to a soluble protease. Discovery of p3 to be the product of sequential α- γ- cuts [52], and detection of Aβ in cerebrospinal fluid and plasma of healthy subjects throughout life [53, 54] demonstrated that, indeed, production of Aβ is a normal metabolic event.

a crucial role.

The nine known missense mutations (fig. 1.9) have been found to increase Aβ production by subtly different mechanisms. In particular, the so-called *Swedish mutation*, characterized by double mutation in the two amino acids (K670N, M671L) preceding β-secretase cleavage site, induces increased cleavage by β-secretase to generate aberrant production of Aβ proteins.

1.2.3 β-secretase: a key enzyme in AD development

Overproduction of Aβ and increase of its capability of self-aggregation are the crucial steps that lead to development of, at least, familial forms of AD. Some missense mutations (as the so-called) London mutations in APP lead to alteration of the relative production of the 40- and 42-isoforms of Aβ. The 42 aa long form is more prone to aggregation [57], however, Aβ_{40} is usually colocalized with Aβ_{42} in the plaques. Nonetheless, these mutations do not affect the total amount of Aβ formation [58]. These results have suggested that amino acids close to the γ-cleavage site are not critical to the total cleavage.

In fact, in healthy individuals the correct regulamentation of Aβ protein is guaranteed by the low proficiency of β-secretase catalytic action ($k_{cat}/K_M \approx 40$ s^{-1} M^{-1}). Swedish mutation acts on this factor dramatically, as in that case, the k_{cat}/K_M ratio increases up to 2450 s^{-1} M^{-1} [36]. Identification of β-secretase as the rate-limiting enzyme for Aβ production has meant identification of a highly promising pharmaceutical target for a future development of AD drug therapies.

The search for enzymes that specifically cleave at the β-cleavage site in APP was initiated long before there was any cellular evidence for the presence of such a metabolic pathway [59]. The search ended in 2000, when Tang *et al.* first, identified the extracellular C- domain the membrane-anchored memapsin 2 protein as the β-secretase enzyme [36], and then, provided also its first crystallographic structure [60].

1.2.4 Structural characterization of β-secretase

β-secretase (BACE hereafter) is an aspartic protease whose bilobal structure (fig. 1.10) has the general fold of pepsins (fig. 1.1). Compared to human pepsin, the most significant differences in the structure of BACE are related to six insertions and a 35-residue long extension in the C-terminus. The six insertions are all loops but the first (labelled A in fig. 1.11), which contains a short solvent-exposed α-helix. Four of these insertions (A,C,D and F in fig. 1.11) are located on the side of the molecular surface where

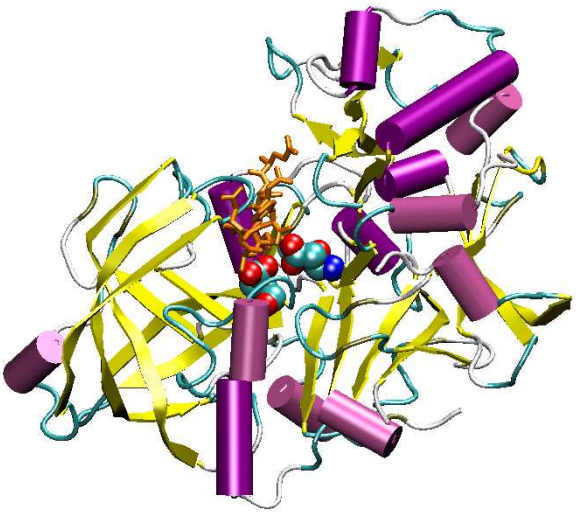


Figure 1.10: 3D picture of β-secretase. α-helices are drawn in magenta cylinders, β-sheets in yellow arrows, the ligand in orange sticks, and the Asp dyad in spheres.

the N-terminus of the substrate-chain is bound. One of these insertions (F) brings a high negative charge, as it contains four acidic residues. Insertions B and E are located in the opposite side of the protein surface (fig. 1.11). The C-terminal extension is highly structured, being composed by a loop, a β-sheet, another loop, and a short α-helix. As previously discussed (see sec. 1.1.1), BACE has differentiated from other pepsins in the position of two of its three disulfide bridges; in fact, they fasten the C-terminal extension to the C-terminal lobe (residues 155-359 and 217-382).

The active-site cleft of BACE is slightly more opened than that of the other pepsins of known structure. This characteristic is related to modifications in the S_4, S_2, and S_1' subsites, and to a deletion of six residues on a loop across the flap over the active-site cleft [60]. These modifications may be useful in designing of inhibitors selective for BACE towards other human aspartic proteases. Determination of the molecular transition-state geometry for the activated Michaelis Complex could also be a very important first-step for a rational design of different BACE inhibitors, as peptidomimetic inhibitors like the ones present in BACE 3D structures [60, 61] could not be pharmacologically relevant *in vivo*, as they were not enough small to be able to pass the blood-brain barrier.

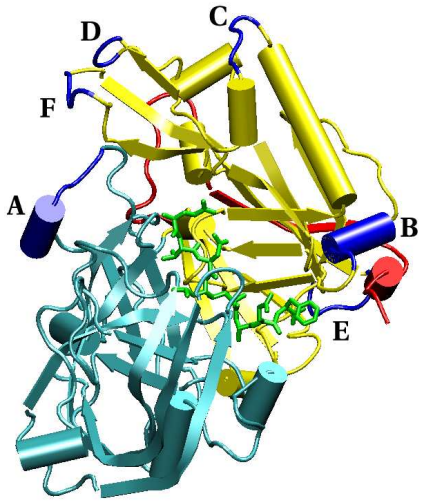


Figure 1.11: Insertions in BACE structure, as compared to pepsin. The N- and C- lobes are coloured in cyan and yellow, respectively, the six helices / loops insertions are coloured in blue, and labelled with capital letters (A-F), the C-terminal extension is coloured in red.

Chapter 2

Computational Methods

The constant increase along the years of computer performances in terms of both computational fastness and data storage/accessibility has lead computer simulations to be a powerful tool for the study of biologically relevant systems. The theory and the implemented algorithms used throughout this work are described below. Because of the intrinsic complexity of research targets in life sciences, all the developed methodologies require different degrees of approximation, which are also discussed. This work has taken advantage from a broad variety of computational tools, ranging from molecular dynamics (MD) simulations (both *ab initio* and *classical*) to coarse-grained computations, and to database-search and multiple-sequence alignment tools.

2.1 The Many-Body Problem and the Density-Functional-Theory approach

2.1.1 The many-body Hamiltonian

The equations of quantum mechanics are required to rigorously describe microscopic properties of molecular systems. In the non-relativistic formulation of quantum mechanics, a group of M atoms and N electron is described by its wave-function Ψ, a function of the 3M+3N geometric coordinates, the nuclear and electronic spin states, and of the time, which solves the time-dependent Schrödinger Equation [62]

$$i\hbar\frac{\partial}{\partial t}\Psi = \mathcal{H}\Psi \tag{2.1}$$

in the absence of external perturbations, and neglecting the spin couplings (which is reasonable for closed-shell systems), the Hamiltonian operator $\mathcal{H}$ is given by the

summation of the nuclear and electronic kinetic operators and the coulombic interaction among nuclei and electrons:

$$\mathcal{H} = -\sum_{i=1}^{M} \frac{\hbar^2 \nabla_i^2}{2m_i} \quad - \quad \sum_{j=1}^{N} \frac{\hbar^2 \nabla_j^2}{2m_e} - \sum_{i=1}^{M} \sum_{j=1}^{N} \frac{Z_i}{|\mathbf{R_i} - \mathbf{r_j}|} + \\ + \quad \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{1}{|\mathbf{r_i} - \mathbf{r_j}|} + \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \frac{Z_i Z_j}{|\mathbf{R_i} - \mathbf{R_j}|} \tag{2.2}$$

In this case, the time-dependent problem can be simplified into a time-independent formulation, which leads to the eigenvalue equation:

$$\mathcal{H}\psi = E\psi \tag{2.3}$$

where ψ now is an eigen-function of $\mathcal{H}$ and a function of only the 3N+3M coordinates.

2.1.2 Born-Oppenheimer approximation: the electronic structure problem

Analytical solutions to the former N+M bodies problem are not accessible, therefore, approximations to its original formulation have to be introduced.

Within the time-independent formulation, the most straightforward approximation consists into separating the nuclear degrees of freedom from the electronic ones (Born-Oppenheimer approximation) [63]. Here, the wave-function ψ is expressed by the following ansatz:

$$\psi(\mathbf{r}, \mathbf{R}) \approx \Phi(\mathbf{r}; \mathbf{R})\chi(\mathbf{R}) \tag{2.4}$$

Where χ is the nuclear wave-function, and Φ is the electronic Wave-function, which depends only parametrically on the nuclear coordinates. The electronic wave-function will solve the "fixed nuclei" electronic Schrödinger equation:

$$\mathcal{H}_e\Phi = E_e\Phi \tag{2.5}$$

where $\mathcal{H}_e$ is:

$$\mathcal{H}_e = T_e + V_{ee} + V_{en} \tag{2.6}$$

and $E_e = E_e(\mathrm{R})$.

The nuclear-coordinate dependent electronic energy is then included in the eigenvalue problem for the nuclear part[1]:

$$[T_n + V_{nn}]\chi = E - E_e(\mathbf{R})\chi \tag{2.7}$$

Thus, within the Born-Oppenheimer approximation, the solution of the N+M many-body problem is reduced, first, to the solution of the M electrons problem at fixed nuclear positions, and then to the solution of the N nuclei problem, in which the electrons contribute as an external field.

The solution of the N electrons problem itself, nonetheless, is a major task of quantum physics. Therefore, different theories and approximations have been developed to face this task. The work done in this thesis has got profit from Density Functional Theory [64], which allows treatment of relatively large systems with a reasonable computational cost.

2.1.3 Density Functional Theory

Density Functional Theory (DFT) is a rigorous theory of the ground state of many particle system [64]. The main idea lies in assumption that the ground-state properties of a quantum system of N particles can be described starting from its density and not from its explicit wave-function, thus, from a three-dimensional function $\rho(\mathrm{r})$, rather than a 3N dimensional one:

$$\rho(\mathbf{r}) = \mathrm{N} \int |\Psi(\mathbf{r_1}, \mathbf{r_2}, ...\mathbf{r_N}|^2 \, \mathrm{d}\mathbf{r_2}\mathrm{d}\mathbf{r_3}...\mathrm{d}\mathbf{r_N} \tag{2.8}$$

The Hohenberg-Kohn theorems

DFT is based on the two Hohenberg-Kohn theorems, enunciated in the early sixties [64]. The first theorem demonstrates that, given a Hamiltonian characterized by a general external potential V_{ext}, the ground-state density $\rho(\mathbf{r})$ associated to it is unique. As V_{ext} univocally determines the Hamiltonian of the system, it follows that the ground state energy and all the observables are functionals of the electron density ρ.

[1] neglecting the contributions of the nuclear kinetic operator over the electronic wave-function; this approximation is justified by the large difference in the electronic over nuclear mass ratio.

The second theorem provides a variational principle for the ground state density. In fact, it demonstrates that, given any trial density $\bar{\rho} > 0$ for which $\int \bar{\rho}(\mathbf{r})\mathrm{d}\mathbf{r} = \mathrm{N}$, it follows that $E[\bar{\rho}] \geq E[\rho]$.

From this result, one can get a variational equation to get the ground-state density. In fact, re-writing the electronic Schrödinger equation 2.5 in terms of the density:

$$E[\rho] = T[\rho] + E_{ee} + E_{en}[\rho] = \int \rho(\mathbf{r})\nu(\mathbf{r})\mathrm{d}\mathbf{r} + \mathrm{F_{HK}}[\rho] \tag{2.9}$$

where $\nu(\mathbf{r})$ is the external potential, and $F_{HK}[\rho] = \langle\Phi[\rho]|\mathcal{H}_e|\Phi[\rho]\rangle$ is the Hohenberg-Kohn functional. $\mathcal{H}_e$ includes the kinetic energy operator and the electron-electron repulsion functionals. Ground-state ρ fulfils the following stationary principle:

$$\delta\{E[\rho] - \mu[\int \rho(\mathbf{r})\mathrm{d}\mathbf{r} - \mathrm{N}]\} = 0, \tag{2.10}$$

and μ solves the following Euler-Lagrange equation:

$$\mu = \nu(\mathbf{r}) + \frac{\partial \mathrm{F_{HK}}[\rho]}{\partial \rho} \tag{2.11}$$

F_{HK} is not dependent on the external potential and, thus, can be assumed as a universal functional of the density. Although DFT is formally rigorous, the lack of knowledge of an analytical expression for this functional leads to necessary approximations to make it useful from a practical point of view.

The Kohn-Sham equations

The Hohenberg-Kohn theorems were re-formulated into a computationally accessible approach in 1965 by Kohn himself and Sham [65]. The main idea of that paper lies in the possibility of mapping a system of N interacting particles into a non-interacting ones, characterized by the same density.

For such systems, the density can be written as a summation over single-particle contributions:

$$\rho(\mathrm{r}) = \sum_{i=1}^{N} \left|\varphi_i^{KS}(\mathbf{r})\right|^2 \tag{2.12}$$

and the kinetic energy functional has an analytical expression:

$$T_s[\rho] = \sum_{i=1}^{N} \left\langle \varphi_i^{KS} \left| -\frac{1}{2}\nabla^2 \right| \varphi_i^{KS} \right\rangle \tag{2.13}$$

The Hohenberg-Kohn functional can be rewritten as:

$$F_{HK}[\rho] = T_s[\rho] + J[\rho] + E_{xc}[\rho] \tag{2.14}$$

where J is the classical part of the particle-particle interaction, and E_{xc} is defined as:

$$E_{xc}[\rho] = T[\rho] - T_s[\rho] + E_{ee}[\rho] - J[\rho] \tag{2.15}$$

and it is named 'Exchange-correlation' functional, as it is obtained by the sum of the correction in the kinetic energy and of the non-classical part of the particle-particle interaction.

Equation 2.11, turns out to be:

$$\mu = \nu_{eff}(\mathbf{r}) + \frac{\partial \mathrm{T_s}[\rho]}{\partial \rho} \tag{2.16}$$

with

$$\nu_{eff} = \nu(\mathbf{r}) + \int \frac{\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} \mathrm{d}\mathbf{r}' + \frac{\delta \mathrm{E_{xc}}[\rho]}{\delta \rho} \tag{2.17}$$

The total energy of the system is expressed as :

$$E = \sum_{i=1}^{N} \varepsilon_i - \frac{1}{2} \int \frac{\rho(\mathbf{r})\rho(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}' \mathrm{d}\mathbf{r} + \mathrm{E_{xc}}[\rho] - \int \frac{\delta \mathrm{E_{xc}}[\rho]}{\delta \rho} \mathrm{d}\mathbf{r} \tag{2.18}$$

where ε_i are the KS orbital energies of the non-interacting system, which are obtained solving self-consistently the Kohn-Sham equations:

$$\left[\frac{1}{2}\nabla^2 + \nu_{eff}(\mathbf{r})\right] \varphi_i = \varepsilon_i \varphi_i \tag{2.19}$$

The KS equations have the great privilege of being single-particle equations which, in principle, exactly collectively solve a many-body problem. Anyway, the lack of an explicit analytical formulation of the exchange-correlation functional leads to requirement of approximated forms for it.

Exchange-Correlation functionals

LDA approximation. A first successful approximation for the exchange-correlation functional has been proposed already in the original paper by Hohenberg and Kohn [64],

and it is based on the uniform electron gas properties. In fact, within this approximation, the electron density is thought to behave locally as a uniform electron gas (LDA: Local Density Approximation); thus, the exchange-correlation functionals reads:

$$E_{xc}^{LDA}[\rho] = -\frac{3}{2}\left(\frac{3}{4\pi}\right)^{\frac{1}{3}} \int \rho \frac{4}{3}(\mathbf{r})\mathrm{d}\mathbf{r} \tag{2.20}$$

The LDA approximation, although successful in the description of a broad variety of highly covalent systems, exhibits heavy deficiencies in describing hydrogen-bonds, which are crucial for studies on biologically relevant objects.

Gradient-Corrected approximations The typical approximation introduced to go beyond a LDA description is to explicitly introduce the gradient of the density in the functional form of E_{xc}. These correction are grouped under the name of "General Gradient-corrected approximations" (GGA).

Becke exchange functional This correction was introduced by Becke [66] to reproduce the exact asymptotic behaviour od the exchange energy. The analytical formula reads:

$$E_x[\rho] = E_{xc}^{LDA}[\rho] - \beta \int \rho^{\frac{4}{3}}(\mathbf{r}) \frac{x^2}{1 + 6\beta \sinh^{-1} x} d\mathbf{r} \tag{2.21}$$

with

$$x = \frac{|\nabla\rho|}{\rho^{\frac{4}{3}}} \tag{2.22}$$

the parameter β is determined by a fit on the exact HF data, and was fixed by Becke as β=0.0042 a.u.

Lee Yang Parr correlation functional The Lee-Yang-Parr (LYP) functional for the correlation energy was derived from the Colle-Salvetti formula [67], and computes correlation energies from HF second order density matrices. Its expression reads:

$$E_c[\rho] = -a \int \frac{1}{1 + d\rho^{-\frac{1}{3}}} \left\{ \rho + b\rho^{-\frac{2}{3}} \left[C_F \rho^{\frac{5}{3}} - 2t_W + \left(\frac{1}{9} t_W + \frac{1}{18} \nabla^2 \rho \right) \right] e^{-c\rho^{-\frac{1}{3}}} \right\} d\mathbf{r} \tag{2.23}$$

where $C_F = \frac{3}{10}(3\pi^2)^{\frac{2}{3}}$, $t_W = \frac{1}{8}\frac{|\nabla\rho|^2}{\rho} - \frac{1}{8}\nabla^2\rho$. a=0.04918, b=0.132, c=2533 and d=0.349 parameters are obtained by fitting the functional formula into HF results for the Helium atom.

The BLYP exchange-correlation functional, obtained by summing the Becke and LYP corrections, has been used in the DFT calculations during this work, as it has been shown to provide reasonable results for chemical bonds characteristic in organic and biological systems, without being too computationally expensive.

Plane waves and pseudopotentials

The use of a plane wave basis set is extremely convenient for solving the Kohn-Sham equations in periodic systems. Following the Bloch theorem, each KS wave-function can be expanded as:

$$\varphi_i^{\mathbf{k}}(\mathbf{r}) = \frac{1}{\sqrt{\mathrm{V}}}\mathrm{e}^{\mathrm{i}\mathbf{k}\cdot\,\mathbf{r}}\sum_{\mathbf{g}}\mathrm{c}_{\mathrm{j}}^{\mathbf{k}}(\mathbf{g})\mathrm{e}^{\mathrm{i}\mathbf{g}\cdot\,\mathbf{r}} \tag{2.24}$$

where V is the volume of the cell, **k** vectors belong to the first Brillouin zone, g is a reciprocal lattice vector, c is the first Fourier component of the plane waves expansion, and the summation is extended to infinite lattice vectors. This approach offers the advantage that results may be systematically converged with respect to the basis set by varying a single parameter, namely, the cut-off energy E_{cut} for the kinetic energy T_{PW} that determines the accuracy of the calculation. Moreover, the use of basis set with a non localized origin provides that the incomplete-basis set corrections (or "Pulay forces") are analytically annihilated, and there is no basis set superposition error.

Fast-Fourier-Transform (FFT) techniques allow efficient handling of the computational efforts for increasingly larger systems.

In the treatment of isolated clusters with a low symmetry, such as, organic molecules or the active sites of enzymes, the Γ-point approximation (k=0) still guarantees a good accuracy, leading to a relevant reduction of the computational cost. The simulation of isolated clusters within a periodic boundary condition scheme needs some care, as self-interaction among replicas has to be cancelled. In our calculations, we have used the procedure developed by Martyna and Tuckerman [68] that allows treating the system as isolated. This algorithm is based on the general possibility of writing an electrostatic potential Φ as the sum of two arbitrary functions:

$$\Phi(\mathbf{r}) = \phi^{long}(\mathbf{r}) + \phi^{short}(\mathbf{r})$$

These functions are defined so that $\phi^{short}(\mathbf{r})$ vanishes exponentially quickly at large distances from the system, while $\phi^{long}(\mathbf{r})$ contains all the long-range components of the physical potential. It can be demonstrated that the average potential energy for a cluster can be written in the reciprocal space as:

$$\langle \Phi \rangle = \frac{1}{2V} \sum_{\mathbf{g}} |\bar{\rho}(\mathbf{g})|^2 \left[\tilde{\Phi}(-\mathbf{g}) + \hat{\phi}^{screen}(-\mathbf{g}) \right] \tag{2.25}$$

Where $\tilde{\Phi}$ is the Fourier transform of the potential function, and $\bar{\rho}$ is the Fourier series expansion of the density. The screen function $\hat{\phi}^{screen}$, defined as:

$$\hat{\phi}^{screen}(\mathbf{g}) = \bar{\phi}^{long}(\mathbf{g}) - \tilde{\phi}^{long}(\mathbf{g}) \tag{2.26}$$

is the difference between the Fourier series $\bar{\phi}^{long}(\mathbf{g})$ and the Fourier transform $\tilde{\phi}^{long}(\mathbf{g})$ of the long-range potential, and its computation in the reciprocal space is efficient for all $\mathbf{g}$ vectors ($\mathcal{O}$ $(N \log N)$). The screen function has the meaning of "screening" the interaction of the system with an infinite array of periodic images, thus canceling the self-interaction among replicas of the system.

Pseudopotentials The greatest drawback in using a plane-wave basis-set comes from the impossibility, from a practical point of view, of describing core electrons within a reasonable computational expense. Indeed, the sharp spatial oscillations of their wave-functions nearby the nuclei would require an extremely high number of plane-waves for an accurate characterization. On the other hand, the core levels are well separated in energy from valence electrons, and, at a first level of approximation, do not play any role in the chemical properties of molecular systems. Thus, the core electron orbitals can be frozen in the KS equations and only the valence electrons are described explicitly. The core-valence electron interactions are implicitly included into the nuclear potential, which will assume an "effective-potential" or "pseudopotential" form.

Pseudopotential are usually derived from all electron (AE) atomic calculations, and several recipes have been proposed to date. In this work, we have used "norm-conserving" pseudopotentials derived from the Martins-Troullier method [69].

Pseudopotentials have to satisfy the following conditions:

- The valence pseudo-wave-function should not contain any radial nodes.
- The valence AE and pseudopotential eigen-values from the radial KS equations must be the same:

$$\varepsilon_\ell^{PP} = \varepsilon_\ell^{AE}$$

where ℓ is the angular momentum.

- The pseudo and AE atomic radial wave-functions must be equal for r greater than a chosen cut-off distance r_{cut}.

  These conditions ensure that the pseudo-atom behaves like the real one in the region of interaction with other atoms while forming chemical bonds.

- the integrated electron density within the cut-off radius for the two wave-functions must be the same.

  This requirement guarantees the transferability and the norm conserving rule of the MT pseudopotential.

- at $r = r_{cut}$, the pseudo wave-function and its first four derivatives should be continuous

- the pseudopotentials should have zero curvature at the origin.

Considering these conditions, the general form for a pseudopotential wave-function is:

$$\varphi_\ell^{PP}(r) = \begin{cases} \varphi_\ell^{AE}(r)\ ;\ r > r_{cut} \\ r^\ell e^{p(r)};\ r \leq r_{cut} \end{cases} \tag{2.27}$$

where $p(r) = c_0 + \sum_{i=1}^{6} c_i r^{2i}$, and the coefficients are obtained by imposing the first three conditions.

The functional form of the pseudopotentials is

$$V_{pseudo} = V_{val}(r) + \sum_{m,l} |Y_{l,m}\rangle\, V_l(r)\, \langle Y_{l,m}|$$

where $|Y_{l,m}\rangle$ are spherical harmonics. The "semilocality" of this functional form (local in the radial coordinate, non local in the angular ones), implies an increase in the computational cost. This difficulty is overcome in our calculation by the method of Kleinman-Bylander [70], which implies addition and subtraction of an "ad-hoc" radial function $\mathrm{V}_L(r)$ to the pseudopotential, which leads to a new functional form, where the local and non-local parts can be completely separated.

2.2 Car-Parrinello molecular dynamics

DFT, plane waves and pseudopotentials allow computation of the electronic structure of large enough biologically relevant molecular systems. Anyway, as the processes of life involve dynamical transformations that occur at finite temperature, it is crucial, in principle, to solve eq. 2.1, which explicitly takes into account the time variable, and, moreover, to somehow introduce temperature effects in the time-evolution.

2.2.1 nuclear dynamics: principal approximations

TD-SCF equations An approximate solution to eq. 2.1 can be written introducing an explicit time dependence into the time independent ansatz 2.4 for eq. 2.3:

$$\Psi(\mathbf{r},\mathbf{R},t) \approx \Phi(\mathbf{r};\mathbf{R},t)\chi(\mathbf{R};t)\exp\left[\frac{i}{\hbar}\int_{t_0}^{t} E_e(t')dt'\right] \tag{2.28}$$

where the nuclear and electronic wavefunctions are separately normalized at every time, respectively.

This ansatz, when inserted into eq. 2.1, leads to the following relations:

$$i\hbar\frac{\partial\Phi}{\partial t} = -\sum_i \frac{\hbar^2}{2m_e}\nabla_i^2\Phi + \langle\chi|V_{en}|\chi\rangle\Phi \tag{2.29}$$

$$i\hbar\frac{\partial\chi}{\partial t} = -\sum_i \frac{\hbar^2}{2m_i}\nabla_i^2\chi + \langle\Phi|\mathcal{H}_e|\Phi\rangle\chi \tag{2.30}$$

These equations define the basis of the time-dependent self-consistent field (TD-SCF) method introduced by Dirac in 1930. Both electrons and nuclei move quantum-mechanically in time-dependent mean-field potentials obtained from expectation values of the other degrees of freedom.

Ehrenfest dynamics It can be shown [71] that the nuclear equations of motion (EoM), which, in principle, follow a quantum-mechanical evolution, in the classical limit ($\hbar \to 0$) can be reduced to standard newtonian equations:

$$M_i\ddot{\mathbf{R}}(t) = -\nabla_i V_e^E(\mathbf{R}) \tag{2.31}$$

where $V_e^E(\mathbf{R}) = \langle\Phi|\mathcal{H}_e|\Phi\rangle$. Thus, the nuclei move according to classical mechanics in an effective potential generated by the electronic configuration at a given time t. The nuclear wave-functions in eq. 2.29, in the classical limit, are replaced by

delta functions centered at the instantaneous positions of the classical nuclei. Eq. 2.29 then, reads:

$$i\hbar\frac{\partial\Phi}{\partial t} = -\sum_i \frac{\hbar^2}{2m_e}\nabla_i^2\Phi + V_{ne}\Phi = \mathcal{H}_e\Phi \tag{2.32}$$

Thus, the quantum M+N bodies problem is again reduced to the N electrons quantum problem and to its time-evolution.

The equations proposed come from a mean-field theory, nonetheless, since explicit time-dependence in the electronic wave-function is kept, transitions between electronic states are accessible during time-evolution.

Born-Oppenheimer dynamics The major back-draw in the Ehrenfest dynamics scheme lies in its implementation. In fact, a numeric integration of the EoM for an electronic wave-function would require an extremely small time-step.

Thus, an alternative approach is to introduce a "ground-state" approximation for the electronic wave-function: in this scheme, the main assumption is that if the electronic excited stated are well separated in energy from the ground state, then, electronic excitations are extremely improbable. Thus, electrons will always remain in the ground state. This assumption leads to the need to solve *time-independent* electronic problems at each time, provided a set of *fixed* nuclear positions is given.

$$\mathcal{H}_e(\mathbf{r};\mathbf{R},t)\Phi = E_0\Phi \tag{2.33}$$

$$m_i\ddot{\mathbf{R}}_i(t) = -\nabla_i \min_{\Phi}\{\langle\Phi|\mathcal{H}_e|\Phi\rangle\} \tag{2.34}$$

In practice, starting from a nuclear configuration, one has to solve the time-independent Schrödinger equation for the electrons, move the nuclei within the electronic effective potential and iterate the process.

The greatest advantage in this scheme is that the nuclei move on the "frozen" time-independent Born-Oppenheimer electronic energy surface, and thus, the time-step for the numerical integration of the EoM can be chosen accordingly only to the nuclear degrees of freedom. On the contrary, the largest bottleneck comes from the necessity of solving the electronic Hamiltonian at each step. The occurrence, in both dynamical schemes, of serious bottlenecks (e.g., the small time-step in Ehrenfest dynamics, and the electronic Hamiltonian diagonalization at each step in BO dynamics) has lead to the need of developing other schemes that could overcome these limitations.

2.2.2 CP lagrangian

The Car-Parrinello approach [11] exploits the quantum-mechanical adiabatic time-scale separation of fast electronic and slow nuclear motion by transforming that into classical-mechanical adiabatic energy-scale separation in the framework of dynamical systems theory. The two-component quantum / classical problem is mapped onto a purely classical problem with two separate energy scales at the expense of losing the explicit time-dependence of the quantum subsystem dynamics.

In classical mechanics the force on the nuclei is obtained from the derivative of a Lagrangian with respect to the nuclear positions. This suggests that, if the quantity $\langle\Phi|\mathcal{H}_e|\Phi\rangle$ is considered as a functional of the wave-function, a functional derivative with respect to the orbitals, interpreted as classical fields, might yield the force on the orbitals, provided the correct Lagrangian is written.

The Lagrangian proposed by Car and Parrinello takes the following expression [11]:

$$\mathfrak{L}_{\mathrm{CP}} = \sum_i \frac{1}{2} m_i \dot{\mathbf{R}}_i^2 + \sum_i \frac{1}{2}\mu\langle\dot{\varphi}_i|\dot{\varphi}_i\rangle - \langle\Phi_O|\mathcal{H}_e|\Phi_0\rangle + constraints \tag{2.35}$$

Where μ is a fictitious "electron mass" associated to the electronic orbital, and the constraints refer to imposition of orthonormality of the electronic orbitals.

2.2.3 CP equations of motion

Following the Euler-Lagrange equations, CP equations of motion take the form:

$$m_i\ddot{\mathbf{R}}_i(t) = -\frac{\partial}{\partial\mathbf{R}_i}\langle\Phi_0|\mathcal{H}_e|\Phi_0\rangle + \frac{\partial}{\partial\mathbf{R}_i}\{constraints\} \tag{2.36}$$

$$\mu_i\ddot{\varphi}_i(t) = -\frac{\delta}{\delta\varphi_i^*}\langle\Phi_0|\mathcal{H}_e|\Phi_0\rangle + \frac{\delta}{\delta\varphi_i^*}\{constraints\} \tag{2.37}$$

According to the CP equations of motion, the nuclei evolve at a certain instantaneous "physical" temperature proportional to the nuclear kinetic energy, while the electronic degrees of freedom evolve in time at a certain "fictitious" temperature $\propto \sum_i \mu\langle\dot{\varphi}_i|\dot{\varphi}_i\rangle$. Then, if the starting electronic configuration is optimized to the Born-Oppenheimer surface, during its time evolution it will remain close to it, provided the adiabatic hypothesis is verified.

A simple way of monitoring whether the adiabatic conditions are fulfilled is to consider the KS orbital dynamics as a superposition of harmonic modes with frequency proportional to:

$$\omega_{ij} = \left[2 \, \frac{\varepsilon_i - \varepsilon_j}{\mu} \right]^{\frac{1}{2}}$$

thus, the lowest frequency of the electronic system is $\sqrt{2\frac{E_{gap}}{\mu}}$. Thus, adiabaticity can achieved with a suitable choice of the fictitious mass μ[2]. For insulators, such as all the system treated in this work, where separation in energy between the HOMO and LUMO levels is large, it turns out that the electronic and ionic frequencies are well separated, and the Car-Parrinello dynamics can be productively performed.

2.2.4 Nosé thermostats

The CP equations of motion, as they are written in eqs. 2.36-2.37, map a micro-canonical ensemble[3], as the total energy is strictly conserved during time evolution. Constant temperature dynamics can be achieved by coupling the ionic degrees of freedom to a Nosé-Hoover thermal bath [72]. Within this scheme, a dynamical friction coefficient is added to the CP lagrangian:

$$T_{nosé} = \frac{1}{2} Q \dot{s}^2 \tag{2.38}$$

$$V_{nosé} = (F+1) k_B T \ln s \tag{2.39}$$

where Q is a parameter, s is the additional degree of freedom of the thermal bath, and F is the total number of the degrees of freedom of the system. The equations of motions for the system take the form:

$$\ddot{\mathbf{R}}_i = \frac{1}{m_i s^2} \mathbf{f}_i - 2\dot{s} \frac{\dot{\mathbf{R}}_i}{s} \tag{2.40}$$

$$Q\ddot{s} = \sum_i m_i \dot{\mathbf{R}}_i^2 s - (F+1) k_B \frac{T}{s} \tag{2.41}$$

Q can be arbitrarily chosen to determine the strength of the coupling; high values result into a low coupling and vice-versa. Although the dynamics is strictly "non-Hamiltonian", the friction coefficient has the effect of accelerating/decelerating the system, keeping the energy around the KT value. As a result, the total energy is

[2] It has to be mentioned that the largest applicable time-step is proportional to the inverse of $\sqrt{\mu}$, thus, an excessive increase of its value would not be recommendable

[3] assuming that ergodic hypotheses are verified

conserved, and it can be shown that the dynamics maps correctly the canonical ensemble, provided the system is ergodic. It has also been shown that the correct sampling of the canonical ensemble can be obtained also for non-ergodic systems, such as the harmonic oscillator, by adding a series of coupled thermostats, rather than only one [73, 74]. This more refined scheme (called "Nosé-Hoover chains") is, actually, the one used in our simulations.

2.2.5 Integration of the nuclear EoM

Analytical integration of the nuclear EoM for real systems is unaffordable; therefore, numerical schemes have to be implemented:

Verlet algorithm The atomic coordinates at a time $t + dt$ can be written as a Tayor series expansion around a time t:

$$R(t + dt) = R(t) + \dot{R}(t)dt + \frac{1}{2}\ddot{t}dt^2 + \mathcal{O}(dt^3) \tag{2.42}$$

Summing or subtracting the expansions of $R(t + dt)$ and $R(t - dt)$, one gets the Verlet algorithm:

$$R(t + dt) = 2R(t) - R(T - dt) + \frac{1}{2}\ddot{R}(t)dt^2 + \mathcal{O}(dt^4) \tag{2.43}$$

$$\ddot{R}(t) = \frac{R(t + dt) - R(t - dt)}{2dt} + \mathcal{O}(dt^2) \tag{2.44}$$

which needs positions at time $t - dt$ and t, and accelerations at time t to obtain the new positions at time t+dt, within an accuracy of order $\mathcal{O}(dt^4)$ and velocities at time t within an accuracy of $\mathcal{O}(dt^2)$. The main advantages of this numerical algorithms lie in its straightforwardness and modest storage requirements, although the velocity calculations are not very accurate.

Leap-frog algorithm A better precision in the velocity estimates can be achieved by the so-called leap-frog algorithm, used in our dynamics simulations. In this scheme, first, the velocities are calculated at half-integer time steps $t + \frac{1}{2}dt$, then, these are used to compute the positions:

$$v(t + \frac{1}{2}dt) \approx v\left(t - \frac{1}{2}dt\right) + \ddot{R}(t)dt \tag{2.45}$$

$$R(t+dt) \approx R(t) + v\left(t + \frac{1}{2}dt\right)dt \tag{2.46}$$

The advantage of this algorithm is that the velocities are explicitly calculated and used to obtain the positions, the disadvantage is that they are not calculated at the same time as the positions. [4]

2.2.6 Long-range interactions evaluation

Car-Parrinello MD is performed under periodic boundary conditions. Within this condition, each particle, in principle, interacts with all the other N-1 particles into the simulation box and with all the N particle images in an infinite 3D array of periodic cells. The electrostatic potential energy V of the infinite system, therefore, takes the form:

$$E = \frac{1}{2}\sum_{|\mathbf{n}=0|}^{\infty}\sum_{i=1}^{N}\sum_{j=1}^{N} E(d_{ij}+nL) \tag{2.47}$$

where L is the length of the periodic box, and **n** are the direct lattice vectors. Convergence of this series is extremely slow, and calculating E from this expression is inefficient.

Simplification of the electrostatic calculations is achieved by splitting the energy functions in two parts: a short-range term (involving charges within a sphere with $r < r_c < L/2$, that is, within the primitive cell) and a long-range term (for atoms that are further than r_c).

Ewald summation techniques [76, 77] are of great help in performing this computation more efficiently: at each charge position $\mathbf{R}_i$, a gaussian charge distribution of opposite total charge -q_i is added (ϱ_i^s) and subtracted (ϱ_i^G), so that the total point charge distribution of the system $\varrho(\mathbf{R})$ may be wrtitten as:

$$\begin{aligned}\varrho(\mathbf{R}) &= \varrho^s(\mathbf{R}) + \varrho^G(\mathbf{R}) - \varrho(\mathbf{R}) = \\ &= \sum_{i=1}^{N} q_i\delta(\mathbf{R}-\mathbf{R}_i) + \sum_{i=1}^{N} q_i\left(\frac{\alpha}{\sqrt{\pi}}\right)^3 e^{-\alpha^2(\mathbf{R}-\mathbf{R}_i)} - \sum_{i=1}^{N} q_i\left(\frac{\alpha}{\sqrt{\pi}}\right)^3 e^{-\alpha^2(\mathbf{R}-\mathbf{R}_i)}\end{aligned} \tag{2.48}$$

At this point, $\varrho + \varrho^s$ and ϱ^G contributions to the total charge distribution can be considered separately. From the first term one gets the following potential:

[4] All the DFT and CP calculations are performed with the CPMD code developed by Hütter *et al.* [75].

$$V^S = \frac{1}{4\pi\varepsilon_0}\frac{1}{2}\sum_{|\mathbf{n}=0|}^{\infty}\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{q_i q_j \mathfrak{f}(\alpha|\mathbf{r}_{ij}+\mathbf{n}|)}{|\mathbf{r}_{ij}+\mathbf{n}|} \tag{2.49}$$

where $\mathfrak{f}$ is the complementary error function: $\mathfrak{f}(x) = \frac{2}{\pi}\int_x^\infty e^{-t^2}dt$. Computation of V^s in real space is now easier, as the opposite charged distribution shields the point charges, giving a negligible long-range contribution.

The long range electrostatic potential contribution is therefore contained in the potential that comes from the second term of the charge density, which, now, can be more easily evaluated in the reciprocal space:

$$V^G = \frac{1}{4\pi\varepsilon_0}\frac{1}{2\pi L^3}\sum_{\mathbf{k}\neq 0}^{\infty}\sum_{i=1}^{N}\sum_{j=1}^{N} q_i q_j \frac{4\pi^2}{k^2} e^{-\frac{k^2}{4\alpha^2}}\cos(\mathbf{k}\cdot\mathbf{r}_{ij}) \tag{2.50}$$

This expression may be calculated with $\mathcal{O}(\mathrm{N}^{3/2})$ operations, and can be further improved in efficiency through approximate algorithms.

2.3 Classical Molecular Dynamics

The large computational costs in the implementation of *Ab-initio* molecular dynamics end up into two major limitations: in the relatively small size of the systems that can be simulated (hundreds of atoms), and in the relatively short simulation timescale (tenths of picoseconds). As relevant biological processes usually involve large systems (thousands of atoms or more), and occur in relatively long timescales (from nano to microseconds or more), it is necessary to develop effective parameterized potentials, which are faster to integrate, albeit less accurate, in order to study this kind of systems.

2.3.1 Empirical force-fields

Force-fields based simulations originate from the assumption that the Born-Oppenheimer potential energy surface can be approximately described by additive parameterized many-body terms that can be obtained by fitting experimental and high-level quantum chemical data into simple functional forms. In this work, the AMBER force field [78] for description of proteins in solution has been used, while the GROMACS package [79, 80] has been used to integrate the EoM. The functional form of this particular force-field can be written as:

$$
\begin{aligned}
E &= E_{stretch} + E_{bend} + E_{torsional} + E_{VdW} + E_{electrostatic} = \\
&= \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\vartheta(\vartheta - \vartheta_{eq})^2 + \\
&+ \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{(d_{ij} - d_{ij}^{eq})^{12}} - \frac{B_{ij}}{(d_{ij} - d_{ij}^{eq})^6} + \frac{q_i q_j}{4\pi\varepsilon_0 d_{ij}} \right]
\end{aligned}
\tag{2.51}
$$

- $E_{stretch} + Ebend$ represent bond stretching and angle bending as harmonic energy terms. The equilibrium bond lengths and angle values r_{eq}, ϑ_{eq}, as well as the spring constants K_r, K_ϑ are fitted to reproduce structural parameters and normal mode frequencies.

- $E_{torsional}$ is the Fourier decomposition of the torsional energy. Dihedral parameters are calibrated on small model compounds, comparing the energies with those obtained by quantum chemical calculations. Improper dihedral angles are inserted between non-covalently bonded atoms to preserve planarity in aromatic rings.

- E_{VdW} describes the Van der Waals interactions via a Lennard-Jones potential. Parameters act as to reproduce chemical-physical properties (e.g., densities, enthalpies of vaporization, solvation free-energies, etc) in organic liquids.

- $E_{electrostatic}$ is the electrostatic energy evaluated by assuming the dielectric constant equal to 1, and using the Restrained Electrostatic Potential partial atomic charges (RESP) [81] as model atomic point charges.

Van der Waals and electrostatic interactions are calculated between atoms belonging to different molecules or for atoms in the same molecules separated by at least three bonds.

Constraints

The time-step in the integration of the EoM is limited by the highest frequency motions, which, in the case of parameterized force fields, are the bond stretching vibrations (in particular, those involving hydrogen atoms). As these motions are of little interest in classical simulations of biological systems, constraining these bond lengths

during the simulation is of great help, as the time-step can be increased without affecting the accuracy of the simulations.

The set of N_c molecular constraints σ_k, one for each constrained bond, can be written as:

$$\sigma_k(\mathbf{R}_1, ..., \mathbf{R}_N) = 0 \qquad k = 1, ..., N_c \tag{2.52}$$

Satisfaction of the N_c constraints can be accomplished by applying the Lagrange multipliers method. An extra term is added to the potential energy function, and the equations of motions become:

$$m_i \ddot{\mathbf{R}}_i = -\nabla \left(E + \sum_{k=1}^{N_c} \lambda_k(t) \sigma_k(\mathbf{R}_N) \right) \qquad i = 1, ..., N \tag{2.53}$$

The LINCS algorithm [82], as implemented in the GROMACS program, allows satisfaction of the set of holonomic constraints by an efficient parallelizable matricial representation of the equations, and it has been used in classical MD calculations done in this work.

Long-range interactions evaluation

Classical MD is usually performed under periodic boundary conditions, in order to minimize boundary effects and to mimic the presence of the bulk.

The VdW interactions can be treated as short-range interactions, since their energy functions decay rapidly; in fact, although these potential functions are not rigorously zero for $r \geq r_c$ truncation at r_c in the calculation will result in a systematic error which may be corrected adding a tail contribution. For Coulombic interactions, instead, the tail correction diverges, as the electrostatic potential function goes slowly to zero as $r \to \infty$. Therefore, interactions with all periodic images must be explicitly considered. Ewald summations presented in sec. 2.2.6 can be efficiently used also in classical MD to solve this computational task.

In the GROMACS package, an efficient Particle mesh Ewald implementation is used, consisting of an interpolation of the reciprocal space potential on a grid using smooth functions. Therefore, the sum can be evaluated by using $\mathcal{O}(N \ln N)$ Fast Fourier Transform algorithms.

Sampling the NPT ensemble

Nuclear EoM as written in 2.2.4 allow MD trajectories to sample the NVT canonical ensemble. Nonetheless, thermodynamic properties of biological systems should be computed at constant pressure, rather than at constant volume conditions. It is, therefore, important to introduce in the EoM the Parrinello-Rahman algorithm [83, 84], which is similar to the Nosé-Hoover temperature coupling. With the Parrinello-Rahman barostat, the box vectors as represented by the matrix $\mathbf{b}$ obey the matrix equation of motion:

$$\ddot{\mathbf{b}} = V\mathbf{W}^{-1}\mathbf{b}'^{-1}(\mathbf{P} - \mathbf{P}_{ref}) \tag{2.54}$$

where V denotes the volume of the box, W is a matrix parameter that determines the strength of the coupling, and $\mathbf{P}$ and $\mathbf{P}_{ref}$ are the instantaneous and reference pressures, respectively. The EoM for the particles change as for the Nosé-Hoover coupling:

$$\ddot{\mathbf{R}}_i = \frac{\mathbf{f}_i}{m_i} - \mathbf{M}\dot{\mathbf{R}}_i \tag{2.55}$$

$$\mathbf{M} = \mathbf{b}^{-1}\left[\mathbf{b}\frac{d\mathbf{b}'}{dt} + \frac{d\mathbf{b}}{dt}\mathbf{b}'\right]\mathbf{b}'^{-1} \tag{2.56}$$

In the GROMACS implementation, the coupling matrix is determined providing the compressibility β and the pressure time-constant τ_p as input:

$$(\mathbf{W}^{-1})_{ij} = \frac{4\pi\beta_{ij}}{3\tau_p^2 L} \tag{2.57}$$

where L is the largest box matrix element.

In the GROMACS package, EoM, integrated *via* the leap-frog algorithm, can be coupled to both Nosé-Hoover and Parrinello-Rahman baths, thus, they can provide a correct sampling of the NPT ensemble.

2.3.2 Analysis of MD trajectories

Data from MD trajectories can be collected and used to calculate various structural properties, as reported here:

Root mean square displacement

The *root mean square displacement* (rmsd) of a set of N atoms at time t, with respect to the initial structure, reads:

$$rmsd(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \Delta\mathbf{R}_i(t)^2} \tag{2.58}$$

where $\Delta\mathbf{R}_i(t)$ is the displacement of the *i-th* atom at time t.

Root mean square fluctuation

The root mean square deviation can be computed with respect to the average positions during dynamics, rather than to the initial positions, and mediated over time. This other quantity is called *root mean square fluctuation* (rmsf) :

$$rmsf = \sqrt{\langle \Delta\mathbf{R}^2 \rangle} = \sqrt{\langle (\mathbf{R} - \langle \mathbf{R} \rangle)^2 \rangle} \tag{2.59}$$

where the brackets $\langle ... \rangle$ stand for a temporal average. This quantity can be compared to crystallographic B-factors through the Debye-Waller relation:

$$B = \frac{8\pi^2}{3} rmsf^2 \tag{2.60}$$

Large-scale motions

Description of the protein motion in terms of over-damped dynamics appears to be particularly valid for protein's low frequency vibrations, which are relevant in protein functional dynamics [85, 86, 87, 88]. *Large-scale* or *essential* motions derived from over-damped dynamics can be calculated on the basis of the C_α atoms covariance matrix $\mathcal{C}$:

$$\mathcal{C}_{ij} = \langle (X_i - \langle X_i \rangle)(X_j - \langle X_j \rangle) \rangle \tag{2.61}$$

where X_i, X_j indicate the instantaneous values of the 3N cartesian coordinates of the N C_αs. The symmetric matrix $\mathcal{C}$ can be diagonalized by an orthonormal coordinate transformation $\mathcal{Q}$.

$$\begin{aligned} R - \langle R \rangle &= \mathcal{Q} q \\ \mathcal{Q}\mathcal{C}\mathcal{Q}^T &= \Lambda = \langle q q^T \rangle \end{aligned} \tag{2.62}$$

which transforms $\mathcal{C}$ into a diagonal matrix of 3N eigenvalues λ_i. Provided a sufficient number of independent configurations is available (at least 3N+1), six of these eigen-values, representing the roto-translational motions will collapse to zero. The roto-translational degrees of freedom can be eliminated by performing a rmsd fit between each MD snapshot and the initial configuration.

In general, long-range motions of a protein are sufficiently described by only M $\ll$ 3N+1 eigenvectors corresponding to the largest M eigen-values. Indeed, it turns out that most of the total motion is usually spanned over the first 2-4 eigenvectors. The correct description of the protein dynamics in terms of essential motions depends strongly on the convergence of the covariance matrix $\mathcal{C}$. A useful way of checking the well definition and relevance of essential modes (that is, of checking whether the MD simulation is long enough) is performed comparing covariance matrices computed in different time-windows during the run, using the same reference structure [89, 90]. So, if $\mathcal{A}$ and $\mathcal{B}$ are the covariance matrices computed in the first and second half of the trajectory, we can define their "distance" as:

$$\begin{aligned} d(\mathcal{A},\mathcal{B}) &= \sqrt{tr\left[\left(\mathcal{A}^{\frac{1}{2}}-\mathcal{B}^{\frac{1}{2}}\right)^2\right]} = \\ &= \left[\sum_{i=1}^{N}\left(\lambda_i^{\mathcal{A}}+\lambda_i^{\mathcal{B}}\right)-2\sum_{i=1}^{N}\sum_{j=1}^{N}\sqrt{\lambda_i^{\mathcal{A}}\lambda_j^{\mathcal{B}}}\left(\mathcal{Q}_i^{\mathcal{A}}\cdot\mathcal{Q}_j^{\mathcal{B}}\right)^2\right]^{\frac{1}{2}} \end{aligned} \tag{2.63}$$

Overlap is finally defined as:

$$\mathcal{O}(\mathcal{A},\mathcal{B}) = 1-\frac{d(\mathcal{A},\mathcal{B})}{\sqrt{tr\mathcal{A}+tr\mathcal{B}}} \tag{2.64}$$

and it is 1 for identical matrices, while is 0 for completely orthogonal subspaces.

Another useful check is the cosine content. It has been proved that the principal components of random diffusion are cosines with the number of periods equal to half the principal component index [89, 90]. The eigen-values are proportional to the index to the power -2. The cosine content is defined as:

$$\frac{2}{T}\left[\int_0^T\cos(k\pi t)p_i(t)dt\right]^2\left[\int_0^T p_i^2(t)dt\right]^{-1} \tag{2.65}$$

If the cosine content is close to 1, that fluctuation is not connected with the potential, but with random diffusion.

2.4 Hybrid QM/MM molecular dynamics

Although very useful in overcoming the small-size/time-limit bottlenecks, parameterized potentials are not adequate for description of a variety of phenomena, for example, where the force-field transferability is low (e.g. in the presence of strong and/or varying electric fields) or during chemical reactions.

In those systems where the inadequacy of a parameterized description occur only for a small number of atoms, hybrid quantum mechanics-molecular mechanics methods allow to overcome the problem. In fact, in these computations, part of the system can be described at a quantum-chemical level, whereas the rest is treated by parameterized Hamiltonians (Warshel and Levitt, 1976; Singh and Kollman, 1986; Field *et al.*, 1990; Sherwood, 2000). The Hamiltonian for a hybrid system can be written as:

$$\mathcal{H} = \mathcal{H}_{QM} + \mathcal{H}_{MM} + \mathcal{H}_{QM/MM} \tag{2.66}$$

where $\mathcal{H}_{QM}, \mathcal{H}_{MM}$ are the Hamiltonians for the quantum and classical systems, respectively, and $\mathcal{H}_{QM/MM}$ contains the interactions between the two parts.

Mixed QM/MM approaches have been developed applying both semi-empirical and *ab initio* levels of theory. In this thesis, the method developed by U. Roethlisberger and co-workers has been used [91], in which the quantum part is treated at DFT/BLYP level (see section 2.1.3) and dynamics is performed following Car-Parrinello EoM, whereas the classical part is described by the AMBER force-field.

2.4.1 The Interface Hamiltonian

The crucial issue in developing a QM/MM method is in the definition of the $\mathcal{H}_{QM/MM}$ part of the Hamiltonian. The scheme adopted in this thesis is the one developed by Laio *et al.* [91], which will be briefly introduced below.

Bonded interactions

The bonded interactions at the interface (e.g. those among quantum and classical atoms directly connected by chemical bonds) are included in the classical force field. Stretch, bend and torsional terms are treated as in the classical MM Hamiltonian (see section 2.3.1). These terms allow the geometry at the interface to be kept stable, nonetheless, since their parameterization does not take into account changes due to chemical reactions, they might not be very accurate, if the boundary region gets crucial distortions upon chemical rearrangements. QM atoms that are directly linked to

MM atoms by chemical bonds are left with unsaturated valence orbitals. This problem can be overcome by saturating the valence with "capping" dummy-hydrogen atoms, which can fill the valences, but whose interactions with the MM part are not considered.

Non-bonded interactions

The electrostatic and steric contributions between non-bonded atoms are taken into account following:

$$\mathcal{H}_{QM/MM}^{non-bonded} = \sum_{i \in MM} q_i \int \frac{\varrho(\mathbf{r})}{|\mathbf{r} - \mathbf{R}_i|} d\mathbf{r} + \sum_{i \in MM} \sum_{j \in QM} V_{VdW}(\mathbf{d}_{ij}) \tag{2.67}$$

The electrostatic term takes into account the interaction of the classical point-charges with the total charge-density ϱ, which is the sum of the electronic density ρ and the nuclear point charges. Two major problems are related to calculation

of the electrostatic interactions: (i) the electron spill-out phenomenon and (ii) a very high computational cost. The overcoming of these two problems is achieved by treating in different ways short and long range interactions.

Short-range electrostatic interactions The spill-out problem derives from the anomalous rearrangement of the electron density that is attracted onto the MM point charges nearby the interface region. This problem is particularly pronounced as the wave-functions are expanded in a delocalized plane-wave basis set. A simple solution is given by modifying the Coulomb potential in the core region of MM atoms: whereas the 1/r behaviour is maintained for large r, for values of r shorter than the covalent radius, the Coulomb potential goes to a finite value. Thus, the electrostatic term is rewritten as:

$$\mathcal{H}_{QM/MM}^{el} = \sum_{i \in MM} q_i \int \varrho(\mathbf{r}) \frac{r_{c_i}^4 - r^4}{r_{c_i}^5 - r^5} d\mathbf{r} \tag{2.68}$$

where r_c are the covalent radii. This choice of the smoothing function has been tested to produce accurate results for the structural properties of a quantum water molecule in a box of classical water molecules without any *ad hoc* re-parametrization of the force field.

Long range electrostatic interactions The explicit calculation of the long-range electrostatic term is too expensive in a plane wave based approach. In fact, the calculation of about $N_r \cdot N_{MM}$ integrals, where N_r is the number of real-space grid points ($\approx 10^3$) and N_{MM} is the number of classical atoms ($\approx 10^5$),would be required. This problem is solved by lowering the degree of accuracy of the calculation as the MM atoms get further from the QM region. The MM region is partitioned into three regions by providing two cut-off radii r_1 and r_2.

- Interaction of the QM charge distribution with all MM atoms within r_1 is explicitly calculated.
- For MM atoms with $r_1 < r < r_2$, the Coulomb interaction is approximated to interaction between the MM atomic point charges and the D-RESP charges [15] of the QM system (see below for D-RESP charges derivation). Thus, the electrostatic Hamiltonian reads:

$$\mathcal{H}^{el}_{QM/MM} = \sum_{i,j}^{r_1 \leq r \leq r_2} \frac{q_i Q_j^{RESP}}{|R_i - R_j|} \tag{2.69}$$

- for $r > r_2$ The MM charges interact with a multi-polar expansion (up to the quadrupolar term) of the QM charge distribution. In this case, the Hamiltonian takes the form:

$$\begin{aligned} \mathcal{H}^{el}_{QM/MM} &= C \sum q_i \frac{1}{|r - R_i|} + \sum_{\alpha} D^{\alpha} \sum q_i \frac{(R_i^{\alpha} - \bar{r}^{\alpha})}{|r - R_i|^3} + \\ &+ \frac{1}{2} \sum_{\alpha\beta} Q^{\alpha\beta} \sum q_i \frac{(R_i^{\alpha} - \bar{r}^{\alpha})(R_i^{\beta} - \bar{r}^{\beta})}{|r - R_i|^5} + ... \end{aligned} \tag{2.70}$$

  where $\bar{r}$ is the origin of the multi-polar expansion (e.g., the geometrical center of the quantum system) and C, D and Q are the total charge, the dipole and the quadrupole quantum charge distributions, respectively.

The only free parameters in this approach are the cut-off radii r_1 and r_2. This approach allows a fully Hamiltonian description of the electrostatic interactions. Thus, energy conservation during dynamics is achieved if the potential energy and the forces are consistently computed. The correctness of the implementation can be directly verified by monitoring the conservation of the energy during MD simulations. Some examples are given in refs. [91, 92], and exhibit negligible shifts.

2.4.2 Partitioning the system

Partitioning the system (i.e. defining the quantum region) is a crucial issue in modelling biological complexes within a QM/MM approach. As a larger QM region ensures a better accuracy and a better predictive power for the simulation, it also heavily increases the computational cost of the whole calculation, which scales as the third power of the QM region size. When studying an enzymatic reaction, as in the work presented in the following chapter, all the reactive moieties have to be within the QM part (i.e. the nucleophile and the electrophilic centre). Moreover, as in our specific case different acidic hydrogens are present in the active site, all hydrogen bonded groups to the active site were introduced in the QM region, as polarization effects over different groups may play a crucial role in the whole proton dynamics of the system.

2.4.3 Electronic structure calculations

Electronic structure calculations are of crucial importance both for providing the correct parameters in long-range part of the interface Hamiltonian and for data analysis of the trajectories.

D-RESP charges

Within the QM/MM scheme there exists the possibility of calculating, at each time-step, *restrained electrostatic potential derived charges* (RESP) associated to the QM atoms [15]. RESP charges are fitted in order to reproduce the electric potential due to the QM electronic density on the MM atoms within the first shell of the hierarchical partition as described in section 2.4.1. Since the explicit contribution has to be computed in any case at every time-step, RESP charges can be computed *on the fly* with no additional computational cost. For this region, they are re-called "dynamically generated RESP" (D-RESP). The instantaneous electrostatic field on the *i-th* MM atom generated by the electron density is:

$$V_i(\mathbf{R}_i) = \int \rho(\mathbf{r})\nu(|\mathbf{r} - \mathbf{R}_i|)d\mathbf{r} \tag{2.71}$$

the collection of V_i's is used as target for a least-square fit. As D-RESP can be defined as the set of point charges $\{q_i^D, i \in QM\}$, located on the QM nuclei which reproduce in the best-possible way the electrostatic field on the MM atoms, they can be derived minimizing the norm:

$$E = \sum_{j \in MM} \left(\sum_{i \in QM} \frac{q_i^D}{|\mathbf{R}_i - \mathbf{R}_j|} - V_j \right)^2 + W\left(\{q^D\}\right) \tag{2.72}$$

where W is a restraining quadratic function:

$$W\left(\{q^D\}\right) = w_q \sum_{i \in QM} \left(q_i^D - q_i^H\right)^2 \tag{2.73}$$

where w_q is a free parameter fixed to 0.1, and $\{q^H\}$ are the Hirschfeld charges of the QM atoms defined as:

$$q_i^H = \int \rho(\mathbf{r}) \frac{\varrho_i^{at}(|\mathbf{r} - \mathbf{R}_i|)}{\sum_j \varrho_j^{at}(|\mathbf{r} - \mathbf{R}_j|)} d\mathbf{r} - Z_i \tag{2.74}$$

Where ϱ_i^{at} is the pseudo-valence charge density of the *i-th* atom, and Z is its valence.

Maximally localized Wannier functions

"Traditional" chemical representations (bond orders, lone-pairs etc..) find a straightforward relation to electronic structures as the latter are expressed in terms of localized orbitals. Boys orbitals [93] (BO's) are broadly used in the literature to perform such characterization. In a plane waves approach, a generalization of BO's are introduced to map electronic density in highly localized molecular orbitals. The maximally localized Wannier functions (WF [94, 95, 96]) are derived by a unitary transformation from the KS wavefunctions. The positions of the WC centers allow an accurate description of polarization effects during quantum (i.e. CP or QM/MM) simulations.

According to Marzari and Vanderbilt [95], the Wannier functions are defined in terms of Bloch functions in the Γ point approximation:

$$w_i(\mathbf{r}) = u_{i\mathbf{k_0}}(\mathbf{r}) = \sum_{\mathbf{g}} c_i^{\mathbf{k_0}}(\mathbf{g}) e^{-i\mathbf{g}\cdot\mathbf{r}} \tag{2.75}$$

the Wannier function set in not unique, because of the phase term in Bloch functions, thus, as a general unitary transformation

$$|u_{i\mathbf{k_0}}\rangle \rightarrow \sum_j U_{ij}^{\mathbf{k_0}} |u_{j\mathbf{k_0}}\rangle$$

preserves the WF center positions $\langle \mathbf{r} \rangle_i = \langle w_i | \mathbf{r} | w_i \rangle$ but not their square, $\langle r^2 \rangle_i = \langle w_i | r^2 | w_i \rangle$, it is plausible to find the set of WF's that is maximally localized on their centers, by minimizing with respect to the unitary transformation, the spread functional:

$$\Omega = \sum_i \left[\langle r^2 \rangle_i - \langle \mathbf{r} \rangle_i^2 \right] \tag{2.76}$$

2.5 Free-energy profile reconstructions

Many biologically relevant events (conformational changes, enzymatic reactions, etc) require the crossing of relatively large free energy barriers. As a result, is too improbable to observe these processes during typical timescales of quantum of classical mechanics MD simulations.

It is therefore necessary to somehow force the system to undergo the requested transformation. This task implies two major problems that can be resumed into: *(i)* the eventual simplification of the phase space into few (or even one!) relevant parameters (so-called *"Reaction coordinates"*) along which drive the transformation, and *(ii)* the definition of efficient schemes for sampling the (eventually) *reduced* phase-space.

Different algorithms that allow answers to these points have been developed in the last decades [12, 97, 98, 99, 100, 101, 102].

Although reaction coordinates are, in principle, combinations of large sets of atomic coordinates, in some cases, as in enzymatic reactions, monodimensional approximations can be reasonable, as the other degrees of freedom influence the system on much more longer-timescales. The monodimensional approximation has been used in this thesis, and the algorithms used to address question *(ii)* are here briefly presented.

2.5.1 Constraint dynamics

The *Constraint MD* method has been developed by Ciccotti *et al.* [101] in 1998, and has been used to a variety of molecular systems up to date.

In this scheme, the reaction free energy is calculated by rigidly constraining the chosen reaction coordinate. The force on the constraint (ƒ) is time-averaged upon successive variations of the reaction coordinate value $\mathcal{R}$, and then integrated along $\mathcal{R}$. The Free energy, then, reads:

$$\Delta G(\mathcal{R}) = \int_{\mathcal{R}_0}^{\mathcal{R}} \bar{\mathfrak{f}}(\mathcal{R}')d\mathcal{R}' \tag{2.77}$$

For the simple case of an interatomic constraint, like in the case of the simulations of this work, the force on the constraint is equal to the corresponding Lagrange multiplier:

$$\bar{\mathfrak{f}}(\mathcal{R}) = \langle\lambda(\mathcal{R})\rangle = -\left\langle \frac{\partial V}{\partial R_{AB}} - \frac{2k_BT}{R_{AB}} \right\rangle_{R_{AB}=\mathcal{R}} \tag{2.78}$$

where the brackets stand for an ensemble average over the trajectory, and the reaction coordinate $\mathcal{R}$ is defined by the distance between atoms A and B.

The validity of this thermodynamic integration scheme is independent on the nature of the interatomic forces, and it has been successfully extended to a variety of chemical systems (e.g. [20]) Practically, the value of $\mathcal{R}$ is increased or shortened in a small amount starting from the final configuration of the simulation at the former $\mathcal{R}$ value. The amount of these increments determines the accuracy of the free energy estimation for e.g. a chemical transformation: nonetheless, the shorter they are, the more time-consuming the simulation will be. This methodology has been used in this work to compute the free energy of the first-step of the enzymatic reaction of β-secretase, as to compare the results obtained with the same techniques in refs. [5, 7] on the HIV1 protease.

2.5.2 Multiple Steering Molecular Dynamics (MSMD)

In 1997, Jarzynski established a relation between non-equilibrium dynamics and equilibrium properties [12, 13]: let $\mathcal{H}(\mathbf{r}, t)$ be the Hamiltonian of a system that is subject to an external time-dependent perturbation, and let $\Delta G(t')$ and $W(t')$ be, respectively, the change in free-energy and the external work performed on the system as the system evolves from $t = 0$ to $t = t'$. Here $\mathbf{r}$ indicates a configuration of the whole system. According to Jarzynski [12], $\Delta G(t')$ and $W(t')$ are related to each other by the following identity:

$$\mathrm{e}^{-\beta\Delta G(t')} = \langle \mathrm{e}^{-\beta W(t')} \rangle, \tag{2.79}$$

where the brackets represent an average taken over an ensemble of trajectories. The previous equality can be also reformulated as [103, 104]:

$$\mathbf{e}^{-\beta[\mathcal{H}(\mathbf{r}',t')-G(0)]} = \langle \delta[\mathbf{r}(t') - \mathbf{r}'] \mathrm{e}^{-\beta W(t')} \rangle. \tag{2.80}$$

The Hamiltonian $\mathcal{H}(\mathbf{r},t)$ can be written as the sum of the time-independent Hamiltonian of the unperturbed system, $\mathcal{H}_0(\mathbf{r})$, plus the time-dependent external potential. In the present study, as well as in refs. [105, 14, 106, 104] the perturbation has been chosen to be a harmonic potential, which minimum position moves at constant velocity v according to:

$$\mathcal{H}(\mathbf{r},t) = \mathcal{H}_0(\mathbf{r}) + \frac{k}{2}[z(\mathbf{r}) - z_0 - vt]^2, \tag{2.81}$$

where $z(\mathbf{r})$ represents a chosen reaction path. In this way, at any given time, only a small window around the equilibrium position $z(\mathbf{r}) = z_0 + vt$ is sampled. By substituting Eq. 2.81 into Eq. 2.80 one finds:

$$G(z') = -\beta^{-1} \ln \left\langle \delta[z(\mathbf{r}(t')) - z'] \, \mathrm{e}^{-\beta W(t')} \right\rangle . \tag{2.82}$$

Thus, the free-energy of a process along a selected reaction coordinate can be computed performing a representative number of finite-time transformations, collecting at each time-step the work done, and then properly averaging it.

Eq. 2.82 requires the convergence of an exponential, which is slow. Thus, it is more convenient to use a second-order cumulant expansion [107]

$$G(z') \approx -\beta^{-1} \ln\langle \delta[z(\mathbf{r}(t')) - z']\rangle + \langle W(t')\rangle_{z'} - \frac{\beta}{2}\sigma^2 \langle W(t')\rangle_{z'}, \tag{2.83}$$

where $\langle \cdots \rangle_{z'}$ denotes an average restricted to trajectories satisfying the condition $z[\mathbf{r}(t')] = z'$ and σ^2 is the variance of the average work done.

An experimental validation of the reliability of Jarzynski's methodology has also recently appeared [108].

This methodoloy has been used in this work to compute the free energy of the hydrolysis of formamide in aqueous solution, as a reference parameter for the enzymatic reaction of β-secretase.

2.6 Coarse Grained computations

The capabilities of actual computer calculations allow, within a reasonable cost, to follow the dynamical evolution of few hundred residues large proteins for tenths of nanoseconds. Although such system dimensions and simulation timescales allow to gain considerable insight into relevant aspects of protein dynamics from simulation data, they are too small, or short, for studying other major aspects of dynamics of biological systems such as larger conformational changes or molecular recognition

processes [109]. In addition, MD runs might not be sufficiently long to legitimate a time average / thermodynamic average equivalence [90].

Several studies have attempted to bridge the gap between the dimensions or timescales of feasible MD simulations and the ones of relevant biological mechanical processes by recurring to a mesoscopic rather than a microscopic approach [110].

In fact, the large-scale dynamical features encountered in MD trajectories can be interpreted, at first approximation, as a superposition of independent harmonic modes [109] that can be achieved by simply substituting detailed force-field interactions by harmonic couplings with the same spring constant [110].

2.6.1 β-Gaussian Model

The results found by Tirion lead to development of coarse-grained schemes where amino acids are represented by effective centroids corresponding to the C_α atoms and the Hamiltonian is a summation of harmonic couplings between pairs of spatially close centroids [111, 112, 113].

At variance with previous approaches, in this thesis the so-called β-Gaussian model [22] has been used. The main difference consists in introducing effective C_β centroids, tethered to the C_α's, that mimic the side-chain, allowing a better control of the directionality of pairwise interactions in the protein, leading to an improved vibrational description.

Modeling the Hamiltonian

The Hamiltonian of a protein is described as an expansion in terms of deviations of the amino acids from their reference positions:

$$\mathcal{H}(\Gamma) = \mathcal{H}_B(\Gamma) + \mathcal{H}_{\alpha\alpha}(\Gamma) + \mathcal{H}_{\alpha\beta}(\Gamma) + \mathcal{H}_{\beta\beta}(\Gamma) \tag{2.84}$$

where

$$\begin{aligned}
\mathcal{H}_B(\Gamma) &= k\sum_i V^{C_\alpha - C_\alpha}\left(d_{i,i+1}^{C_\alpha - C_\alpha}\right) \\
\mathcal{H}_{\alpha\alpha}(\Gamma) &= \sum_{i<j} \Delta_{ij}^{C_\alpha - C_\alpha} V^{C_\alpha - C_\alpha}\left(d_{i,j}^{C_\alpha - C_\alpha}\right) \\
\mathcal{H}_{\alpha\beta}(\Gamma) &= \sum_{i,j} \Delta_{ij}^{C_\alpha - C_\beta} V^{C_\alpha - C_\beta}\left(d_{i,j}^{C_\alpha - C_\beta}\right) \\
\mathcal{H}_{\beta\beta}(\Gamma) &= \sum_{i<j} \Delta_{ij}^{C_\beta - C_\beta} V^{C_\beta - C_\beta}\left(d_{i,j}^{C_\beta - C_\beta}\right)
\end{aligned} \tag{2.85}$$

Δ_{ij} is the native contact matrix that loads the values of 1 or 0 if the native distance between two particles i and j is below a certain cutoff value *R*.

To account for the higher strength of the peptide bonds with respect to non-covalent contact interactions between amino acids, an explicit chain term $\mathcal{H}_B$ has been added, where the interaction of consecutive C_α's is controlled by $k > 1$.

By construction, the global minimum of this Hamiltonian coincides with the native state. For small fluctuations around the native structure, the potential interaction energy of two particles can be expanded in a Taylor series. If $\overrightarrow{r}_{ij}$ is the native distance vector and $\overrightarrow{x}_{ij}$ is the deviation vector, so that the total distance vector is $\overrightarrow{d}_{ij} = \overrightarrow{r}_{ij} + \overrightarrow{x}_{ij}$, the pairwise interaction can be approximated as:

$$V(\overrightarrow{d}_{ij}) \approx V(\overrightarrow{r}_{ij}) + \frac{V''(\overrightarrow{r}_{ij})}{2}\sum_{\mu,\nu}\frac{r_{ij}^{\mu} r_{ij}^{\nu}}{r_{ij}^2} x_{ij}^{\mu} x_{ij}^{\nu} \tag{2.86}$$

where μ and ν run over the three Cartesian components. Based on this quadratic expansion, the Hamiltonian in eq. 2.84 can be approximated as:

$$\begin{aligned}
\mathcal{H} \approx \frac{1}{2}\sum_{ij,\mu\nu} x_{i,\mu}^{C_\alpha}\mathcal{M}_{ij,\mu\nu}^{C_\alpha - C_\alpha} x_{j,\nu}^{C_\alpha} &+ \sum_{ij,\mu\nu} x_{i,\mu}^{C_\alpha}\mathcal{M}_{ij,\mu\nu}^{C_\alpha - C_\beta} x_{j,\nu}^{C_\beta} + \\
&+ \frac{1}{2}\sum_{ij,\mu\nu} x_{i,\mu}^{C_\beta}\mathcal{M}_{ij,\mu\nu}^{C_\beta - C_\beta} x_{j,\nu}^{C_\beta}
\end{aligned} \tag{2.87}$$

Where $\mathcal{M}$'s are symmetric matrices.

However, the location of the C_β atoms in a protein structure is almost uniquely determined by the geometry of the peptide chain [114]. Thus, the degrees of freedom can be reduced to only those of the C_α's, whereas the fluctuations of the C_β's are dictated by those of the former.

The location of the *i-th* C_β is given by:

$$\overrightarrow{r}_i^{C_\beta} = \overrightarrow{r}_i^{C_\alpha} + \ell \frac{2\overrightarrow{r}_i^{C_\alpha} - \overrightarrow{r}_{i+1}^{C_\alpha} - \overrightarrow{r}_{i-1}^{C_\alpha}}{|2\overrightarrow{r}_i^{C_\alpha} - \overrightarrow{r}_{i+1}^{C_\alpha} - \overrightarrow{r}_{i-1}^{C_\alpha}|} \tag{2.88}$$

where ℓ= 3 Å. Deviations of the the C_β positions are then fully derived from those of the C_α's:

$$\overrightarrow{x}_i^{C_\beta} = \overrightarrow{x}_i^{C_\alpha} + \ell \frac{2\overrightarrow{x}_i^{C_\alpha} - \overrightarrow{x}_{i+1}^{C_\alpha} - \overrightarrow{x}_{i-1}^{C_\alpha}}{|2\overrightarrow{x}_i^{C_\alpha} - \overrightarrow{x}_{i+1}^{C_\alpha} - \overrightarrow{x}_{i-1}^{C_\alpha}|} \tag{2.89}$$

In this way, C_β are localized as a function of the C_α's[5], and equation 2.87 takes the form of:

$$\mathcal{H} \approx \frac{1}{2} \sum_{ij,\mu\nu} x_{i,\mu}^{C_\alpha} \tilde{\mathcal{M}}_{ij,\mu\nu}^{C_\alpha - C_\alpha} x_{j,\nu}^{C_\alpha} \tag{2.90}$$

where $\tilde{\mathcal{M}}$ is a new 3Nx3N symmetric matrix, and the elastic response of the system is uniquely dictated by the eigenvalues and eigenvetors of $\tilde{\mathcal{M}}$.

Equilibrium properties

Several theoretical, experimental and computational studies have demonstrated that the dynamics of a protein is over-damped by the solvent [85, 86, 87, 88].

For small deviations from the reference position, the dynamics of the coarse-grained amino acids can be written as:

$$\dot{x}_{i,\mu}(t) = -\sum_{j,\nu} \tilde{\mathcal{M}}_{ij,\mu\nu} x_{j,\nu}(t) + \eta_{i,\mu}(t) \tag{2.91}$$

the stochastic noise satisfy:

$$\begin{aligned} \langle \eta_{i,\mu} \rangle &= 0 \\ \langle \eta_{i,\mu} \eta_{j,\nu} \rangle &= \delta_{ij} \delta_{\mu\nu} 2k_B T \end{aligned} \tag{2.92}$$

These two conditions ensure the onset of canonical thermal equilibrium, so that the probability of a given configuration $\{x\}$ for the particles in the system is controlled by the Boltzmann factor.

[5] C_β positions are determined for *all* residues but Gly, and for the initial and final residues in the chain, as they lack one of the flanking C_α's.

Covariance matrices and temperature factors

The main observable quantity that can be calculated within the Gaussian model is the correlation between displacements of pairs of C_α atoms. The thermodynamic averages of the correlated displacements are obtained from the inversion of the $\tilde{\mathcal{M}}$ matrix. In fact, if the thermal factor β^{-1} is set to 1, one has:

$$\langle x_{i,\mu} x_{j,\nu} \rangle = \tilde{\mathcal{M}}^{-1}_{ij,\mu\nu} \tag{2.93}$$

the inverse matrix $\tilde{\mathcal{M}}^{-1}_{ij,\mu\nu}$ is referred as the covariance matrix $\mathcal{C}$, and its eigenvectors represent the three-dimensional independent modes of structural distortion for the protein (see section. 2.3.2).

The mean-square fluctuations (rmsf^2) for each aminoacid are easily obtained as:

$$\left\langle \left| \overrightarrow{x}_i^2 \right| \right\rangle = \sum_\mu \mathcal{C}_{ii,\mu\mu} \tag{2.94}$$

which are related to experimental temperature-factors (or B-factors), as already stated in section 2.3.2.

2.7 Evolutionary patterns in protein families

From determination of the DNA structure [115] to decodification of the amino acidic alphabet, evolutionary theories coupled to molecular biology and genetics have been being able to shed light on the mechanisms followed by nature that lead to differentiation, speciation, and eventually familiar genetic diseases.

2.7.1 Evolution of a protein sequence

The analysis of evolution at molecular level must consider the processes which alter the DNA sequences (e.g. errors in DNA replication of DNA repair). These changes provide the *mutations* upon which natural selection can act. The results of genotypic changes provide the molecular record of evolution: the more closely related two species are, the more similar are their genome sequences. In the case of proteins, similarity relationships can be extended from their sequence to their structure. In fact, as the function of a protein is strictly related to its three-dimensional structure, conservation of the shape is usually a better index of the evolutionary relationship between two proteins rather than sequence identity itself.

Evolutionary relationships between members of protein families can be classified according to the concepts of *homology* and *analogy*: sequences are homologous if they are related by evolutionary divergence from a common ancestor, analogous if the have acquired shared features (e.g. protein fold or function) by convergent evolution. Homologous sequences lead to structurally similar proteins.

A critical task confronting genome sequencing projects is the structural and functional characterization of new proteins. This task can be solved by recognizing the evolutionary pattern of newly discovered protein sequences, and thus reconstructing their "family trees" , starting from already characterized proteins.

2.7.2 Protein homology and sequence similarity

The major postulate in divergent evolution is that each single mutation process is not likely to change very much the structure/function of the original protein. As a result, protein sequences can highly mutate while keeping their functionality, as long as the *functional regions* of the protein (e.g. the active site of an enzyme) remain untouched. This feature has been proved by Sander and Schneider in 1991 [116], when they showed that natural proteins are surely "homologous" up to a threshold of around 70-75% of mutations in their sequence.

Thus, when discovering a new protein sequence, it is crucial to identify in the protein database whether there exists any homologous protein which has been already characterized, in order to reconstruct the structure and define the function of the new sequence.

Identity matrix

The only way to recognize whether two proteins belong to the same family, and how distant they are along the evolutionary path, is to match the amino acids in the two sequences, and look at how many of them are conserved. If s^A and s^B are two protein sequences, their identity can be defined as:

$$\mathfrak{I} = \sum_i \delta_{s_i^A s_i^B} - \mathcal{P}_{gap} \tag{2.95}$$

As evolution may have caused insertions or deletions in the sequences, it is necessary to include a $\mathcal{P}_{gap}$ penalty scoring function in the formula, in order to allow handling of these features.

Similarity matrices

The identity matrix provides a *lower limit* for homology recognition, as every mutated aminoacid scores zero. In practice, a better evaluation of mutations in an alignment can be achieved by using "similarity matrices" $\mathcal{S}$, which assign a different score depending which residue is mutated into which. The total score $\mathfrak{Z}$ will then be written as:

$$\mathfrak{Z} = \sum_i \mathcal{S}_{s_i^A, s_i^B} - \mathcal{P}_{gap} \tag{2.96}$$

the definition of the matrix elements of $\mathcal{S}$ is not unique, and several schemes are proposed; the most common ones are usually based on the frequency two amino acids are mutated into each other in families of homologous proteins [117].

Multiple sequence alignment

The schemes presented so far allow to align two protein sequences, provided they are homologous. Anyway, finding the evolutionary connections in distantly related proteins (so-called: *remote-homolog detection*) is a complex and difficult task. In this case, a single alignment could be not accurate enough, as the best alignment score might be heavily dependent on the gap penalty function, or on the chosen $\mathcal{S}$ matrix. Therefore, it can be useful to try to align as many sequences as possible (up to the whole family) into a *multiple sequence alignment*.

Moreover, as proteins are subject to evolutive pressure, functional relevant regions should be better conserved than the others; therefore, an alignment of the whole family surely results extremely useful in identifying with high accuracy these regions.

Different algorithm which are able to efficiently search in the protein database and align homologous sequences are nowadays available [118, 119]. The method used in this thesis is the SAM-T99 Hidden Markov Model (HMM) algorithm, implemented by Karplus *et al.*[6] and briefly resumed here.

Hidden Markov Models Profile Hidden-Markov Models [120, 121, 122] have been demonstrated to be very effective in detecting conserved patterns in multiple sequences [123]. The main algorithm follows the scheme as in fig. 2.1: the profile is a chain of match (square), insert(diamond) and delete (circle) nodes, with all transitions between nodes, and character costs in the insert and match nodes trained to

[6] free server available at http://wwww.cse.ucsc.edu/research/compbio/HMM-apps/T99-query.html

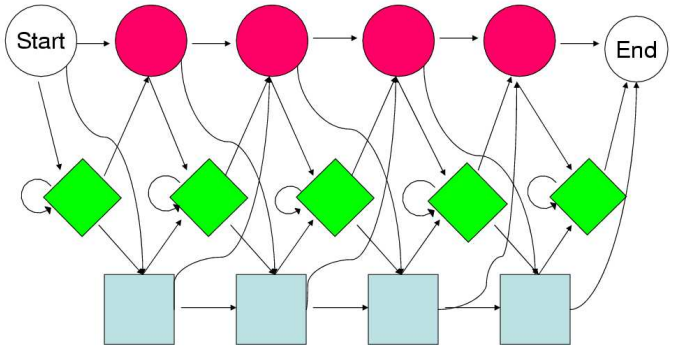


Figure 2.1: Scheme of an HMM, where sequences go from *Start* to *End* passing through match(Square), insert(diamonds) or delete(circles) states.

specific probabilities. The single best path through an HMM corresponds to a path from "start" to "end" in which each character of the sequence is related to a successive match or insertion state along that path. Delete state indicate that the sequence has no character corresponding to that position in the HMM. The accumulation of data from aligned sequences *trains* the transition probabilities; therefore, if a HMM is trained on sequences that are members of the same family, the resulting HMM will identify the regions in the sequence of amino acids that are conserved and characteristic of the family. This scheme is also able to discriminate between family and non-family members in the sequence-database search, or to identify cross-conserved domains in proteins belonging to different families. In this thesis HMM has been used to identify conserved sequence regions in the pepsin family.

Chapter 3

Enzymatic activity of β-secretase

The identification of Memapsin2 being the long sought key-pharmaceutical target β-secretase [36] has resulted in a large concentration of efforts on this protein. In these last years, a large amount of structural information [60, 61, 124] has been produced. The determination of the molecular details of its catalytic action and the identification of the transition-state geometry may have a positive fallout in the determination of possible model-drugs.

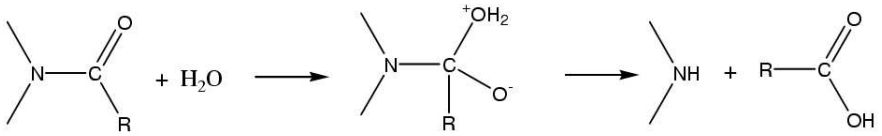
In the most characterized retropepsin, HIV-1 AP[125], conformational fluctuations play a major role for the enzymatic catalysis, as the protein acts as a sophisticated machinery capable of steering the substrate toward the catalytic Asp dyad to favor a reactive conformation [5, 6, 49]. The key question is, therefore, whether also BACE enzymatic activity could be modulated by mechanical fluctuations.

Here, different molecular dynamics techniques have been used as to find answers to these key questions. The enzymatic reaction has been studied by, first, inferring a model reference reaction in water (section 3.1) through full QM computations; then, structural and dynamical features of the enzyme-substrate complex are identified by performing 20 ns of classical MD simulations on the enzyme in complex with a model substrate. In particular, conformational changes in the protein-substrate complex which affect the structure of the active site have been evidenced (section 3.2). Finally the reaction in the Michaelis complex (section 3.3) has been investigated *via* a QM/MM scheme.

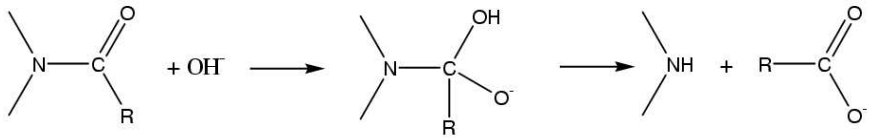
3.1 Reference reactions: formamide hydrolysis

A reliable model reference for the enzyme-catalyzed hydrolysis of peptides is provided by hydrolysis of amides in aqueous solution [126].

Small amides are very stable in aqueous solution and hydrolysis of their amidic bond does not occur at room temperature and physiological pH [127]; in these conditions, the uncatalyzed reaction,



which is initiated by water addition, is so disadvantaged that, even though there is some experimental evidence for this reaction to occur [127], kinetic data favor a base-catalyzed mechanism even at neutral pH [127, 128]. On the other hand, both acid or basic solutions exhibit catalytic activity on the reaction. From a biochemical point of view, the base-promoted reaction



is more important than the acid-promoted one because of its closer analogy with most enzymatic mechanisms [126]. The experimental activation free-energy for the base-catalyzed reactions is of $\approx$ 20-30 kcal mol^{-1} [129, 130] depending on the substrate and the temperature.

The first step of the reaction at both neutral and alkaline pH has been the subject of many computational studies. The ab-initio calculated activation free-energy of the uncatalyzed reaction is about 55 kcal mol^{-1} in vacuo [131, 132], and 48-51 kcal mol^{-1} in the presence of the solvent [133], whereas the free-energy barrier for the base-catalyzed reaction decreases to 18 kcal mol^{-1}. Furthermore, for the base-catalyzed reaction the activation barrier has been shown to be completely solvent induced [134, 135, 136]. Empirical valence bond calculations by Warshel and coworkers [126, 137, 138] on the base catalyzed reaction confirm this proposal.

Here, first principles molecular dynamics simulations have been carried out within the Car-Parrinello (CPMD)[11] approach, using the CPMD code [75], and free-energy computations have been done following the MSMD scheme, as reported in chapter 2.5.2.

Simulation parameters

The model for the reaction in water has been composed by a formamide molecule (FOR) and 56 water molecules enclosed in a periodic box of
13.5 Å×11.8 Å×11.8 Å edges (Fig. 3.1).

This corresponds to take explicitly into account two complete solvation shells of formamide plus a part of the third shell. The OH^- attack has been investigated in a simulation cell obtained from the previous one by mutating a randomly chosen water molecule into a hydroxyl ion (Fig. 3.2). Here, an uniform positive background has been added to neutralize the total electric charge. In this way, it has been possible to simplify the model and avoid the continuous presence of the counter-ion nearby the reaction site, which could nonphysically alter the dynamics because of the small size of the cell. Previous CPMD studies of charged aqueous solutions, carried out with simulation cells comparable in size, number of atoms and charge, showed that finite size effect are not crucial in determining the main dynamical properties of an ionic aqueous solution [139, 140].

Accuracy of the quantum-chemical calculations The majority of the studies of chemical reactivity carried out so far using DFT seem to indicate that the reaction barriers are underestimated (see for example Refs. [141, 142]). To have an estimate of a possible intrinsic error in our pseudo-potential-DFT/BLYP calculations, we have computed the activation energy for the addition of water to formamide in vacuo (concerted mechanism). No zero-point-energy effects have been taken into account. The calculation has been performed at DFT/BLYP level, as described in sec. 2.2, in a 8.5 Å×9.0 Å×9.0 Å cell. The value obtained (40.5 kcal mol^{-1}) is only slightly lower than the value reported by Jensen et al. [132] at MP2/6-311G** level (42.8 kcal mol^{-1}). This result makes us confident that the present DFT/BLYP calculations are adequate to reasonably describe the energetics of the reactions investigated in water solution.

3.1.1 Hydrolysis by H_2O addition

Energetics. The free energy profile of the direct addition of water to formamide converges within few steering trajectories (Fig. 3.3a). Furthermore, within our statistical uncertainty, it coincides with that of the reverse reaction. Hence, for this system, hysteresis does not seem to be relevant. The calculated free energy of the reaction is of 44 kcal mol^{-1}, which is fairly good agreement with previous calculations [133]. The inverse reaction features a small activation energy (2 kcal mol^{-1}). Although this suggests that the intermediate is very labile, it must be pointed out that recent findings [7]

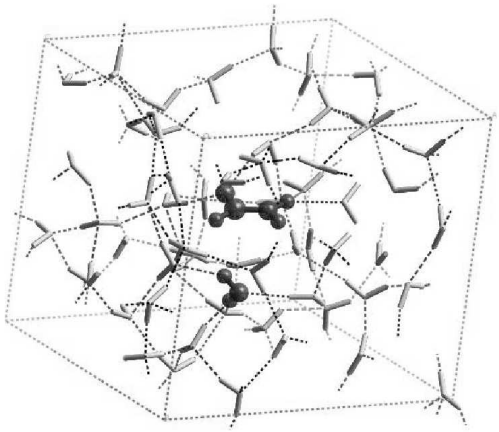


Figure 3.1: Simulation box for WAT addition to FOR. FOR and WAT are drawn in dark-grey balls and sticks, the other water molecules are drawn in light-grey cylinders.

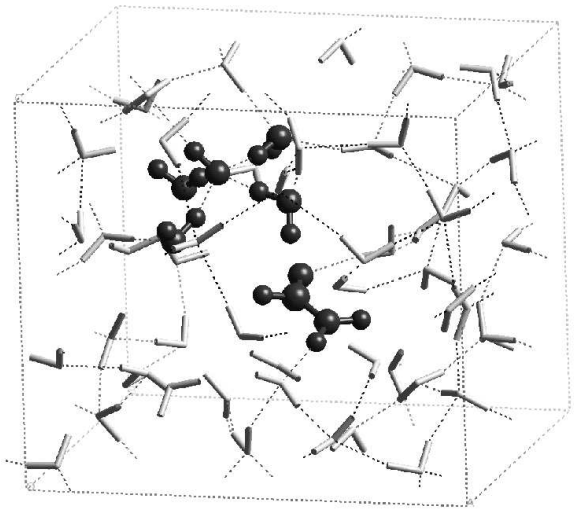


Figure 3.2: Simulation box for HYD addition to FOR. FOR, HYD and its hydration shell are drawn in dark-grey balls and sticks, the other water molecules are drawn in light-grey cylinders.

indicate that the free energy of the intermediate in this reaction may be overestimated by the BLYP calculations by several kcal mol^{-1}.

Structural features. At large WAT-FOR distances, the solvation of reactants resembles that in the bulk; in particular, WAT hydrogen-bond length is $\approx 1.9 \pm 0.2$ Å, as in CPMD simulations of bulk water. WAT attacks FOR perpendicularly to the FOR molecular plane. The molecular planes of FOR and WAT atoms get almost parallel as the transition state (TS) is approached. Furthermore, as the reaction coordinate goes below 2.5 Å, one of the two H-bonds received by WAT breaks. After breaking, the new solvation structure of WAT is conserved up to the formation of the tetrahedral adduct (Figs. 3.4a and c). The two H-bonds donated by WAT never break during reaction. The tetrahedral angles of the TS depend on the structure of the solvent, without showing a preference for staggered or eclipsed configurations of the two molecules.

The approach of WAT to FOR, up to the TS, neither induces relevant changes in the solvation shell of FOR nor involve interactions between reactants solvation shells. Thus, the dynamical features of the direct reaction are essentially related to the approaching motion of the two reagents that does not requires any complex rearrangement of the solvent close to the TS. The TS is found at a C-O_{WAT} distance of 1.73 Å, with an activation free-energy of 44 kcal mol^{-1}. The equilibrium distance of the intermediate is found at a distance of 1.61 Å. The intermediate adduct resem-

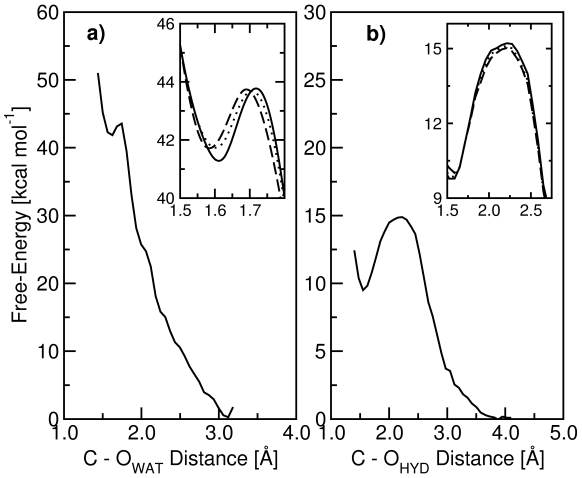


Figure 3.3: Free-energy profile of WAT (a) and HYD (b) addition to FOR. In the insets the TS region is enlarged to show the convergence of the free energy as a function of the number of trajectories (two: dashed line; four: dotted line; six: solid line). In both simulations, no differences in the free-energy profiles obtained averaging over 5 and 6 trajectories have been evidenced.

bles the TS (late TS, Fig. 3.4c). After formation of the tetrahedral intermediate, the N atom acts as H-bond acceptor, whereas the carbonyl oxygen is able to coordinate three H-bonds.

Electronic structures. The electronic structure does not change significantly until the TS is reached (C-O_{WAT} = 1.73 Å). Then, as the π system of FOR is disrupted, one of the two BO's [96, 95, 143] migrates close to the oxygen atom, and the other between the two atoms of the carbonyl group, as evidenced in Fig. 3.5a, where the relative displacement of the two BO's related to the CO bond is shown.

3.1.2 Hydrolysis by OH^- addition

MSMD runs have been performed following the same protocol as for WAT addition. In this case, one of the two reagents brings a negative charge. Because of the long-range nature of electrostatic interactions between the two reagents, the oxygen atom of the hydroxyl ion (O_{HYD}) has been restrained at a longer distance from C (4.5 Å) by a harmonic potential with force constant k = 25 kcal mol^{-1} Å^{-2}. During the equilibration phase, a coordination constraint [144] has been set on O_{HYD} in order to avoid unwanted proton transfers. Six steering simulations have been carried out

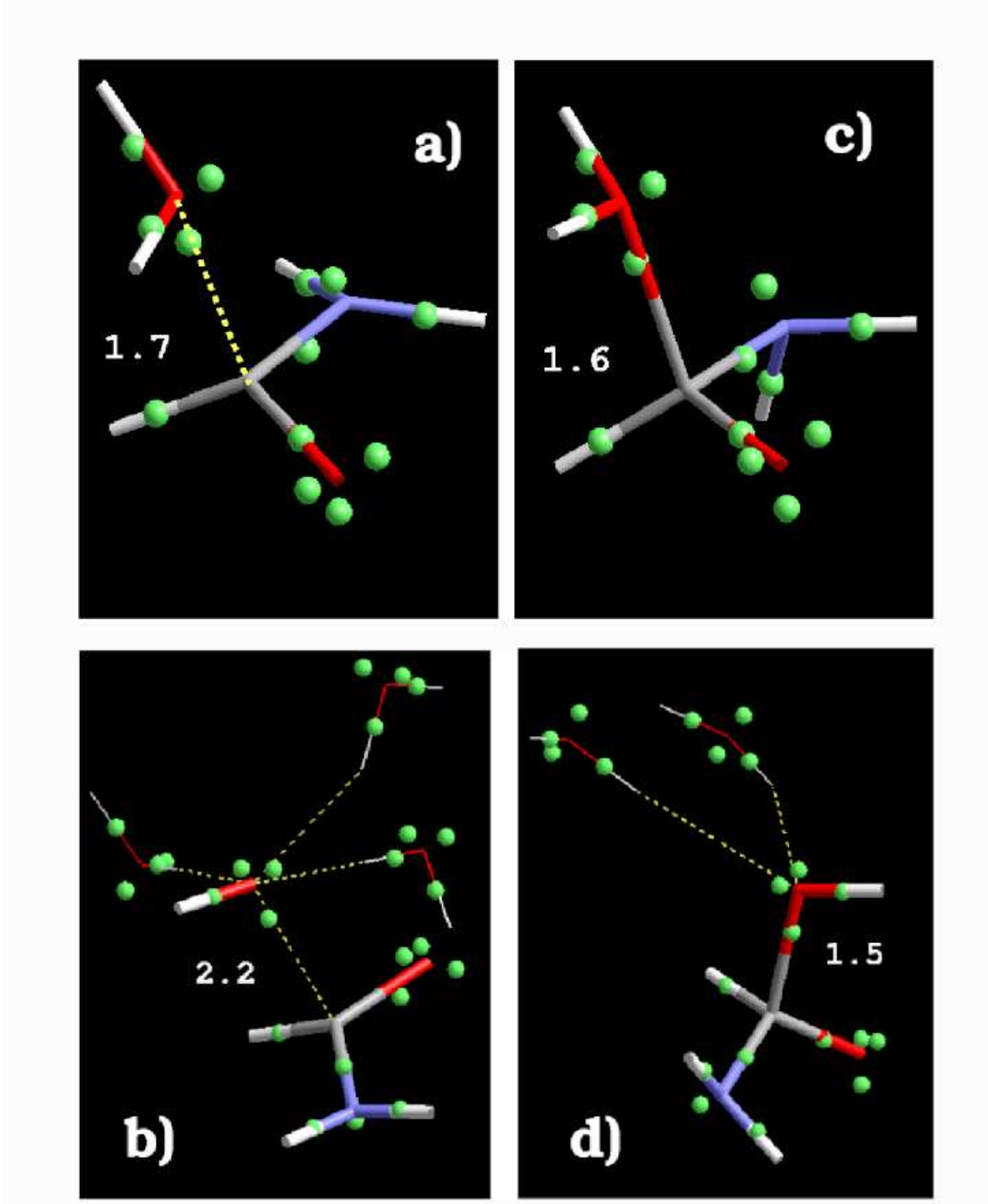


Figure 3.4: Structure of the transition state and the intermediate of WAT (a-c) and HYD addition (b-d). BO's are drawn as spheres. Water molecules H-bonded to OH^- are drawn as sticks. For sake of clarity, the other water molecules are not drawn. The C-O distance (in Å) is also reported.

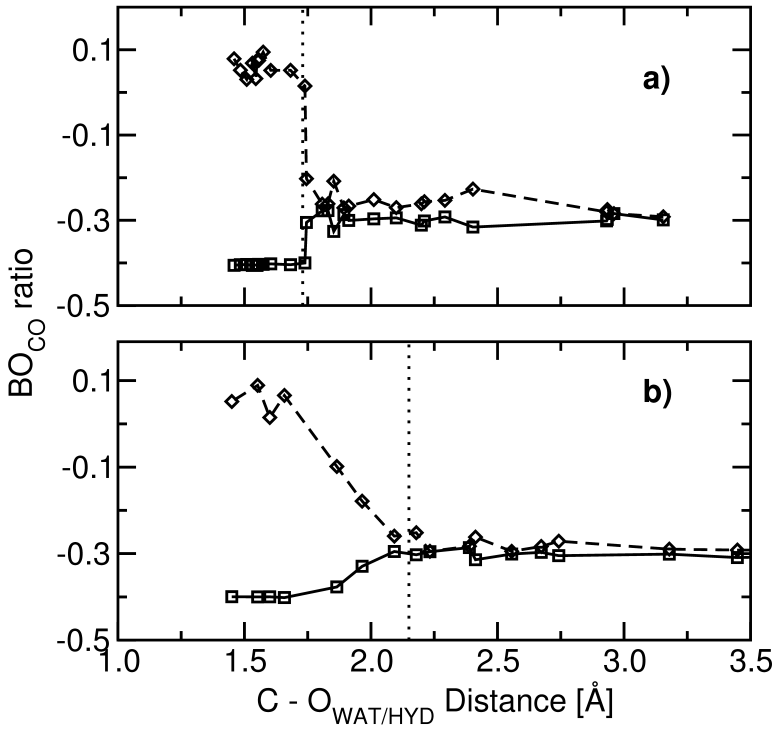


Figure 3.5: Relative displacement of the carbonyl BO's during WAT (a) and HYD (b) approach to FOR. The relative displacement is computed as $(d_{\mathrm{C-BO}} - d_{\mathrm{C-O}})/d_{\mathrm{C-O}}$, where $d_{\mathrm{C-O}}$ is the $\mathrm{C_{FOR}}$-$\mathrm{O_{FOR}}$ bond length and $d_{\mathrm{C-BO}}$ is the $\mathrm{C_{FOR}}$-BO distance. The vertical dotted line indicates the position of the TS.

for both the direct and inverse processes, using as starting point configurations saved every 500 fs from the previous simulation.

In our simulation, OH^- has a four-fold coordination, in agreement with previous calculations at a pH comparable to ours [145, 146, 147, 139, 148], and it has to progressively loose two solvation waters, in order to react to FOR. Both the four and three water solvation structures are very stable [146]. As a consequence, O_{HYD} cannot approach the C atom relaxing correctly with respect to the harmonic restraint; hence, the free energy cannot be reliably calculated with a finite number of trajectories. Simulations performed with a slower pulling speed (v = 0.2 Å ps^{-1}) have manifested the same problem. In contrast, the *inverse* reaction is not affected by this problem. (Fig. 3.3b). In fact, in this case there is a pre-orientation of the reactants (e.g. the H_{HYD}-O_{HYD}-C angle) in the cleavage of the C-O_{HYD} bond, and the subsequent hydration of OH^- occurs readily because the strong H-bond interactions between the hydroxyl ion and water molecules drive a fast formation of the solvation cluster. Hence, the free energy profile of the inverse reaction can be calculated using C-O_{HYD} distance as reaction coordinate.

Energetics.[1] As for the uncatalyzed reaction, the free energy profile converges within a few trajectories (Fig. 3.3b, inset). The calculated activation free energy with respect to solvent-separated FOR and OH^- is 15 kcal mol^{-1} (Fig. 3.3b). This value is in good agreement with the experimental energy of $\approx$ 19 kcal mol^{-1}, measured for FOR derivatives [129, 130]. The inverse reaction features an activation energy of 6 kcal mol^{-1}), similarly to what found in ref. [137].

Structural Features. Although the OH^- ion does not relax properly with respect to the minimum of the harmonic restraint, the *structural* features of the direct simulations have reproduced the same ones found in the inverse reaction. Thus, for sake of clarity, we will describe the reaction as the OH^- attack.

The equilibrium distance of the tetrahedral adduct is found at a C-O_{HYD} distance of 1.55 Å. The TS is located at a C-O_{HYD} distance of 2.20 Å, which is considerably longer than the average TS distance of the water addition. The orientation of the hydroxyl ion with respect to the C atom is correlated to the reaction coordinate; in fact, the average value of the angle formed by H_{HYD}, O_{HYD} and C atoms, as the reactants

[1] It is referred only to the free energy profile of the inverse reaction ,as, during the *direct* reaction of the base-catalyzed attack, the strongly H-bonded OH^- water molecules do not allow, on the time scale of our simulations, for the large rearrangements of the solvation shell that are deemed necessary for the reaction to occur. In other words, the approaching time of OH^- to FOR is too short compared to that required for the OH^- shell to undergo partial disruption.

are further than the TS region, is almost 180°. This geometry is preserved until the C-O_{HYD} distance approaches that of the TS. The reaction occurs as the H_{HYD}-O_{HYD}-C angle bends to a value of about 100°.

The four ligands never leave the solvation shell as the reaction coordinate remains out of the TS region (approximately located between 2.0 and 2.3 Å). The average O_{HYD}-H distance (1.7±0.3 Å) is shorter than H-bonds formed in bulk water (≈ 1.9 Å). The number of water oxygens H-bonded to O_{HYD} decreases from four to three, before the TS (Fig. 3.4b). At the TS, the OH^- reacts when the H_{HYD}-O_{HYD}-C angle bends suitably, due to thermal fluctuations. After the TS, one of the three remaining OH^--bound water molecules is lost (Fig. 3.4d).

The energetics of our calculations are comparable with those calculated by Warshel et al. [126] and Kollman et al. [135]; however, the present study differs for the description of the solvation shell dynamics of the OH^-: indeed, the solvent is implicitly treated in Warshel's calculations [126], whereas the empirical potential describing the OH^--solvent intraraction in [135] provides a different geometry for the solvation structure of OH^- (average number of H-bonds of about 5.8).

Electronic structures. At all the C-O_{HYD} distances, the MD-averaged dipole moment of the four water molecules forming H-bonds with the OH^- anion are appreciably larger than those of bulk water (≈ 3.2 and 3.0 D, respectively). The polarization of the OH^-, for C-O_{HYD} distances larger than that of the TS, results to be negligible (Fig. 3.6b), suggesting that the long-range interaction between OH^- and FOR does not play any crucial role.

Changes in the electronic structure along the reaction are abserved after the TS is crossed (Fig. 3.5b). Thus, the features of the reaction appear to be related the solvent effects, as opposed to electronic properties. This suggests that the activation free-energy for the base-catalyzed reaction is completely solvent induced, as stated in refs. [135, 136].

3.1.3 Summary

The mechanism and energetics of peptide bond hydrolysis differ dramatically upon change of the nucleophile (Water or Hydroxyl ion). In both cases, polarization effects on the nucleophile seem not to play a fundamental role in the reaction, while the solvent has a crucial part in both mechanisms, although for opposite reasons. In fact, in the case of the base-catalyzed reaction, the solvent is entirely responsible for the activation-free energy, as disruption of the hydroxyl ion solvation shell is the *time-*

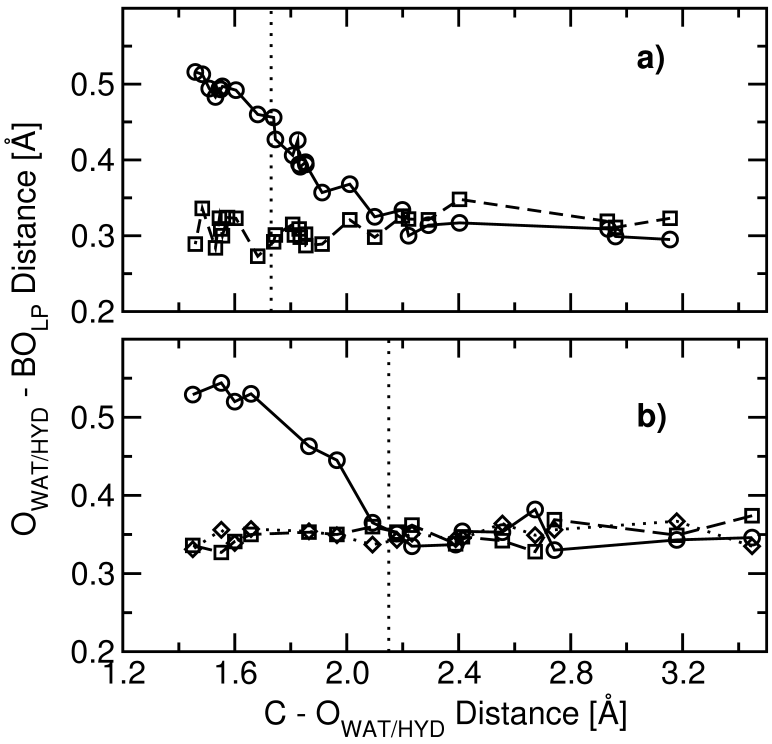


Figure 3.6: Displacement of BO's related to the lone-pairs of WAT (a) and HYD (b) along the reaction coordinate. The vertical dotted line indicates the position of the TS.

limiting event for the reaction. The bottle-neck for water attack is, on the contrary, represented by the extremely low probabilty of water deprotonation during reaction, that is, by the too low pK_b value of the solvent. Therefore, the solvent,or, more generally, the environment, might reduce the free-energy barrier of water addition by enhancing the probability for water to lose one of its protons along the reaction coordinate.

These considerations is validated by results obtained in human caspase-3 [19] and HIV-1 aspartic protease [5], for which similar computational setups have been used. In fact, during reaction in water solution, WAT is practically as polarized as in the bulk, showing a MD-averaged dipole moment of $\approx$ 3.1 D, that is slightly larger than that of the other water molecules in the cell ($\approx$ 3.0 D). Similar values have been calculated, using DFT/BLYP, for peptide hydrolysis catalyzed by these enzymes. In these enzymatic systems, in contrast, efficient proton-shuttle systems, which are able to efficiently deprotonate the reactive water molecule are always present, in particular, the catalytic histidine (His237) in caspase-3 [19], and the catalytic Asp dyad (Asp25, Asp25') in HIV1-PR [5]. In pure water, this efficient proton shuttle is absent, and both WAT hydrogen atoms remain bound to O_{WAT} until the tetrahedral adduct is formed. Only after this event, O_{WAT}-H bonds are weakened and the average bond distance increases from 0.97 Å to 1.05 Å. Hence, deprotonation can happen only after the low-rate step of the reaction has occurred[2].

Therefore, some care will be put in observation of proton dynamics in the enzymatic reaction of BACE, and eventual differences with that of HIV-PR.

3.2 Molecular Dynamics of β-secretase in water

3.2.1 X-ray data

Up to date, there are three structures of BACE deposited in the PDB databank, all solved by J. Tang *et al.* [60, 61, 124]. The structure of the unbound enzyme (pdb entry:1SGZ [124]), solved at 2.0 Å resolution, was the latter to come out (2004), while the first two structures (pdb entries: 1FKN, 1M4H [60, 61]) were crystallized in the presence of peptidomimetic inhibitors. In particular, the inhibitor Om99-2, present in the 1.9 Å resolution structure deposited in 2000 [60] can be easily mutated into an amino acidic sequence that closely resembles that of the substrate β-APP in its

[2]Water deprotonation has been observed within few fractions of ps in two out of six MSMD simulations.

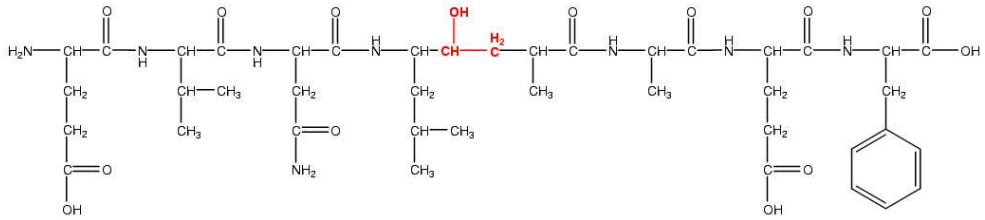


Figure 3.7: Chemical formula of the OM99-2 inhibitor. The hydroxyl-vinyl group, changed into a peptide bond in MD simulations, has been evidenced in red.

swedish-mutation (see section 1.2.2). Therefore, this X-ray structure has been chosen as the starting point for the classical MD simulations.

The PDB file contains two non-equivalent images of the protein-inhibitor complex. One of the two images (labelled as "Chain A" in the pdb file) has two residues more resolved in the N-terminal "pro-sequence", a destructured linker region, nonetheless, the corresponding inhibitor ("Chain C") diverges critically from standard structural parameters (e.g. there are linear peptide bonds); therefore, the "Chain B- Chain D" complex has been chosen as starting structure (Fig. 1.10).

Om99-2 inhibitor is based on a EVNLAAEF octo-peptide chain (fig. 3.7), where the cleavage site, located at the Leu-Ala peptide bond, is substituted by a hydroxyl-vinyl group (fig. 3.7).

The substrate structure was obtained by substituting the hydroxyl-vinyl group with a peptide bond, and minimizing its geometry keeping all other atoms fixed. The hydroxyl group of the inhibitor is putatively occupying the catalytic water molecule position, therefore a water molecule was added to the moiety in between the substrate and the aspartic dyad.

3.2.2 System setup

Classical MD simulations of β-secretase (BACE) were performed applying the following protocol.

Structural model The BACE/EVNLAAEF substrate complex was obtained by replacing the hydroxyl-vinyl group of the OM99-2 inhibitor with the amide group

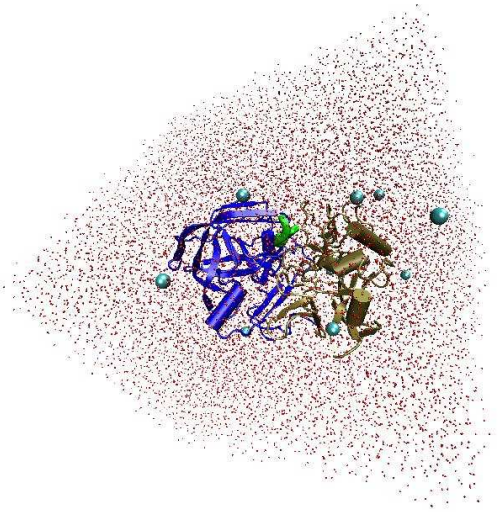


Figure 3.8: Simulation box for the MD run. The protein is depicted as a cartoon, water oxygens are represented by small red spheres, counter-ions are drawn in blue spheres.

in the X-ray structure of BACE (pdb code: 1FKN, X-ray structure:1.9 Å resolution [60]). One of the two catalytic aspartates (Asp32) was protonated, following Ref. [149], whilst other ionisable groups were assigned their charged forms. Histidines were protonated accordingly to their putative H-bond pattern in the X-ray structure.

MD simulations The protein was immersed in a water box of size 75 Å x 87 Å x 90 Å. Charge neutrality was accomplished by adding 9 sodium counter-ions (fig. 3.8). The whole system, composed of about 48,000 atoms, underwent 20 ns of MD at 300 K and 1 atm pressure simulations carried out with the GROMACS program [79, 80]. The Amber force-field (Parm98 [78]) was used to describe the protein and the counter-ions, whilst the TIP3P model was used for water[150]. Particle mesh Ewald routines were used to treat long-range electrostatic interactions [76, 77]. A cut-off of 12 Å was used for the van der Waals interactions and the real part of the electrostatic interactions.

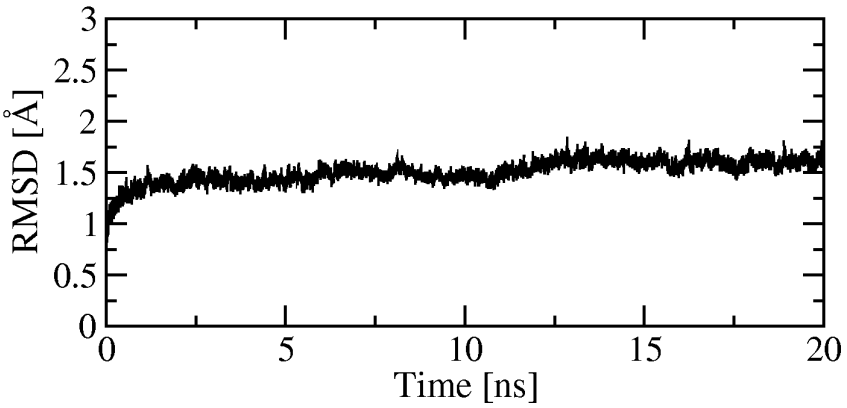


Figure 3.9: The rmsd of the backbone of the protein has been plotted as a function of time.

3.2.3 Structural and dynamical insights

The structure of the Michaelis complex (Fig. 1.10) equilibrates after 0.8 ns, and after this time its rmsd fluctuates around an average value of 1.5 Å (Fig. 3.9). The overall fold of the protein is well maintained. The largest changes occurr in the long loop ranging from Val 309 to Asp317, which is also the less rigid region of the protein, as evidenced from rmsf calculations (see. Fig. 3.10).

Rmsf data are in good qualitative agreement with experimental B-factors, as evidenced in Fig. 3.10. The main discrepancy is in the Thr254-Pro258 loop, which, in MD simulations, is found more mobile than in experimental B-factors. This discrepancy may be attributed to crystal packing effects, as the PDB file evidences that this part of the protein surface is closer than 3 Å to the other image in the crystal.

The rmsf plot evidences that the protein core is rather rigid, whereas hydrated loops are the most mobile regions. The substrate is also allowed to fluctuate in the binding cleft.

3.2.4 Essential motions of β-secretase

In order to rationalize motion of the substrate in the binding cleft, essential modes of the Michaelis complex have been computed from covariance matrix (see section 2.3.2).

The plot of the eigen-values of the covariance matrix denotes that the first ten slowest frequency eigenvectors are sufficient to describe most of the motion of BACE. In fact, the amplitude of the eigen-vectors drops significantly after the third one (see

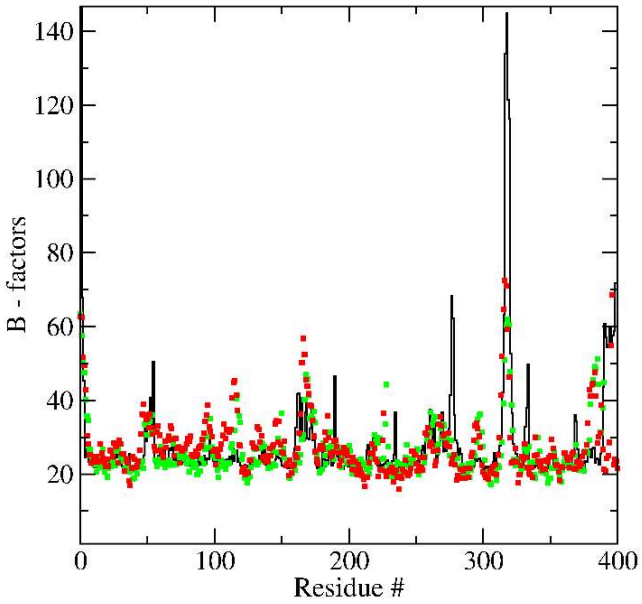


Figure 3.10: Computed B-factors (black line) are matched with experimental values coming from the two independent images in the X-ray crystal [60].

fig. 3.11a).

The second eigenvector describes movements of the hydrated loops, leaving the core of the protein rather immobile. Projection of this mode onto the MD trajectory has a quite high cosine content value (see fig. 3.11b), therefore, this mode should not represent a real "coherent" motion, but, rather, a random diffusion of less structured parts of the protein.

The first and the third eigenvectors, on the contrary, are characterized by low cosine contents (fig. 3.11b), implying that these ones should be *true* over-damped vibrational modes of the protein, and thus, any functional mechanics of the protein may be related to these motions.

Both eigenvectors represent *quasi*-rigid rotations of the two lobes toward each other (see fig. 3.12), in qualitative agreement with the rotation of the C-lobe found in the 3D structure of AP from *Plasmodium Falciparum* [151], and both modes affect the relative distances of the C_αs in the active site (see tab. 3.1).

During both modes, a structurally conserved region, namely, the four anti-parallel β-sheets located opposite the substrate binding pocket, with respect to the cleavage site (see fig. 3.13), remain fixed. In particular, the largest displacement between the

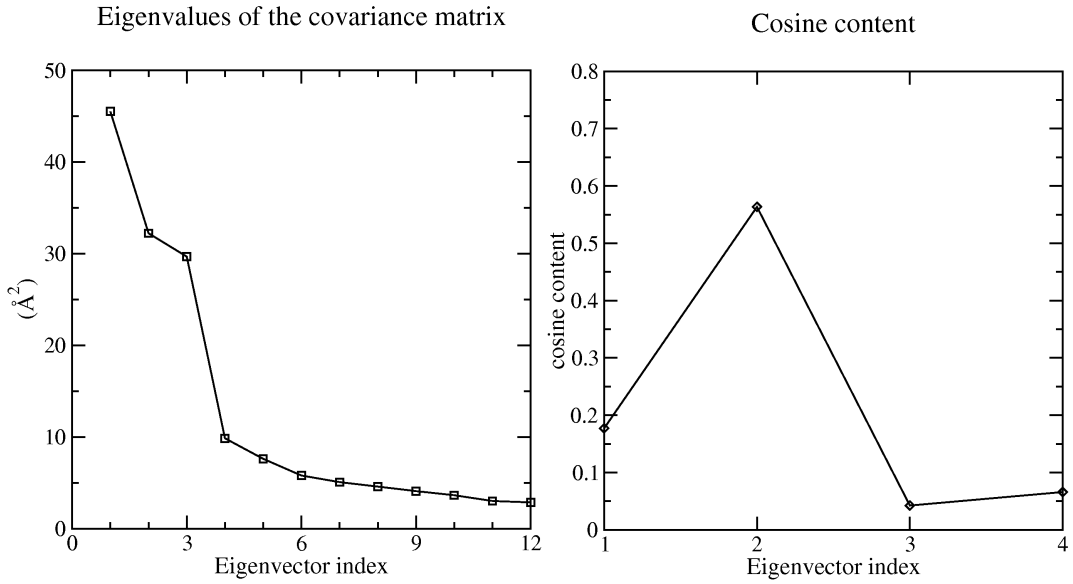


Figure 3.11: On the left, the largest eigen-values of the covariance matrix are plotted; on the right the cosine content of the first 4 eigenvectors is reported.

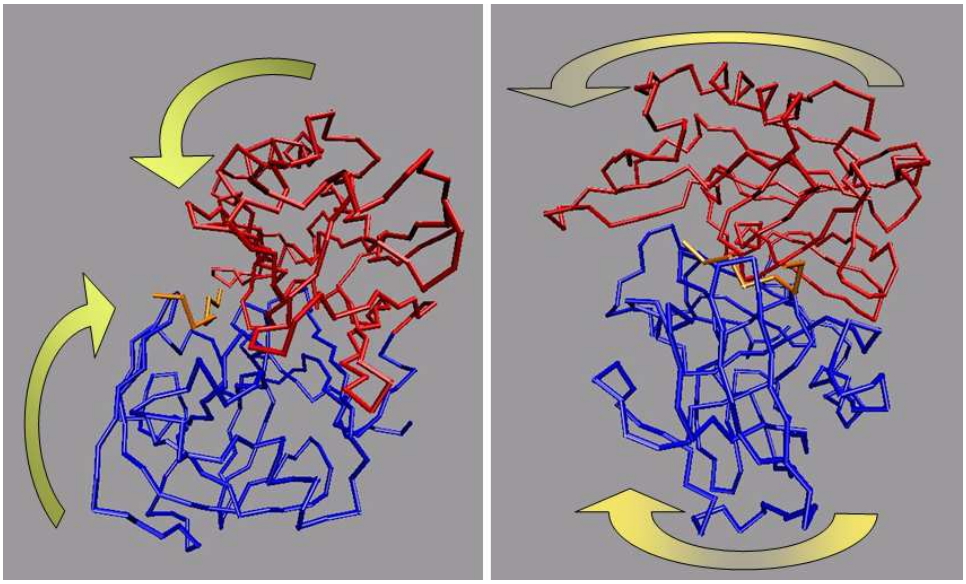


Figure 3.12: Schematic representations of the first and third eigenvectors. The N- lobe is coloured in blue, the C-lobe in red, the substrate in orange. Yellow arrows describe the relative rotations of the two lobes.

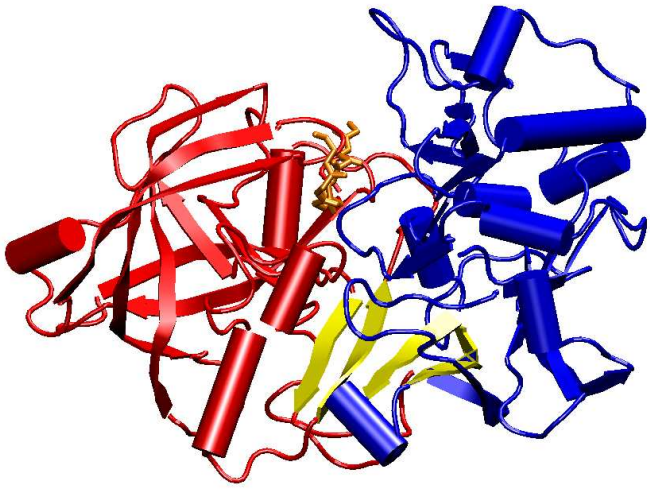


Figure 3.13: Cartoon representation of BACE. The hinge β-sheets that cross-link the two lobes are evidenced by a yellow color.

Mode	Asp32-Leu	Asp32-Ala	Asp228-Leu	Asp228-Ala
1	0.51	1.05	-0.36	0.08
3	0.77	1.31	-0.46	-0.09

Table 3.1: Variation Δ (in Å) of the relative distances in the active site, along the first and third eigenvector of the covariance matrix.

ends of the external sheets is ≈ 0.3 Å. Therefore, this region, which is the only structured region that links the two lobes of the enzyme, can be recognized as the hinge for these two modes.

The analysis of the covariance matrix shows that the substrate motion is highly correlated to its nearest neighbours, and particularly with the cleavage site and the turn of the substrate-binding flap (see fig. 3.14).

Correlation of the substrate with this region is guaranteed by the presence of a conserved tyrosin (Tyr71), which is is thought also to play a key role for substrate binding [124]. Tyr71 is located opposite the cleavage site with respect to the substrate, but it is directly linked to it through a buried water (BW) detected in the X-ray structure [60] (Fig. 3.15).

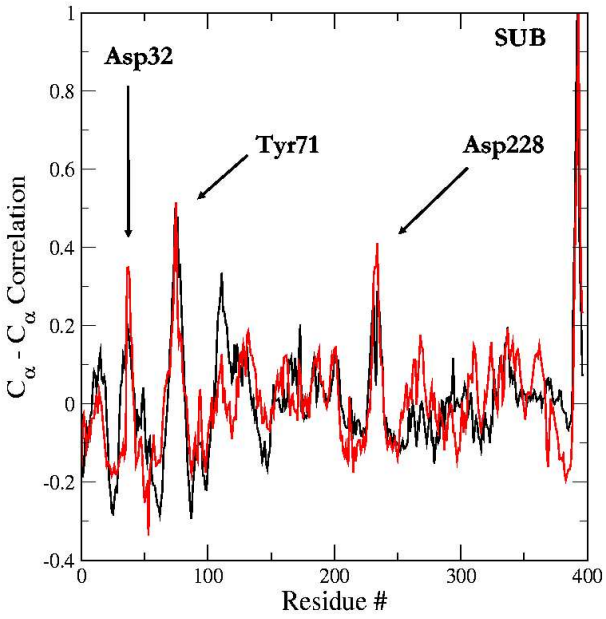


Figure 3.14: Correlation of $\mathrm{Leu_{SUB}}$ and $\mathrm{Ala_{SUB}}$ residues with the protein are drawn in black and red lines. Position of the two catalytic aspartates and Tyr71 are evidenced by black arrows.

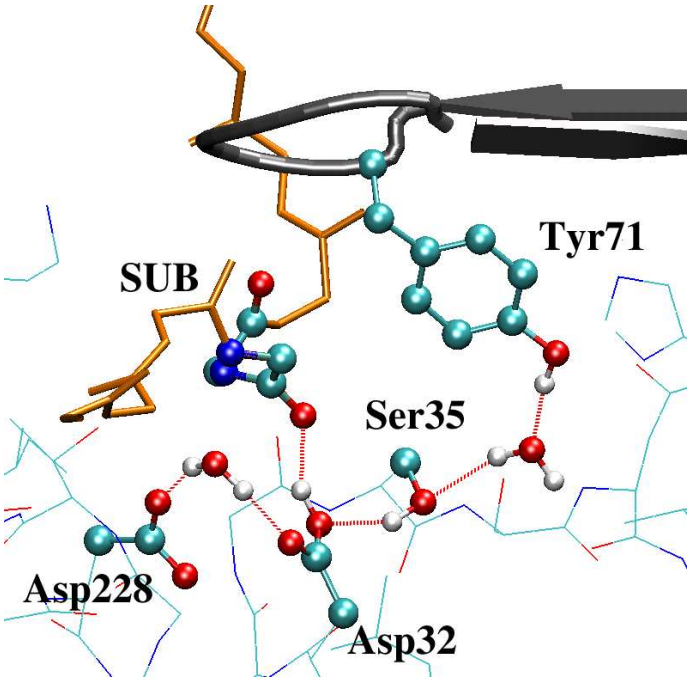


Figure 3.15: Tyr71 H-bond pattern in BACE. Tyr71, Ser35, Asp32 and Asp228, side-chains, an ordered water molecule and the catalytic water molecule are shown in balls-and-sticks. Only hydrogens bound to polar groups are shown for clarity. The substrate backbone is represented in orange cylinders, except for $\mathrm{Leu_{SUB}}$ and $\mathrm{Ala_{SUB}}$ atoms, which are drawn is spheres, the flap is represented by a cartoon. The rest of the protein is drawn in lines. this H-bond pattern is conserved in all pepsins investigated here.

This conserved H-bond pattern provides an efficient way for the enzyme to clamp the substrate into the binding cleft. BW was proposed by Andreeva and Rumsh [46] to have an active role in the enzymatic activity of pepsins, specifically, by stabilizing the intermediate. In fact, BW was thought to undergo conformational changes that could end up, at the TS, into a rotation of the side-chain of Ser35, which should provide an advantageous change in the H-bond pattern in the active site (see fig. 3.16). Indeed, in our simulations we found that BW always donates an H-bond to Ser35, which, in turn, is already bound with its hydroxyl to Asp32. On the contrary, we never found the conformation of Ser35, pointing toward BW. It must be pointed out, anyway, that this conformation should be typical of the free form, rather than the Michaelis complex, and moreover, the overall hydrogen path described by Andreeva and Rumsh is not conserved in BACE, as Trp39 is mutated into an Ala residue.

3.2.5 Conformational changes in the active site

The two modes described in the former section induce relevant variations in the relative distances of the active site. As fluctuations of the substrate in the binding cavity had already been proved to play a fundamental role in the enzymatic activity of HIV1-PR by Piana *et al.* [5, 7], it is interesting to monitor variations of the substrate/aspartic dyad distance during dynamics.

A straightforward way of defining this distance is considering the ξ coordinate, defined as the average of the distances between the C_αs of the aspartic dyad and those of Leu_{SUB}-Ala_{SUB}, as defined in fig. 3.17a.

The plot reported in fig. 3.17b evidences that the substrate fluctuates around two characteristic distances ξ^1=7.7 Å and ξ^2=8.1 Å , which interchange into each other in a timescale of few nanoseconds. Transitions from one to the other ξ value occur very rapidly (few ps), and correspond to conformational rearrangements in the active site: in fact, when ξ=ξ^1, Asp32 H-bonds to Leu_{SUB} carbonyl, while the catalytic water bridges the two aspartates (Fig. 3.18a). At ξ=ξ^2, Asp32 carboxyl group is rotated by 180° and points its hydrogen toward the catalytic water, which acts a H-bond donor toward Leu_{SUB}'s backbone and Asp228 (Fig. 3.18b).

The geometrical configuration of the water molecule seems to be particularly favourable for a nucleophilic attack only for ξ=ξ^1, where its oxygen atom does not receive any H-bond, and it is oriented toward the electrophilic carbon atom.

On the contrary, the geometry in ξ=ξ^2 appear to be not a reactive one, as the nucleophilic attack of water to the carbonyl would require a relevant rearrangement of the groups in the catalytic cavity.

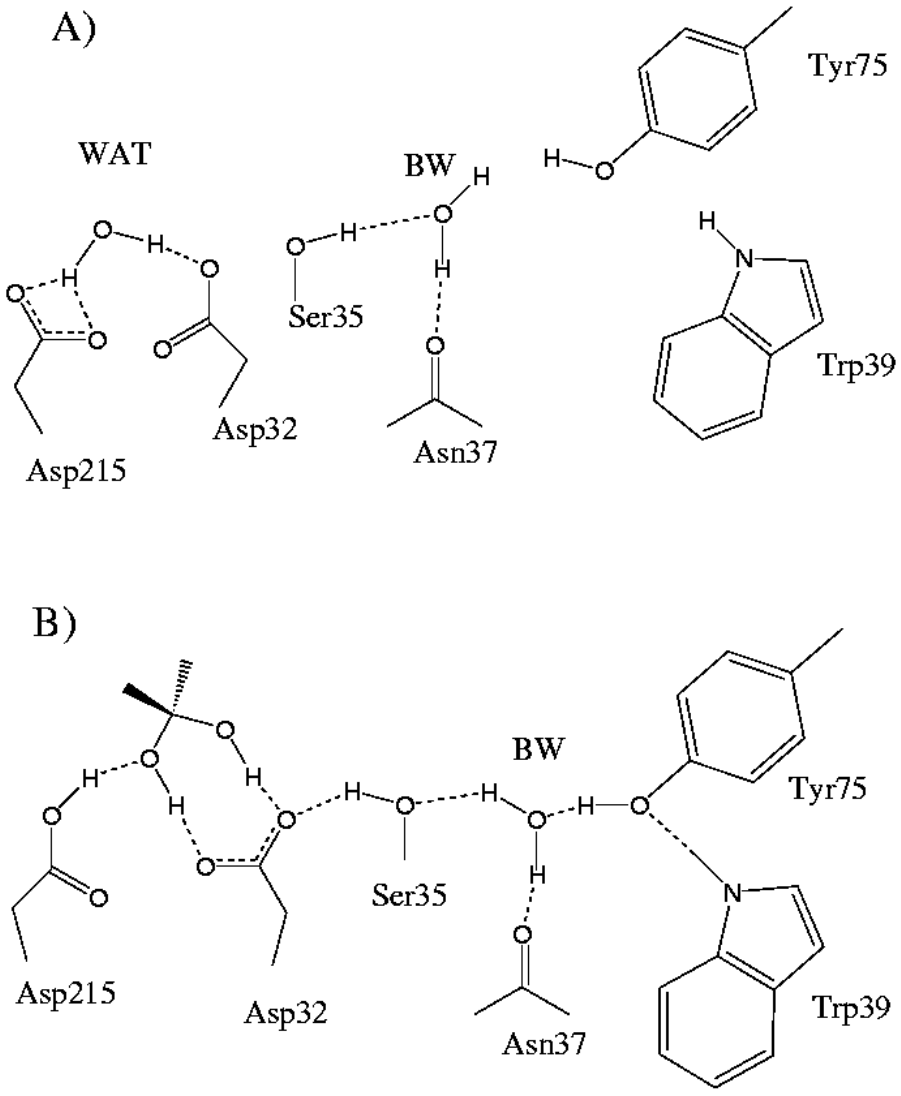


Figure 3.16: H-bond pattern switch mechanism as proposed by Andreeva and Rumsh [46].

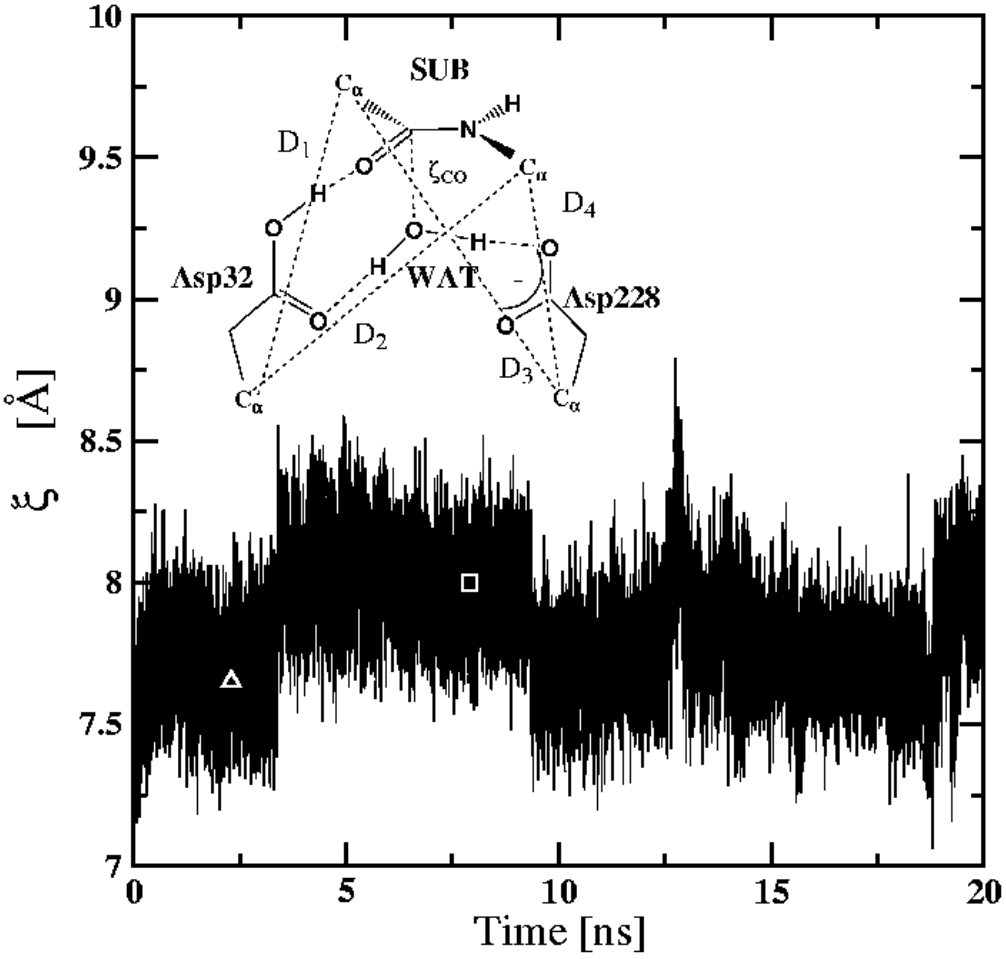


Figure 3.17: Fluctuation of the Substrate in BACE active site. ξ distance (ξ= (D_1+D_2+D_3+D_4)/4)[5] is plotted as a function of the simulated time. The white triangle and square represent the two snapshots used as starting point for the QM/MM simulations, at ξ= ξ^1=7.7 Å and ξ^2=8.1 Å , respectively.

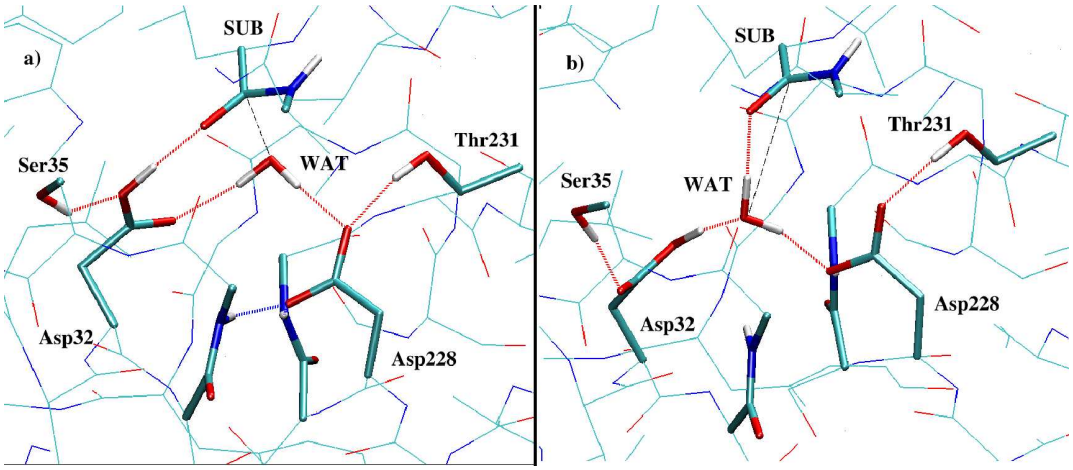


Figure 3.18: BACE active site geometries. Configurations at ξ^1 (a) and ξ^2 (b). Atoms drawn in cylinders are treated as quantum atoms during the QM/MM calculations. Hydrogen bonds are drawn in red and blu lines. Only hydrogens bonded to polar groups are shown for the sake of clarity.

3.3 Enzymatic reaction

A quantitative measurement of the difference in reactivity between the two conformations of the active site has been obtained by coupling hybrid QM/MM molecular dynamics (see section 2.4) and constraint dynamics (section 2.5.1).

3.3.1 Simulation setups

Two selected classical MD snapshots, in which the distance between substrate and cleavage site (see fig. 3.17) was 7.6 and 8.0 Å , respectively, were used as starting point for QM/MM simulations. The system was partitioned into: (i) a QM region, comprising Asp32 and Asp228 side-chains, the Leu-Ala substrate peptide, the catalytic water and the groups that share H-bonds with the Asp-dyad, namely, Ser35 and Thr231 side chains, and Thr33-Gly34 and Ser229-Gly230 peptide backbones (Fig. 3.18); (ii) a MM region, comprising the rest of the system.

The free-energy of the reaction was computed by thermodynamic integration [152], using as approximate one-dimensional reaction coordinate ζ_{CO} the distance between the oxygen atom of the catalytic water and the carbon atom of the peptide bond of the substrate (Fig. 3.18). For both systems, ζ_{CO} was fixed at increasingly shorter distances, from the equilibrium distance up to formation of the C-O bond. For each

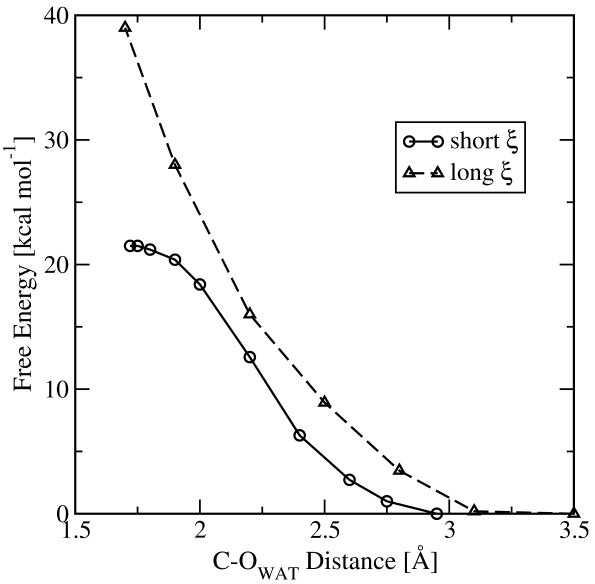


Figure 3.19: Free energy profiles for the first step of the enzymatic reaction of BACE. The free energy has been computed at ξ=ξ^1 and ξ=ξ^2 distances.

computed distance, about 2 ps of QM/MM simulations were carried out, in order to achieve an acceptable convergence of the force on the constraint.

3.3.2 Results

At ξ= ξ^1, the free energy barrier for the first step is around 20 kcal mol^{-1} (see fig. 3.19), in fair agreement with the experimental barrier of the reaction ($\approx$ 18 kcal mol^{-1}) [36], and similar to the one computed in the HIV-1 AP with the same procedure (18-20 kcal mol^{-1} [5, 7]).

During the QM/MM simulations, the H-bond pattern found in the initial state is conserved while shortening ζ_{CO}.

The water molecule is not polarized by the environment, and it does not change its electronic properties until the TS is reached. In fact, even at $\zeta_{\mathrm{CO}} = 1.75$ Å , that is, 0.02 Å before the transition state, water is only weakly polarized as its dipole moment is $\approx$ 3.1 D (fig. 3.20).

Figure 3.21 reports the polarization of the O-H bonds during the constraint dynamics run at $\zeta_{\mathrm{CO}} = 1.75$ Å . This plot remarks that these bonds are on average not polarized yet. Nonetheless, the H-bonds with the the two aspartic acids are at this point strong enough to allow instantaneous transfer of the protons from the water

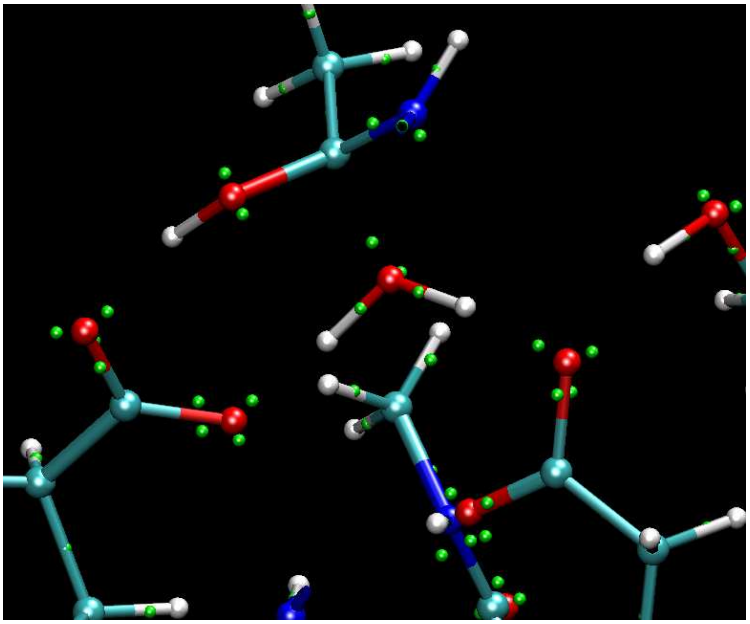


Figure 3.20: Active site of BACE at $\zeta_{CO} = 1.75$ Å . Wannier Centers are drawn in small green spheres.

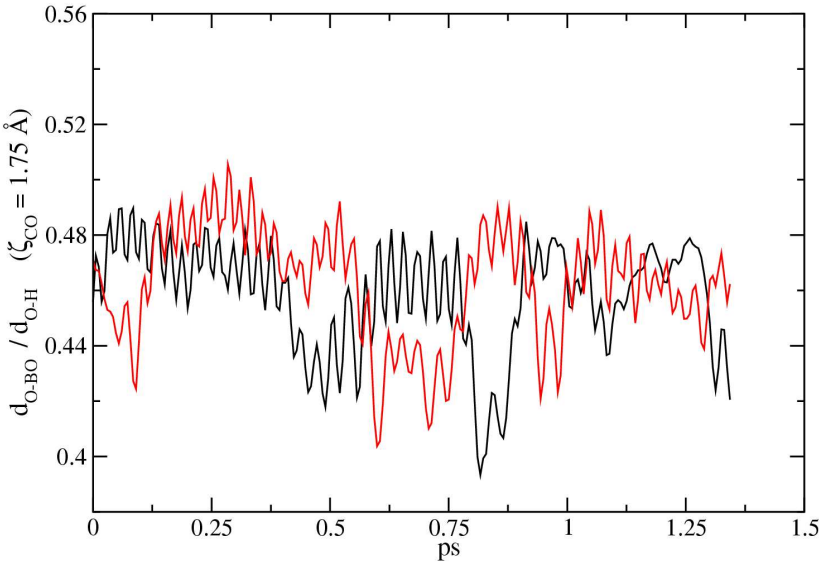


Figure 3.21: Polarization of the OH bond of WAT during MD run at ζ_{CO}, computed as ratio between the distances of the oxyen atom and the corresponding Wannier center or the H atom.

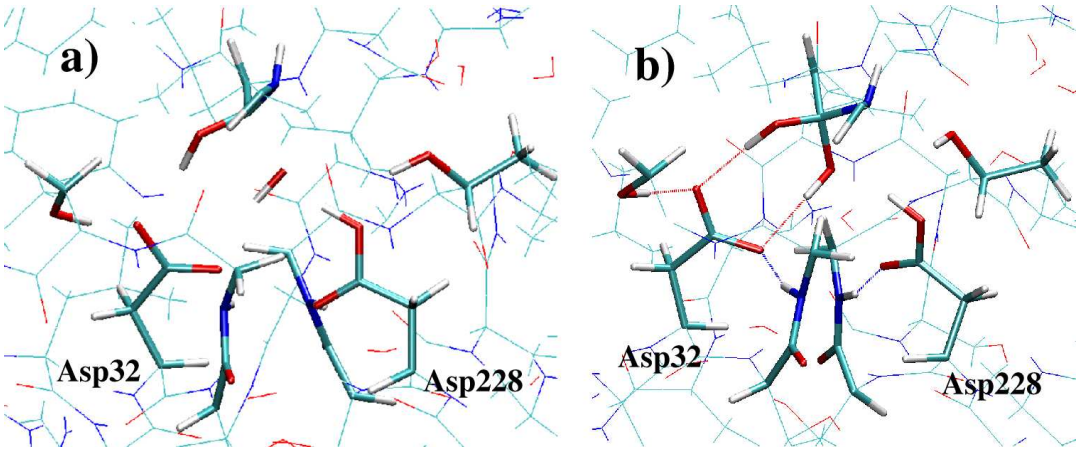


Figure 3.22: BACE TS (a) and intermediate (b)geometries. Atoms drawn in cylinders are treated as quantum atoms during the QM/MM calculations.

molecule to the protein moiety.

At the TS distance ($\zeta_{\mathrm{CO}} = 1.73$ Å , see fig. 3.22a), the O-H bonds of the catalytic water are elongated, with an average distance of 1.1 Å; in particular, one of the two protons is shared with Asp228, (fig. 3.23). Thus, BACE catalyzes the reaction by ensuring a sufficiently strong base to be able to deprotonate the nucleophilic water at the TS.

Another relevant aspect of proton dynamics in the reaction refers to the H-bond between Asp32 and $\mathrm{Leu_{SUB}}$. In fact, Asp32 transfers its proton to $\mathrm{Leu_{SUB}}$'s carbonyl at $\zeta_{\mathrm{CO}} \approx$1.9-2.0 Å, well before the transition state (TS) is reached. The protonation of the substrate carbonyl has the relevant effect of increasing the electrophilicity of the substrate, thus, it positively contributes to the catalytic action. Early protonation of the carbonyl represents also a major difference with reference simulations in aqueous solutions, where the carbonyl was protonated in neither of the simulations also after formation of the intermediate.

The deviation from co-planarity of the Asp dyad is more pronounced than in other ligand-bound pepsin structures [60], as a result, both Thr33-Gly34 and Ser229-Gly230 backbone amide groups point to Asp228's carboxyl, while Asp32 is only H-bonded to Ser35 side-chain. Nonetheless, after deprotonation, Asp32 side-chain rotates and recovers co-planarity with Asp228 side-chain, re-gaining the H-bond with Thr33-Gly34 backbone as well. This event results into a decreasing of the number of hydrogen bonds donated from the protein to side-chain of Asp228; thus, it might also have some positive effect on catalysis, as decreasing the number of the received

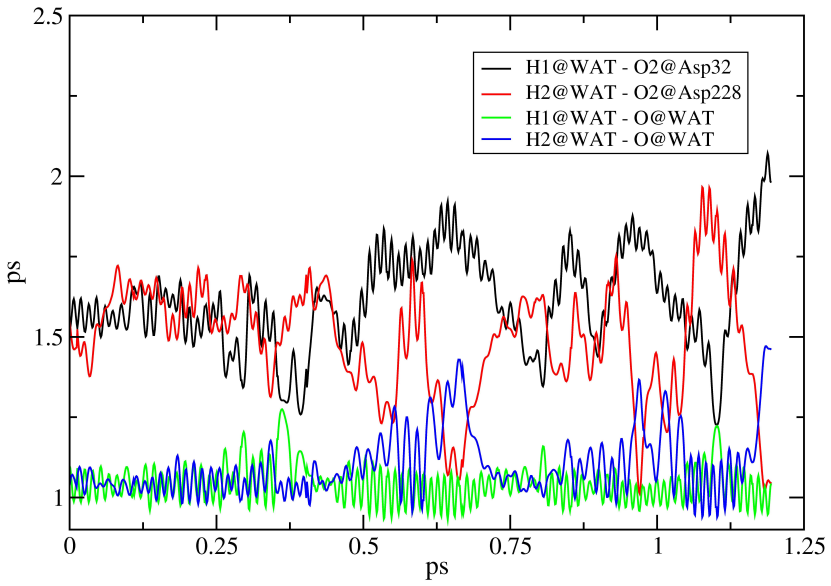


Figure 3.23: Distances between WAT protons and carboxylic oxygens of Asp32 and Asp228 at ζ_{CO}=1.72 Å (TS distance).

hydrogen bonds might increase the basicity of Asp228.

The geometry of the intermediate can be obtained by releasing the constraint after the TS. In this case, the intermediate is a gem-diol that H-bonds with its two hydroxyl groups to Asp32 (see fig. 3.22b).

The conformation associated to ξ=ξ^2 value is found to be non-reactive (ΔG >40 kcal mol^{-1}, see fig. 3.19). In particular, as ζ_{CO} become as short as 2.5 Å , Asp32 H-bonds to Asp228 and does not interact anymore with the water molecule (see fig. 3.24). This conformation, characterized by the presence of an H-bond shared by the two aspartates, has been found also in HIV1-PR, and also in that case it has lead to a non-reactive profile [5].

When $\xi = \xi^2$, on the contrary, the aspartic dyad is not able to interact with the substrate, and when constraining the catalytic water towards the carbonyl, it forms an H-bond between the two aspartates. The formation of this bond has the effect of damping all the catalytic properties of the aspartic dyad, as, in this conformation, it is not able to enhance the electrophilicity of the substrate, and at the same time, crucially, it lowers the basicity of Asp228, thus, hindering deprotonation of the catalytic water at the TS.

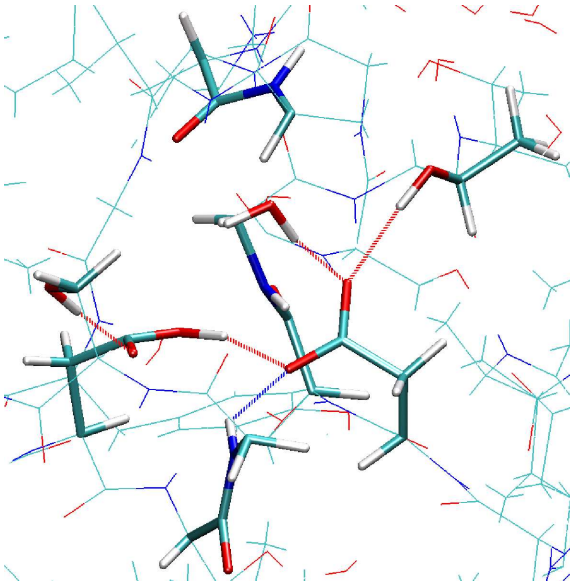


Figure 3.24: Geometry of the Active Site at $\xi = \xi^2$ and ζ_{CO} = 2.5 Å . Quantum atoms are drawn in cylinders.

3.4 Summary

Our calculations suggest that BACE explores different geometric conformations of the active site that are able to enhance or suppress its catalytic power. This efficient way of modulating BACE enzymatic activity is achieved by thermal fluctuations of the whole protein structure. Specifically, as oscillations of the substrate in the binding cavity let Asp32-Leu_{SUB} distance reach relatively short values, Asp32 carboxyl and Leu_{SUB} carbonyl share a hydrogen bond. The following rearrangement of the H-bond pattern leads to WAT reorientation into a favourable geometry for nucleophilic attack to Leu_{SUB} carbonyl, with a resulting free energy of $\approx$ 20 kcal mol^{-1}, in fair agreement with experimental data ($\approx$ 18 kcal mol^{-1}) [36].

As anticipated in section 3.1.3, the reaction free energy for the peptide hydrolysis is strongly decreased by the presence of an efficient proton shuttle system, which, in the case of BACE, is formed by the aspartic dyad. In fact, when the substrate and the two aspartates are relatively close, it provides a strong enough base (Asp228), which is able to cleave the O-H bond of water at the TS. Moreover, the formation of a strong hydrogen bond between the other aspartic acid (Asp32) and the carbonyl of the substrate enhances the electrophilic character of the substrate, thus, it favours the stabilization of the TS. The geometrical pre-orientation of water is also a relevant

aspect of the catalysis, which is achieved as a consequence of the hydrogen bond pattern found for $\xi = \xi^1$.

Calculations done in this work support a reaction mechanism similar to that proposed by Andreeva [46] for the nucleophilic attack of the activated water and the proton dynamics in the active site. On the contrary, while Andreeva proposed a local triggering effect mediated by a buried water molecule (BW), we cannot identify such a subtle effect. While it has to be underlined that a mutation in a possibly non irrelevant position for this mechanism occurs in BACE, MD calculations suggest that reactive conformations of the active site are dynamically stabilized by global motions of the whole protein scaffold.

The reaction mechanism described here is similar to that found in HIV-1 AP with a similar setup ref. [5, 7], being the main difference on protonation the substrate carbonyl. In fact, in that case the proton hopping from Asp32 to the substrate's carbonyl is concerted, as it occurs at the TS.

The evidence that HIV1-PR and BACE share the capability of modulating their enzymatic power by similar mechanical coupling poses the key question of whether this would be a more general feature, evolutionarily conserved in all members of the pepsin and retropepsin families. A possible positive answer to this question is provided through the studies presented in the following chapter.

Chapter 4

Functional Mechanics in Pepsins and Retropepsins

In the last years, the importance of mechanical coupling for enzymatic catalysis has been being unravelled for a large variety of enzymes [153], among which, flavin reductase [8], dihydrofolate reductase [9], acyl CoA dehydrogenase [154], orotidine 5'-monophosphate decarboxylase [155], tyrosyl-tRNA synthetase [156] and xylose isomerase [157], as well as in mitogenic signalling processes [158].

Results presented in the former chapter have showed that also BACE modulates its enzymatic activity by long-range fluctuations of its scaffold. A significatively similar coupling mechanism has been proposed in HIV1-PR by Piana *et al.* [5], and recently confirmed by a variety of studies [6, 49, 7].

The existence of analogous mechanisms in two proteins (BACE and HIV1-PR) belonging to the same enzymatic class, and members of two different families, believed to be evolutionarily related [4, 10], is more than a clue for the hypothesis that this characteristic could be shared by all members of these two aspartic protease families.

The work presented in this chapter tries to shed some light on this fascinating possibility.

First of all, the evolutionary pattern of pepsins is analyzed by means of multiple sequence alignment techniques. Then, mechanical responses of different pepsins are computed by coarse-grained computations. The model for this set of computations, proposed in section 2.6.1, has been first validated by a comparison with results from MD of BACE, then applied to five other pepsins from different organisms.

Dynamical data have been cross-linked with evolutionary information. This operation have lead to determination of structurally conserved regions that have functional

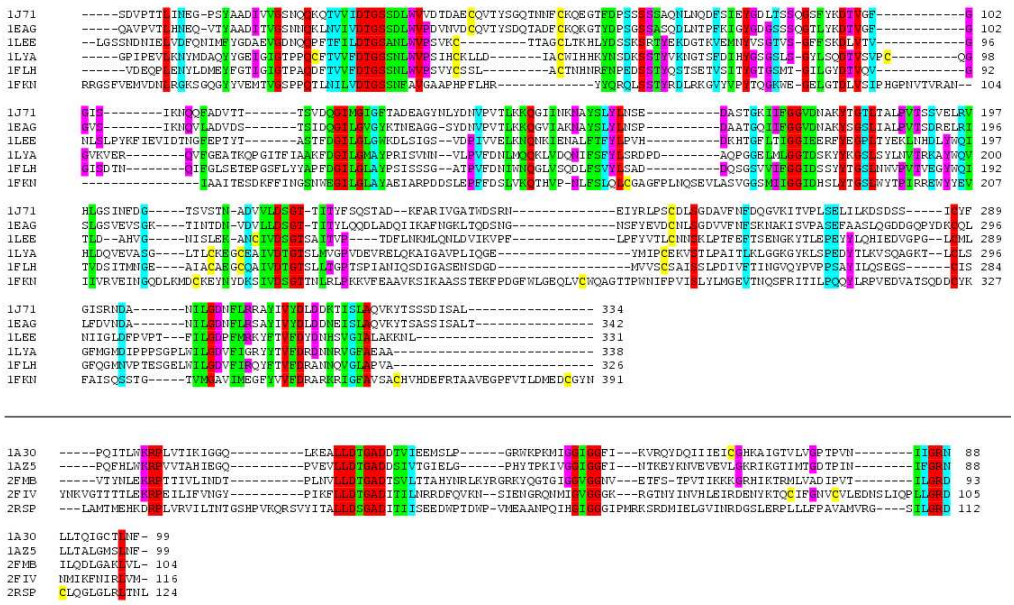


Figure 4.1: Multiple alignment of pesins and retropepsins. (1) Pepsins from candida tropicalis (PDB code: 1J71)[159], candida albicans (EAG)[160], Plasmodium (1LEE)[151] and human cathepsin D (1LYA)[161], uropepsin (1FLH)[162] and BACE (1FKN)[60], and (2) Retropepsins from Human immunodeficiency virus 1 (1A30)[125], simian immunodeficiency virus (1AZ5)[163], equine anemia virus (2FMB)[164], feline immunodeficiency virus (2FIV)[165] and rous sarcoma virus (2RSP)[166] are shown. Conserved residues are coloured in red, conservative mutations in green, blue and magenta, cysteins in yellow.

roles in the protein mechanics. A particular attention has been dedicated to the flap mechanics, which is not easily characterizable by standard MD, as its typical times are too long as compared to affordable simulation times.

Finally, the same analysis has been repeated for retropepsins, underlining their similarities and differences with pepsins.

4.1 Multiple alignment of pepsin sequences

The alignment of the pepsin family has been obtained by a Hidden-Markov-Model (HMM) in the sequence database (see sec. 2.7.2), which ended up with a final number of 868 aligned sequences. As a sample, in fig. 4.1 six of these sequences have been reported. As already stated in chap. 1.1.1, very few residues are conserved or show

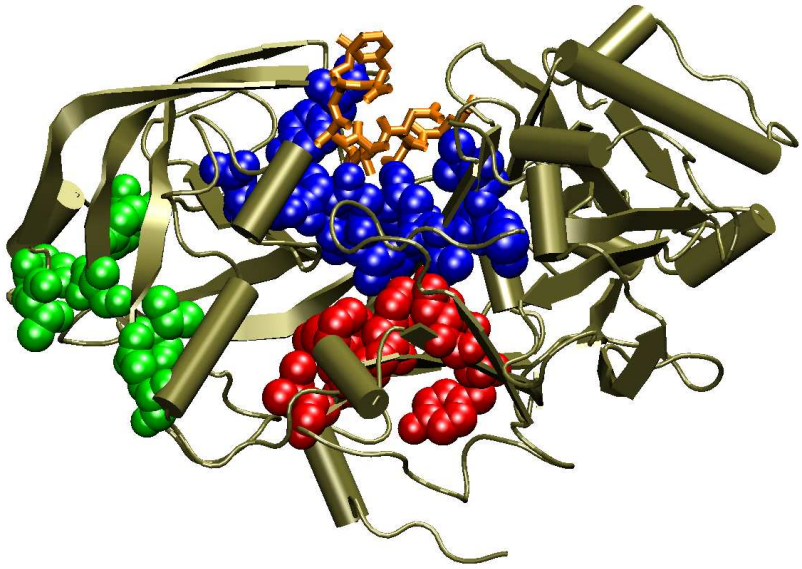


Figure 4.2: The structure of BACE is shown in complex with model substrates (orange sticks). The conserved regions are drawn as spheres: the active site in blue, the cross-linking β-sheets in red, and the surface region in green.

a high homology throughout the whole family. Only very few continuous pieces of the sequences are conserved. These aa stretches are mainly localized in the core of the protein, corresponding to the active site loops (residues number:30-38, 226-231, BACE sequence) and to its first-neighbours (residues: 122-129). In other regions conserved residues show a binary periodicity, typical of conserved β-structures, or are scattered along the sequence.

4.1.1 Conserved regions in pepsins

At first glance, matching amino acid conservation on the primary structure would suggest that AP have conserved only the active site region, the other few conserved residues are more or less randomly localized in the rest of the structure. On the contrary, localization of conserved amino acids in the 3D structure leads to the evidence that more than 80 % of the conserved residues are clustered in three different regions (see fig. 4.2):

- *active site region*. This region comprises the active site loops and their first neighbours. All residues are localized at the bottom of the ligand-binding cleft.

The only exception is represented by Tyr71, which is located on the flap, and whose functional role has already been discussed in sec. 3.2.4.

- *lobe cross-linking β-strands*. This structurally conserved region is conserved as well in aa sequence. Molecular dynamics calculations indicate this region as the hinge in functional fluctuations, (see sec. 3.2.4). The conserved residues are all hydrophobic (Leu, Ile, Met and Phe), in fact, their side-chains provide a compact hydrophobic packing in the protein core. Residues in the β sheets whose side-chains point outside the protein body are allowed to have a larger mutability.

- *residues at the flap ends on the surface of the N-terminal lobe*. A conserved contact pattern correlates some residues localized on the surface of the N-terminal lobe, at the flap Ends. Unlike the two other conserved regions, no direct correlations between it and the enzymatic mechanics/activity could be directly found during MD simulations. A deeper discussion on this region will follow in sec. 4.3.

The other few conserved residues which do not belong to neither of these regions are mainly glycines in turns or in bent β-sheets, therefore their conservations should be related to restrains in protein folding dynamics or in native state geometry stabilization.

4.1.2 Dynamical characterization of conserved residues

In order to get insights on the eventual correlations between protein mobility and conserved regions, residues belonging to these parts of the protein have been evidenced on the RMSF plot obtained from MD simulations of BACE (fig 4.3).

As expected, conserved residues in the active site and in the cross-linking β-sheets are all characterized by a low mobility. Rigidity of these two regions can be easily put into relation with their functional role. In fact, rigidity of the active site is known to be a necessary and general feature, as they are locations where there occur processes at relatively high energy, and therefore, they should not allow dissipation of energy, e.g. through conformational relaxations. The conserved stiff hydrophobic packing in the four beta-sheets, on the other side, plays at the same time a structural role in the stabilization of the bilobal fold, and a dynamical one, providing the fulcrum for functional relative rotations of one lobe toward the other. Most interestingly, also the third conserved region shows a relatively low mobility, although this feature cannot

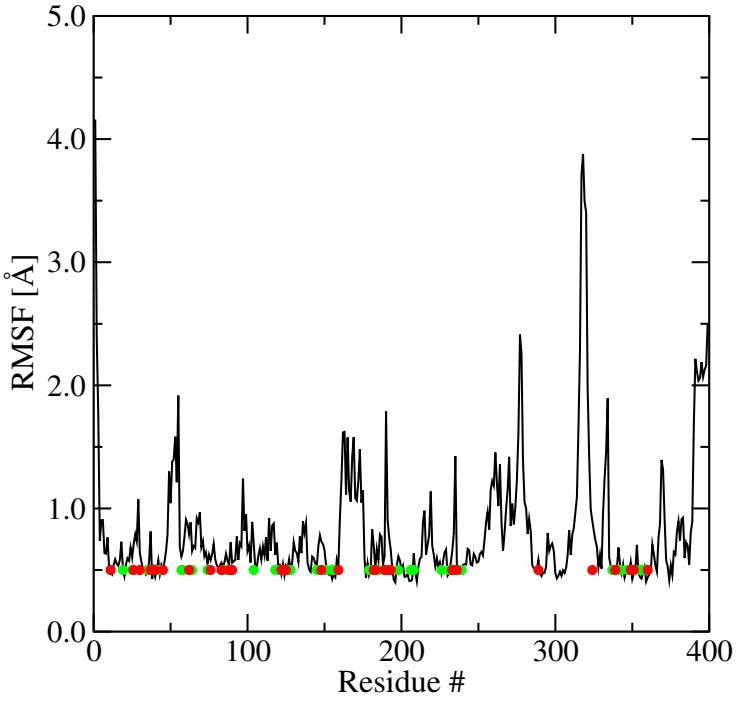


Figure 4.3: Root-mean-square fluctuations of BACE. Fully conserved or highly conserved residues are evidenced by red or green spots, respectively.

be easily related to some functional mechanics for the catalytic action of the enzyme, due to its peripherical position.

4.2 Coarse-grained computations

At least two of the three conserved regions in the pepsin family have found to play a direct functional role in BACE enzymatic activity, and are characterized by a low mobility. Also the other region is peculiarly characterized by a relatively low rmsf; therefore, correspondence between parts of the protein that are conserved in sequence and rigidity of the scaffold might occur not just by chance in BACE, but it might be a conserved property in all members of the pepsin family.

In order to get evidence of this hypothesis, a wide screening on mechanical fluctuations of different pepsin structures could be done, by taking advantage of a fast and yet reliable coarse grained mechanical model, as the β-Gaussian model described in sec. 2.6.1.

4.2.1 β-Gaussian model of BACE

First of all, the model has been tested on BACE, in order to get some insights on its reliability for this kind of folding. As reported in sec. 2.6.1, both covariance matrices and temperature factors can be computed within this model, and thus, can be compared to those obtained by MD simulations (fig. 4.4).

First, we compare the normalized covariance matrices, defined as:

$$\overline{C}_{ij} = \frac{\langle \mathbf{r}_i \cdot \mathbf{r}_j \rangle}{\sqrt{\sigma^2_{\mathbf{r}_i} \sigma^2_{\mathbf{r}_j}}}$$

This quantity has the advantage of being dimensionless, and therefore more convenient in a comparison between atomistic and mesoscopic models. The scatter plot in fig. 4.5 summarizes the degree of accord between the normalized covariance matrices of the MD and β-gaussian model. To avoid artificial biases in the correlation, the diagonal elements of the normalized matrices (by definition, all equal to 1) have been omitted from the plot.

The linear correlation coefficient among the two set of entries is as good as 0.84, and the interpolating line lies very close to the diagonal, with a slope of s= 1.02. It has to be evidenced, nonetheless, that the distribution is only approximately bi-normal, and a certain scatter in the points is present.

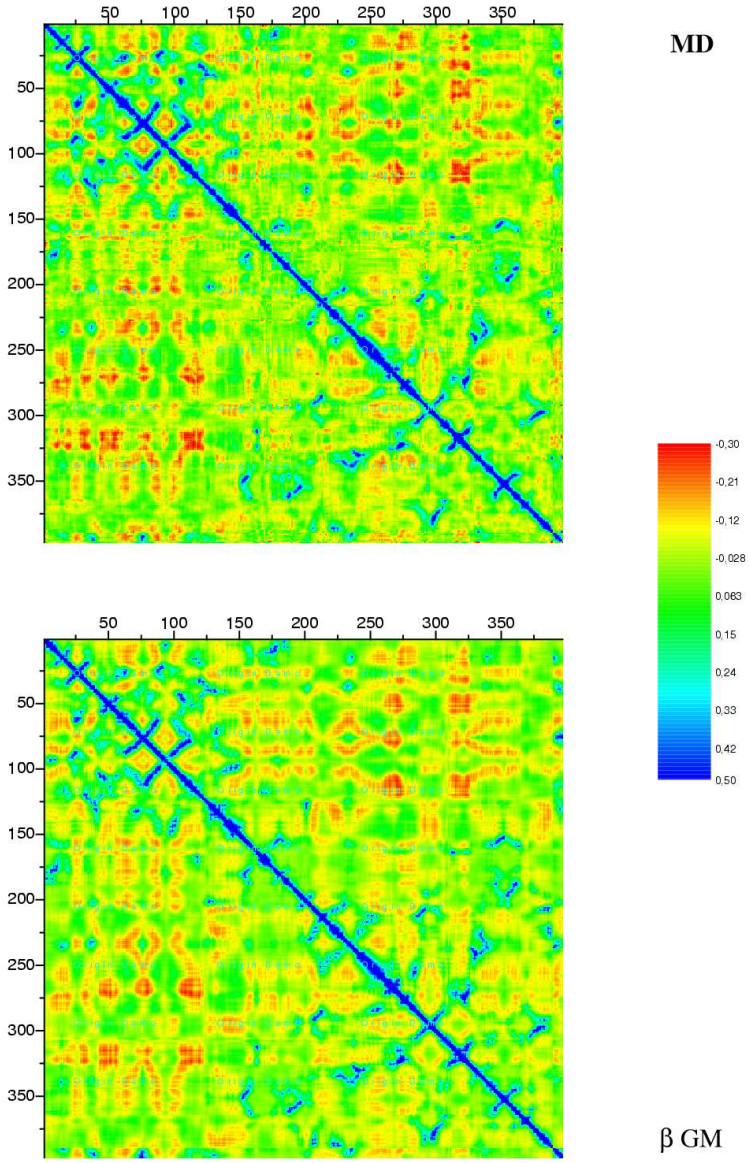


Figure 4.4: covariance matrices for BACE, as obtained from MD simulations (A) or coarse grained computations (B) are shown. Different colors correspond to different values in the matrix entries.

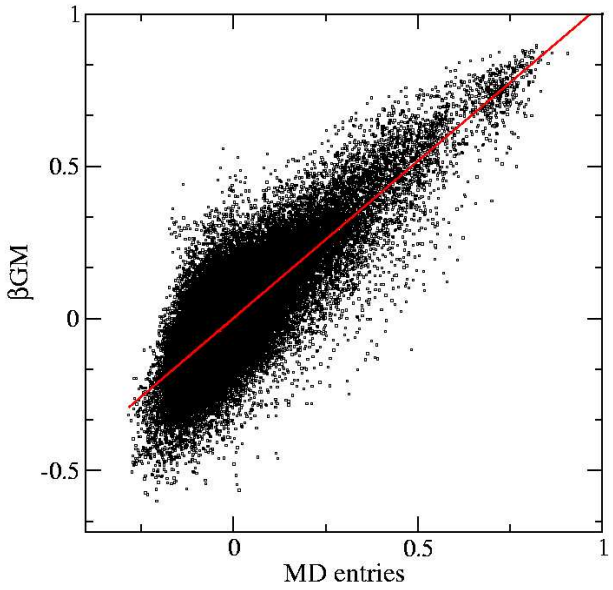


Figure 4.5: Scatter plot of corresponding entries of the covariance matrices obtained by β-GM and MD simulations of BACE. The red line represent the linear regression.

The diagonal elements of the non-normalized covariance matrices have the same meaning of rmsf in MD simulations, and therefore can also be compared. Figure 4.6 shows that indeed, a good agreement between the mobility of the amino acids found in MD and in Coarse-grained computations is found.

As already underlined in secs. 1.1.1,1.2.4, the number and positions of SS bridges in pepsins, and in BACE in particular, have undergone modifications. In the plot reported in fig. 4.6, rmsf for BACE have been computed also switching on explicit disulfide interactions in the model. Indeed, the presence of S-S bond interactions does not affect qualitatively the low-frequency large scale dynamics around the folded state. Thus, their presence might impact more strongly on the folding dynamics or native-state stability rather than the near-native vibrational properties.

4.2.2 Mobility of pepsins

Five randomly chosen pepsins from different organisms have been chosen (Pepsins from candida tropicalis (PDB code: 1J71)[159], candida albicans (EAG)[160], Plasmodium (1LEE)[151] and human cathepsin D (1LYA)[161] and uropepsin (1FLH)[162]), and fluctuations of their scaffolds have been computed by coarse-grained model and

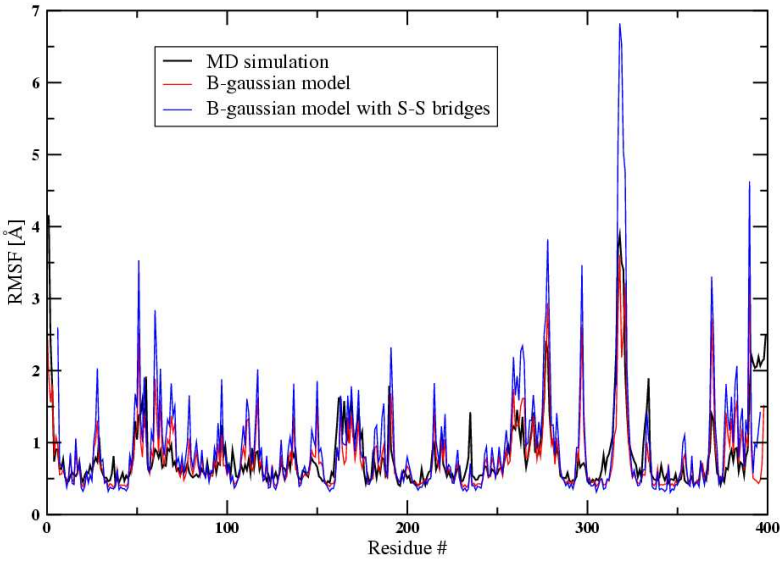


Figure 4.6: rmsf of single residues computed from MD simulation (black line), from B-Gaussian model (without explicit S-S bridges interactions) (red line) and from B-Gaussian model with explicit S-S bond interactions (blue line).

compared to those of BACE. Remarkably, aligned residues among these pepsins exhibit similar mobility properties, as evidenced in fig. 4.7a. In particular, highly conserved regions are rather rigid, while nonconserved regions show a larger mobility.

Visual inspection of the lower-frequency eigenvectors from coarse-grained computations evidence that slowest motions imply relative rotations of the N- and C-lobes also in these other structures. Therefore, rigidity of the active site and the cross-linking β-sheets seems to be a conserved feature that plays a key role in ensuring the correct mechanics between the two lobes of the enzymes, and thus, providing an efficient way of modulating enzymatic a activity in all members of this AP family.

4.3 Flap mechanics

The conserved region at the basis of the flap is not involved in the relevant large scale motions of the enzymes and therefore might not play a significant role for the enzymatic catalysis itself. This region, located roughly perpendicular to the flap ends, comprises mostly polar residues, displaying specific contacts on the protein surface.

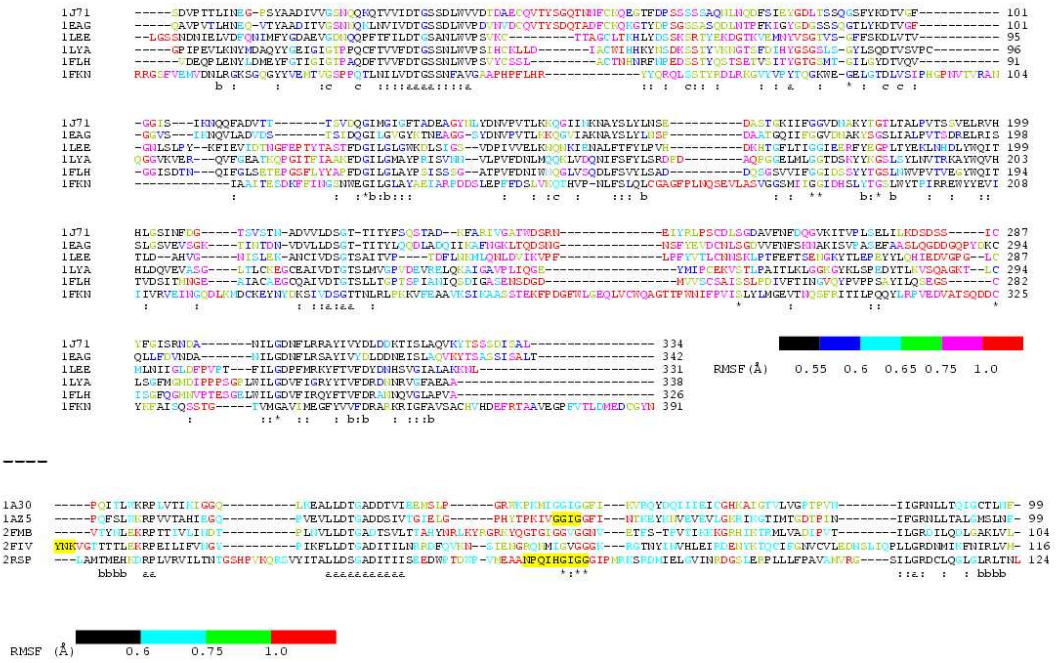


Figure 4.7: Alignment of pepsins and comparison with residue mobilities. (A) Pepsins from candida tropicalis (PDB code: 1J71)[159], candida albicans (EAG)[160], Plasmodium (1LEE)[151] and human cathepsin D (1LYA)[161], uropepsin (1FLH)[162] and BACE (1FKN)[60]. Conserved residues belonging to regions the three conserved regions are marked by a, b, c, respectively; conserved residues not belonging to these regions by asterisks, and conservative mutations by columns. Residues are coloured according to their root-mean-square-fluctuations (RMSF), based on the MD simulations (BACE) or coarse grained computations (all the others). (B) Retropepsins from Human immunodeficiency virus 1 (1A30)[125], simian immunodeficiency virus (1AZ5)[163], equine anemia virus (2FMB)[164], feline immunodeficiency virus (2FIV)[165] and rous sarcoma virus (2RSP)[166]. Conserved residues in regions structurally homologous to the first two conserved regions described in pepsins are labelled by a, b, respectively; conserved residues not belonging to these regions by asterisks, and conservative mutations by columns.

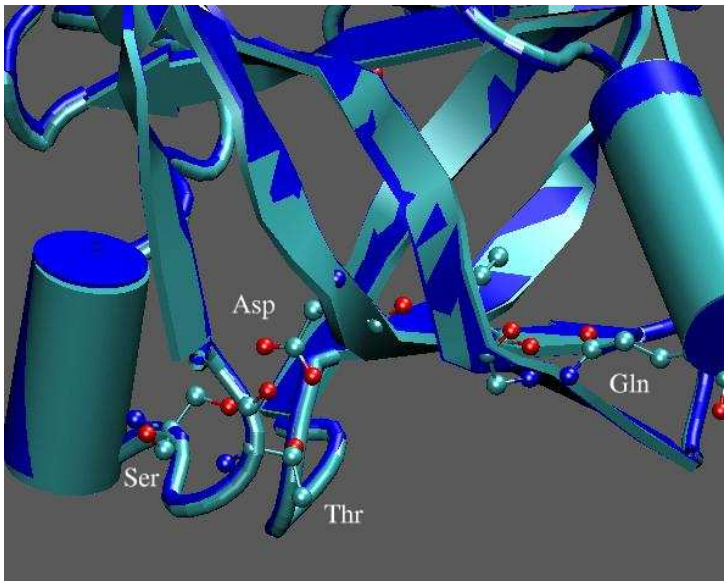


Figure 4.8: Conserved regions at the N-terminal surface. The open and close structures of cathepsin D are superimposed. In particular, the Ser-Thr-Asp pattern at the basis of the flap has been evidenced in a ball-and-sticks draw.

In particular, conserved Ser57, Thr59 and Asp83 (Fig. 4.8), H-bonded to each other, are located, respectively, in the loop preceding the beginning of the first β-sheet of the flap, and at the other β-sheet end. Because of its topological position , it is reasonable to assume that this conserved region might have a role for the flap opening functional mechanics, which, in turn, plays a crucial role for substrate binding [124]. Indeed, the opening of the flap should require major distortions of its geometry, as evidenced by crystallization of the free form of BACE [124] (see fig. 4.9).

The flap opening transition was not observed in our MD, since the relatively high free energy barrier for the close-to-open conformational change of the flap (estimated as $\approx$ 20 kcal mol^{-1} for the HIV-1 isoenzyme [167]) makes this mode unlikely to occur during the typical time-span covered by MD simulations.

Nonetheless, a structural analysis supports this hypothesis: in all pepsin for which the structure has been determined in the free state and in complex with inhibitors [1], flap residues (Val67 to Trp76) rearrange largely (rmsd for single amino acids in the flap up to 4.5 Å)[124], whilst the region under examination does not rearrange

[1]BACE[60, 124], renin [168], endothiapepsin [169, 170], human pepsin [171] and cathepsin D [161]

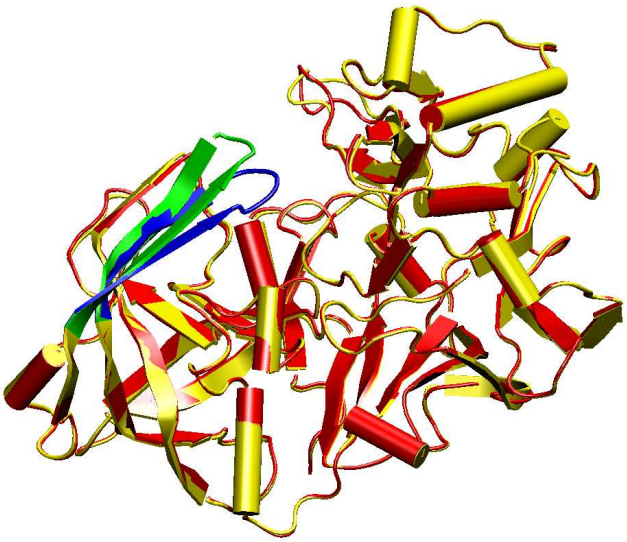


Figure 4.9: Superposition of the BACE structures crystallized with (red) or without a ligand (yellow. Distortion of the flap has been evidenced by colouring it in blue and green for the ligand-bound or unbound structures, respectively.

significantly (rmsd < 0.3 Å , see Table 4.1).

By simply looking at the available crystal structures, the fulcrum of the open-to-close transition could be localized in the kink of the β-sheets that form the flap (see fig 4.9), where a conserved Gly is present (Gly78, BACE numeration). Gly indeed has a major accessibility to different values in the Ramachandran plot, and thus, might work as an efficient hinge for the flap. In this case, the conserved region could act as a rigid cantilever for ensuring the correct bending mode of the β-sheets of the flap. Another possibility is that the conserved region may act directly as the hinge for the flap opening, although this hypothesis would require a larger distortion of the flap with respect to those seen in the available crystallographic structures of free forms of pepsins.

4.4 Insights on retropepsins

Up to now, very few retropepsins have been identified and characterized [27]. Based on this relatively small database, one can establish that only the residues located at the interface, which stabilize the dimeric structure, are conserved [172], as evidenced

Pepsin	Flap	SURF	Total
Renin[a]	1.7 Å	0.2 Å	0.3 Å
Endothiapepsin[b]	0.6 Å	0.1 Å	0.3 Å
Human Pepsin[c]	2.8 Å	0.3 Å	0.7 Å
Cathepsin D[d]	1.9 Å	0.2 Å	0.4 Å
BACE[e]	3.0 Å	0.2 Å	0.6 Å

Table 4.1: Comparison among rmsds between bound and free form of different pepsins in the turn region of the flap, in the region at the surface (SURF), and of the whole structure. PDB codes: [a]: 2REN,1HRN [168]; [b]: 1OEW, 1ENT [169, 170]; [c]: 4PEP, 1PSO [171]; [d]: 1LYA, 1LYB [161]; [e]: 1SGZ[124], 1FKN [60].

in figs. 4.2b,4.10.

The majority of these residues are comprised in the cleavage site loops and their nearest-neighbours, and the terminal antiparallel β-sheets. These regions have been found to show a relatively low mobility, both in MD simulations [5, 6, 49] and by our coarse grained model (Fig. 4.7b). Based on structural [125, 60] and dynamical analysis [5, 6, 49], a structural and functional correspondence between these regions and the regions at the interface of the two lobes in pepsins can be found.

Coarse grained computations suggest that the flaps of the retropepsins (residue numbers 38-60 in the HIV1-AP sequence) seem to be highly mobile, when the enzyme is in the free form. In fact, it has not always been possible to determine their position in the crystals. The flaps in HIV1-AP seem to be rather stiffer the those in other structures. Indeed, the rmsf computations on HIV1-AP refer to a closed conformation of the ligand-bound enzyme [5, 6]. This contrasts the findings in pepsins, where similar mobilities have been evidenced for the flaps in both ligand bound and ligand free enzymes.

Unlike pepsins, there are no conserved residues at basis of the flap, on the protein surface. With this regard, it has to be stressed that in retropepsins this region is less structured that in pepsins. Moreover, structural modifications of the C-lobe of pepsins, with respect to the N- one have produced an asymmetry in the binding pocket and the presence of only one flap. On the contrary, the homodimeric structure of retropepsins leads to a symmetric binding pocket and to the presence of two flaps [1]. The symmetry of the retropepsin fold contrasts with the intrinsic asymmetry of the substrate, and thus, some conformational fluctuations that induce a breaking of the asymmetry upon ligand binding might be necessary. Therefore, it is reasonable to ex-

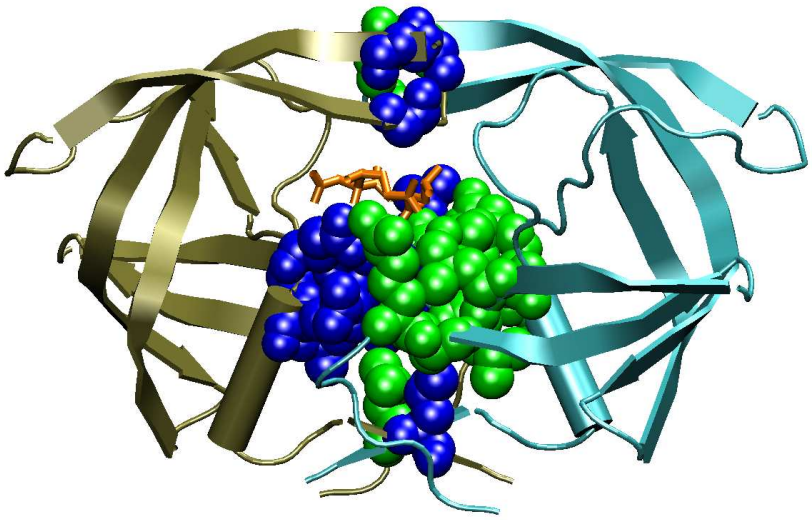


Figure 4.10: The structure of HIV1-AP is shown in complex with model substrates (orange sticks). conserved residues are shown in spheres and coloured in blue and green, according to their subunit.

pect that the mechanisms of molecular recognition and binding between the enzyme and the substrate could be different in the two families, and so, that functional regions implied in the flap mechanics might have been modified in the evolutionary path of the two families.

4.5 summary

Inspection of the alignment of *all* pepsin sequences (and related proteins) present up to date in the database show that, despite the very high sequence variability, at least three conserved regions can be identified. All these three regions appear to be of general relevance in the pepsin family. In fact, 81% of conserved residues in the whole pepsin family are localized in these three regions, whilst only the remaining 19% (mainly involving glycines in loops) do not group into specific regions. These regions have shown a low mobility in BACE MD run, and exhibit low-mobility also in other pepsin structures, as evidenced by coarse-grained computations. Indeed, the low mobility can be related to specific functional roles for the enzymatic catalysis and for substrate binding.

Very interestingly, two of these three regions have structurally conserved related regions in retropepsins, which have been proved to be also characterized by a relative rigidity.

The conserved region of pepsins which has not been found in retropepsins could be related to the substrate binding mechanics, which is expected to be different in the two families, as this process involve regions of the proteins where the major differences in the fold between the two families, like the different number of substrate-binding flaps and the symmetry of the binding cleft, occur.

Concluding remarks

The aim of this work has been to get some general insights on aspartic proteases enzymatic action, from a description of the chemical problem, to the identification of some general feature related to their fold. The chemical reaction has been faced first by exploiting Car-Parrinello / DFT computations on the reference reactions in water. This system has been studied via the recent Multiple-Steering Molecular Dynamics (MSMD) [12, 13], which has provided an efficient tool for determination of the structural, dynamical and energetical features of the raction. In our simulations, it has been possible to identify the de-protonation of the nucelophilic water as the rate-limiting event that avoids reaction to occur at room conditions. In fact, the nucleophilic water molecule yields one of its proton only after reaction has occurred, leading to a quite high free-energy barrier of about 44 kcal mol^{-1}.

Then, the chemistry of the Michaelis complex of a specific pepsin (BACE) by hybrid QM/MM studies has been inferred. We have found that the protein enzymatic action is exploited by providing an efficient proton-shuttle system that enhances both the nucleophilic and electrophilic characters of the catalytic water and the substrate, respectively, efficiently providing the correct protonation states of the reactants along the reaction path. This shuttle system is not present in water, because of its different acid-base properties, as compared to those of the catalytic aspartic dyad, Therefore the free energy of the reaction in aqueous solution is much larger. The free energy barrier in BACE is of about 20 kcal mol^{-1}, in good agreement with experimental data [36], and similar to that found in the reference reaction with OH^- ($\approx$ 18 kcal mol^{-1}). Determination of the TS geometry may help in design of specific drugs for BACE, a key target in Alzheimer's disease pharmaceutical therapy.

Classical MD calculations suggest that mechanical fluctuations of BACE modulate the motion of the substrate in the active site, which, in turn, impacts on the enzymatic activity. An analogous mechanism has already been described in HIV-AP by Piana *et al* [5] and thus could occur also in *all* retropepsins, despite the differences between the folds of the eukariotic and viral enzymes.

The correct functional mechanics of this refined molecular object is achieved by ensuring rigidity in two specific regions, namely, the active site region and the β-sheets that cross-link the two lobes of pepsins. Based on coarse-grained models and sequence alignments of pepsins, it has been possible to propose that pepsins have evolved conserving not only the cleavage site, and the residues in the β-sheets linking the two lobes but also a region located on the N- lobe surface at the flap ends. All these regions should be important for the dynamical properties of this family of enzymes. The region on the surface of the protein, which is putatively relevant for substrate binding, is typical of only the pepsin family. Instead, the other two regions, which are directly involved in the enzymatic activity, have their structural counterparts in the fold of the functionally-related viral isoenzymes. Thus, the folds of retropepsins and pepsins, although very different, feature two fully-conserved similar region related to specific, evolutionarily-selected mechanical functions.

By synergically exploiting structural and dynamical information, it has been possible to characterize three structural regions with different functional roles in pepsins and two related ones in retropepsins. This methodology could be applied in the future as a useful tool for broadening the actual capabilities of identifying function for relevant targets in structural genomics research.

Appendix A

MSMD applied to water exchange at alkali ions

The novel fast-growth or multiple steering molecular dynamics (MSMD) technique has been recently developed by Jarzynski to calculate free energy profiles along general transformation pathways [12, 13]. Here, we apply this approach to calculate free energy barriers involved in the water exchange reaction of Na^+ and K^+ in aqueous solution. We investigate the influence of the key parameters of the MSMD simulations - the steering velocity, the sampling of the initial configurations and the force constant - on the free energy. Furthermore, we use this approach to describe energetical and structural features of the water exchange reaction of Na^+ and K^+ in aqueous solution.

A.1 Introduction

We investigate the water exchange reaction of alkali ions:

$$M(H_2O)_x^+ + H_2O^* \rightleftharpoons M(H_2O^*)(H_2O)_{x-1}^+ + H_2O$$

where M = Na^+ and K^+, and x is a typical coordination number [173]. Our computational approach is the novel multiple steering (or fast growth) molecular dynamics (MSMD) approach proposed by Jarzynski [12, 13, 103, 104], and introduced in sec. 2.5.2.

In the work presented here the MSMD technique is applied in the context of a chemical reaction. We find that the MSMD methodology enables efficient and accurate calculations of activation barriers and constitutes a promising new tool for the investigation of a wide range of chemical processes.

A.2 Methods

This unavoidable truncation on the average ensembles required in the theory (see sec. 2.5.2) introduce errors that can be minimized through an optimal choice of three key parameters, namely (i) the steering velocity v, (ii) the initial conditions for each simulation and (iii) the force constant k. The first influences directly the degree of irreversibility of the transformation; the second enters the ensemble average and affects the weight of each simulation; in fact, the initial bias on the steered system enters explicitly in the exponential of the right-hand side of eq. 2.82; the third affects the dynamics of the systems, and enters directly in the formulas as well. In the $v \to 0$ limit the transformation becomes reversible, the distribution of the argument of the exponential becomes infinitly sharp, and we recover the well known result that the free energy profile $G_0(z)$ coincides with the reversible work done to bring the system from the initial equilibrium state to the final one. We have checked that indeed this is the case for the systems described in the next sections. Moreover, the choice of initial configurations for the multiple MD runs has a direct influence on the efficiency with which the available phase space can be sampled within a limited number of simulations.

In particular, it is crucial that the averaging is performed with individual runs that are largely independent and thus uncorrelated. Thus, it may have a crucial influence on the convergence properties. Finally the choice of the force constant k could influence the convergence of the free energy profile. The convergence behavior with respect to these parameters is tested in this work (see Results section).

All the simulations were performed using the AMBER 5 program with the Åqvist [174], Smith [175] and TIP3P [150] force fields for Na^+, and K^+ cations, the Cl^- anion, and water, respectively. In each simulation, one alkali ion (either K^+, or Na^+) was immersed in a 20.4 x 20.4 x 36.5 $Å^3$ tetragonal box containing 421 water molecules and a Cl^- counterion. The latter, which was initially located 15.0 Å far from the alkali ions, remained always farther than 9.0 Å from the cations during the MD simulations. Periodic boundary conditions were applied. Long-range electrostatic interactions were treated via a particle-mesh Ewald procedure [76, 77] using a grid of 20x20x36 points with a cubic spline interpolation. A cutoff of 10.0 Å for the short-range electrostatic, and van der Waals interactions was used. Constant temperature (300 K) and constant pressure (1.0 atm) conditions were achieved by coupling the systems to a Berendsen thermostat and barostat [176]. The integration time step of our simulation was set to 1.5 fs.

All the systems underwent first 0.3 ns of unconstrained MD. Subsequently, the

Ion	1^{st} max.[a]	1^{st} min[b]	n[c]	1^{st} max.(lit) [d]	1^{st} min.(lit) [e]	n (lit) [f]
K^+	2.70	3.65	6.9 (5-10)	2.7-2.9	3.6-3.8	6.3-7.6-8.0
Na^+	2.40	3.25	5.8 (4-7)	2.4	3.2-3.5	4.9-6.0-6.6

Table A.1: Selected equilibrium properties of K^+ and Na^+ in aqueous solution. [a]) first maximum and [b]) first minimum of $g_{MO_w}(r)$ (Å) between the ion and water's oxygens and [c]) correspondent coordination numbers. The range of coordination numbers is indicated in brackets; ([d],[e],[f]) Properties obtained in previous simulations [177, 178].

position of one water molecule was restrained using a harmonic potential at a distance of 7.0 Å from the metal ion for 0.24 ns. The structural properties (such as the $g_{HO}(r)$ and $g_{OO}(r)$ radial distribution functions) resemble those of bulk water. 10 snapshots of this MD simulations, taken every 12 ps of dynamics, constituted our initial structures for the MSMD simulations.

In the MSMD simulations, the irreversible work *W* was calculated as the work required to bring the system from the initial equilibrium state to its final state. In practice, the approaching water (WAT hereafter) was steered from the bulk towards the alkali ion (M^+) applying the time-dependent potential as described in eq. 2.81. Consistent with the initial conditions, the starting position of the minimum of the harmonic potential $z^0(0)$ was set to 7.0 Å . The calculations were performed until the minimum of the harmonic potential reached a distance of 2.0 Å from the cation. MSMD runs with varying pulling velocities and different choices of the initial MD configurations were performed. Namely: (i) 4 set of 10 runs with different steering velocities on K^+ (v=0.0333, 0.0667, 0.3333, 0.6666 Å ps^{-1}), each of them with initial configurations sampled every 12 ps and with a force constant k = 300 pN $Å^{-1}$;

(ii) 3 sets of 10 runs with different sampling on the starting positions on K^+ (snapshot taken every Δt = 12, 1.5, 0.15 ps) with a pulling speed $v = 0.0666$ Å ps^{-1} and a force constant k = 300 pN $Å^{-1}$;

(iii) 6 sets of 10 runs with different force constants on K^+ (k= 30, 60, 100, 500, 1000, 1500 pN $Å^{-1}$) with v = 0.0666 Å ps^{-1} and initial configurations sampled every 12 ps;

(iv) 1 set of 10 runs on Na^+ with v=0.0333 Å ps^{-1}, k= 300 pN $Å^{-1}$ and with initial configuration sampled every Δt= 12 ps.

Coordination numbers of the alkali ions were calculated as integrals of the oxygen (water) - ion radial pair distribution functions (rdf) $g_{MO_w}(r)$ (Figure 1) from r = 0 to the first minimum of rdf (r = 2.7 and 2.4 Å for potassium and sodium, respectively

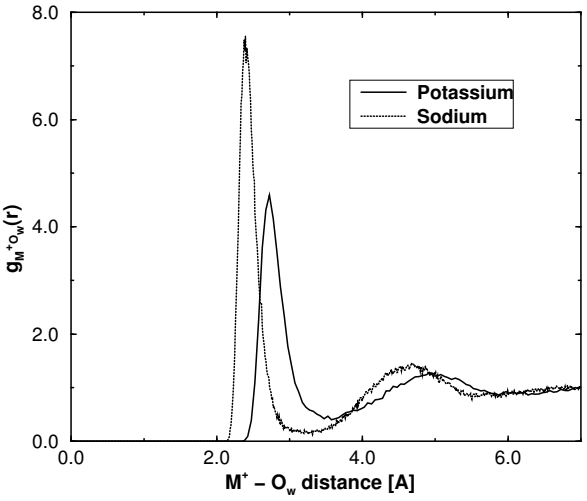


Figure A.1: Calculated ion/water radial pair distribution functions ($g_{MO_w}(r)$) for Na^+ and K^+ in aqueous solution.

(Figure A.1, Table A.1)). Identical integration boundaries were chosen in order to obtain the corresponding values in the MSMD runs.

A.3 Results and discussion

In this section, we first assess the accuracy of our computational scheme. Subsequently, we analyze the structural and energetic properties of sodium and potassium ions in aqueous solution.

Accuracy of the approach.

The accuracy of our computational approach was tested through a series of calculations for the water exchange reaction at K^+, for which a classical force field description is known to perform especially well [179].

In particular, we investigated the dependence of the calculated free energy from (i) the *steering velocities v* , from (ii) the choice of

the *initial configurations* and from (iii) the *force constant* k.

(i) Dependence on v.

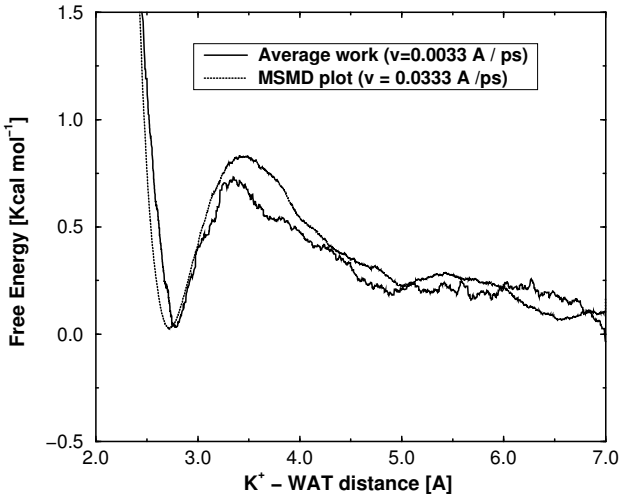


Figure A.2: Average work of the transformation at very slow pulling speed (v = 0.00333 Å ps^{-1}). This profile is used as a benchmark for the free-energy reconstruction profiles obtained at higher pulling speeds.

In this and in the subsequent simulations the initial structures were taken by selecting snapshots from our restrained MD simulation at a time interval of Δt=12 ps after the system was equilibrated.

The first reference simulation consists in using an extremely low steering velocity (v=0.00333 Å ps^{-1}) and computing the average work done in moving a water molecule from the bulk to the metal ion. Under this condition, the transformation can be considered quasi-static. Thus, it provides an approximate, reference free-energy profile of the process (Fig. A.2).

If v is one order of magnitude larger (v=0.0333 Å ps^{-1}), the reconstructed MSMD free energy profile turns out to converge within very few simulations (Fig. A.3a). Indeed, the profile can be considered fully converged after an averaging of ca. 8 trajectories (Fig. A.3a) and matches well our reference (Fig. A.2).

In this regime the free energy profile is qualitatively well described by a single integration of the force along the reaction coordinate (Fig. A.4a). This indicates that the performed work is essentially conservative and already a single simulation appears to contain sufficient information about the equilibrium properties.

By doubling the steering velocity, convergence to the same value of the free energy barrier is achieved (Table A.2, figure A.5). In both this and the previous case,

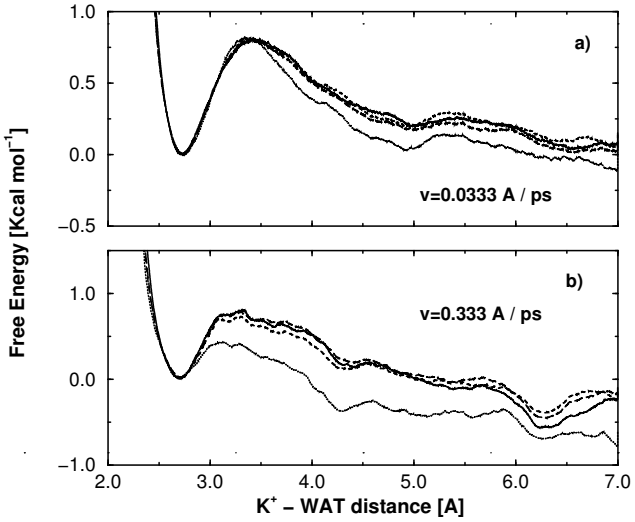


Figure A.3: Dependence of the reconstructed free energy of K^+ water exchange reaction on the steering velocity v. The free energy is obtained averaging data from four (dotted line), six (dashed line), eight (long-dashed) and ten (solid line) MD simulations. (a): v=0.0333 Å ps^{-1}; (b): v=0.3333 Å ps^{-1}.

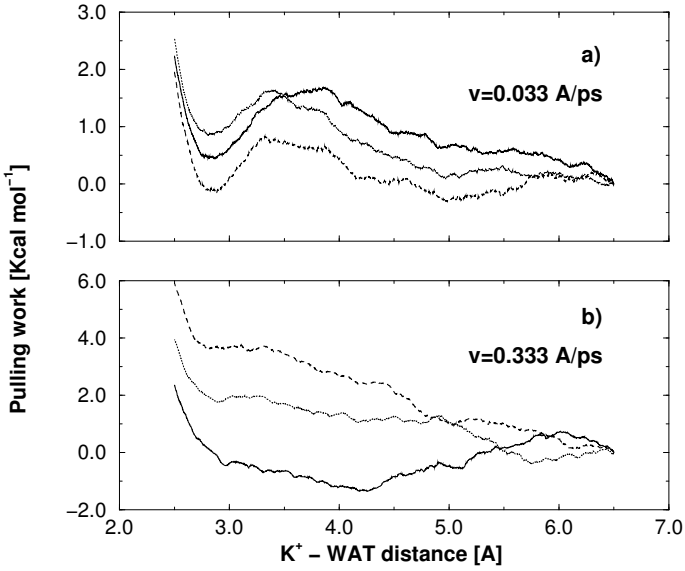


Figure A.4: Work associated with single steering dynamics runs at different conditions with non zero bias (see Tab. 2). (a): v=0.0333 Å ps^{-1}; (b) v=0.3333 Å ps^{-1}.

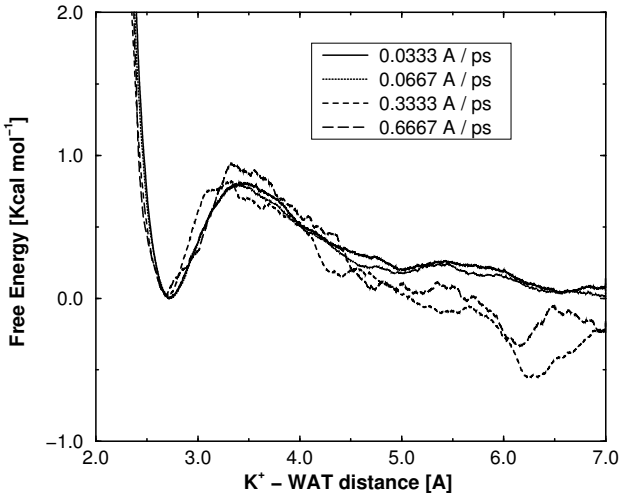


Figure A.5: Free energy profile for K^+ by using different pulling speeds. The free energy reconstruction profiles obtained at different v for the potassium ion are displayed: v=0.0333 Å ps^{-1} (solid line), v=0.0667 Å ps^{-1} (dotted line), v=0.3333 Å ps^{-1} (dashed line) and v=0.6667 Å ps^{-1} (long-dashed line).

hysteresis effects are essentially absent.

For relatively large velocities (v=0.3333 Å ps^{-1}), the free energy profile does not converge in the bulk region (that is between 7.0 Å and $\approx$ 5.5 Å) whereas it is similar to the previous ones in the the well region (between $\approx$ 3.5 and 2.0 Å) (Fig. A.3b). In this case, significant hysteresis effects are present, suggesting that the dominant part of the work associated with the steering force is non-conservative (Fig. A.4b).

Upon further increasing the steering velocity (v=0.6667 Å ps^{-1}), the free energy plot in the repulsive region (from $\approx$ 5.0 Å to $\approx$ 3.5 Å) and in the well region, although rather noisy, is still qualitatively similar to those obtained by pulling at a lower speed (Fig. A.5). This calculation therefore provides an upper limit for $v \approx 0.05$ Å ps^{-1}.

The qualitative shape of the free energy profile can be understood as follows: when the tagged water molecule is at large distance from the ion, it is surrounded by an average number of solvating water molecules as given in table A.1. The same situation is also expected when the tagged water molecule is at a particular distance from the ion substituting, on average, one water molecule from the hydration shell.

We thus expect that the free energy profile shows two minima whose values are the same, one at infinite distance from the ion and the other at a distance typical of

Ion	v [Å ps^{-1}] [a]	Δ G$^{\neq}$ [Kcal mol^{-1}] [b]	R$_0$ [Å] [c]	R$_{TS}$ [Å] [d]
K$^+$	0.0333	0.81	2.72	3.4
K$^+$	0.0667	0.80	2.73	3.4
K$^+$	0.3333	0.83	2.70	3.2
K$^+$	0.6667	0.92	2.70	3.3
Na$^+$	0.0333	1.30	2.43	3.1

Table A.2: *Multiple Steering Molecular Dynamics of K$^+$ and Na$^+$* [a]) Steering velocity, [b]) calculated activation free energy, [c]) M$^+$-WAT distance corresponding to the minimum of the free energy plot, [d]) distance of M$^+$-WAT at the TS.

the first coordination shell. As the tagged molecule approaches the ion a divergent profile due to Van der Waals repulsion is expected.

(ii) Influence of the *initial configurations*.

Varying initial configurations were obtained by selecting snapshots from our restrained MD simulations at different time intervals (Δt). Specifically, three sets of ten snapshots were considered. In the first Δt =12 ps , in the second Δt = 1.5 ps and in the third Δt = 0.15 ps (see Methods) . The same steering velocity (v=0.0666 Å ps^{-1}) was used for all calculations.

The reconstructed free energy profiles of the first two sets are rather similar and well converged (Fig. A.6). In contrast, the free energy profile in the third set does not converge within the ten simulations carried out here. This suggests that the initial configurations are not independent enough, resulting in an insufficient sampling of phase space.

(iii) Choice of the *force constant k*.

Calculations with values of k around the reported value of [105] (from 30 to 1500 pN Å^{-1}) were performed. It is found that from $k > 60$ pN Å^{-1} up to $k < 1000$ pN Å^{-1} the convergence is fast and no significant changes are found in the profiles.

In conclusion, our analysis allows to establish the following criteria for the choice of the three computational key parameters: (i) a steering velocity $v \leq 0.05$ Å ps^{-1}; (ii) initial configurations sampled at a time interval of $\Delta t \geq 1.5$ ps; (iii) the force constant should be of the order of $\approx$ 100-600 pN Å^{-1}. Within these conditions, the convergence on the free energy appears to be rather fast.

For all simulations of reaction (1) we used v=0.0333 Å ps^{-1}, Δ t=12 ps and k=300 pN Å^{-1}. These values compare well with those used in the MSMD simulations of aquaglyceroporin of ref. [105]. The free energy profiles obtained by using these

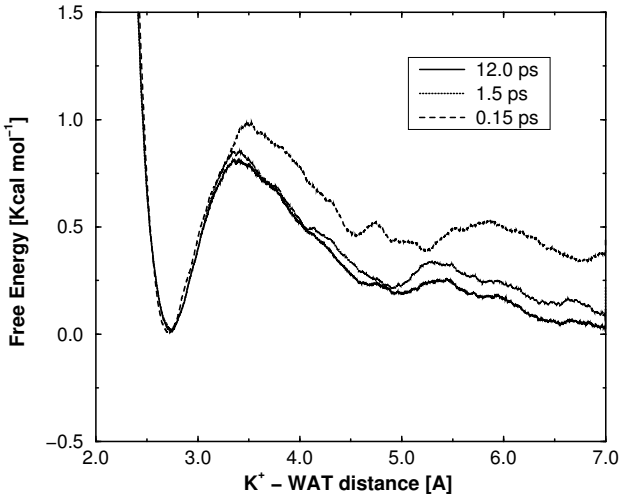


Figure A.6: Influence of the initial bias. The three profiles have been calculated averaging ten simulations. Each initial condition has been chosen sampling a biased dynamics runevery 12 ps (solid line), 1.5 ps (dotted line) or 0.15 ps (dashed line).

parameters for both of the two cations are shown in Fig. A.7.

Potassium. In the unconstrained simulation of a potassium ion in aqueous solution, the metal ion exhibits a labile coordination sphere (average coordination number 6.9). Seven water molecules coordinate K^+ for about half of the simulated time (0.15 ns). The n = 8 coordination is also significant (about 28% of the time). In the rest of the dynamics, K^+ is coordinated to a largely varying number of ligands (Table 1). The observed coordination numbers and their average is in good agreement with other data present in the literature [177] - [178] and with the one of ab initio MD simulations [179].

The exchange reaction (1) occurs spontaneously during the dynamics. The overall process takes place within as short a time as 0.5 ps in qualitative agreement with previous findings [179, 177, 180, 181, 182].

The exchange reaction turns out to occur mostly via a substitution mechanism (Figure A.8). However, also addition and elimination processes are observed.

At the transition state of the substitution mechanism the oxygen belonging to the approaching water interacts with both the metal ion and with a water molecule belonging to the second shell (Figure A.8 and Table A.2). Its presence induces a distortion in the potassium coordination polyhedron, accompanied by a weakening of

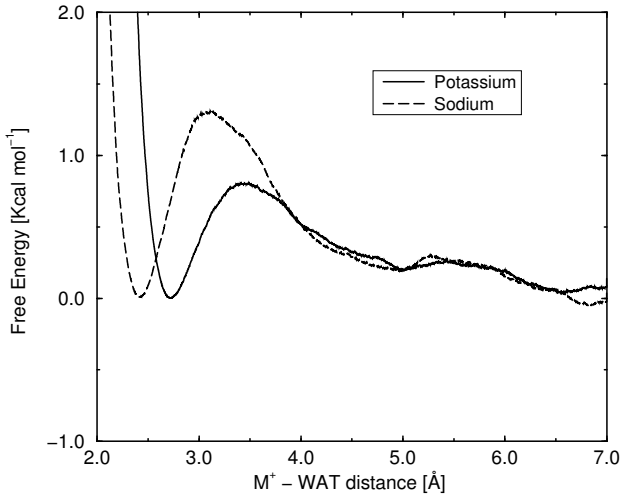


Figure A.7: Free energy plots for the two ions. The free energy plots for Na^+ and K^+ are shown. Ten simulations were averaged setting the pulling speed to v=0.0333 Å ps^{-1}, and obtaining the inital conditions sampling every 12 ps a former biased MD simulation.

the interaction between the metal ion and one of the water binding to it. The latter eventually leaves the K^+ coordination sphere assisted by the formation of a hydrogen bond with another metal-bound water. In the absence of the formation of this last hydrogen bond, addition of the approaching water to the ion is observed. Our MSMD simulations provide the same mechanistic picture for the addition mechanisms.

The calculated free energy of the process has a plateau in the bulk region followed by a maximum at 3.4 corresponding to the activated complex, and a minimum at about 2.7 Å (Table A.2). The latter value represents the optimal ion-water distance and is in agreement with the maximum of the radial pair distribution function (Table A.1, and Figures A.1 and A.7).

Sodium. In our equilibrium simulations Na^+ is coordinated by five, and, more (66% of the time) by six water molecules (average coordination number n=5.7). The penta- and hexa-coordinated complexes are stable for more than 10 ps and exhibit clear trigonal bipyramidal and octahedral equilibrium geometries.

The water exchange mechanism occurs via a associative/dissociative pathway, in which the sodium coordination number changes from 5 to 6 and viceversa. In the associative process, the approaching water molecule forms a hydrogen-bond to one of the water molecules bound to the metal ion; subsequently, it enters the coordina-

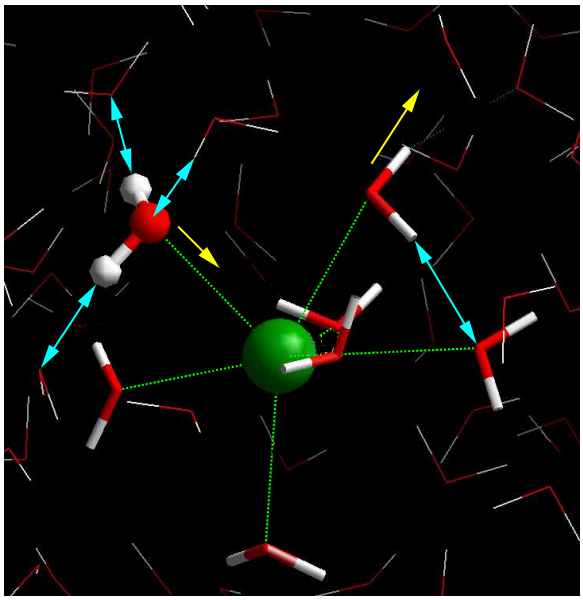


Figure A.8: Transition state geometry of reaction (1) for the potassium ion. The incoming water (WAT) is represented by a ball-and-sticks model and first-shell water molecules are drawn as cylinders and linked to the central ion by green dashed lines. Selected H-bonds are shown as blue arrows. Structural changes of the ligand coordination sphere are displayed as yellow arrows.

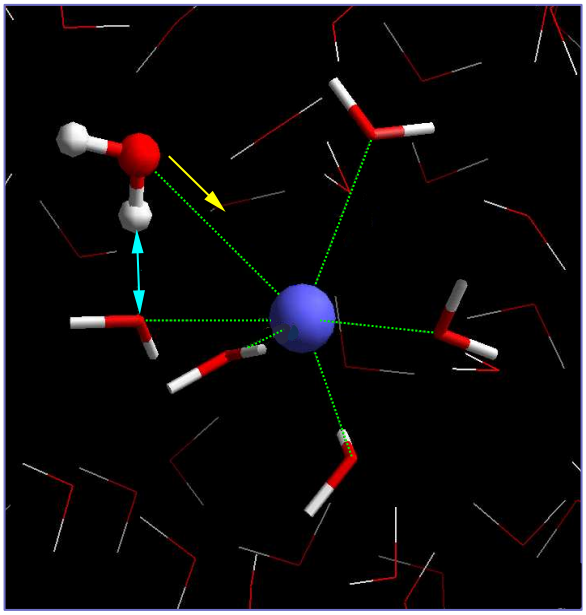


Figure A.9: Transition state geometry of reaction for the sodium ion. The incoming water (WAT) is represented by a ball-and-sticks model and first-shell water molecules are drawn as cylinders and linked to the central ion by green dashed lines. Selected H-bonds are shown as blue arrows. Structural changes of the ligand coordination sphere are displayed as yellow arrows.

tion sphere. The transition state (Figure A.9 and Table A.2) is a distorted octahedral structure, in which the incoming water still forms an H-bond to a metal-bound water molecule (Figure A.9 and Table A.2). Finally, the geometry of the coordination polyhedron relaxes to an octahedral structure.

In the dissociative process, the 6-coordinated complex undergoes relatively large structural fluctuations, which at times cause the formation of an H-bond between two metal-bound water molecules. This induces the release of one of the water ligands and reformation of the trigonal-bipyramidal geometry. Thus, basically, the mechanism of the associative part of the reaction is the direct inverse of its dissociative counter part.

These mechanisms have been observed in an recent ab-initio molecular dynamics simulations [183]. Furthermore, the characteristic time scale of the process (about 0.3 ps) is also similar to that of the ab-initio calculation [183].

In our MSMD simulations exchange reaction occurs via the same mechanisms observed for the unconstrained ones. The free energy profile of the process is plotted in Figure 6. As expected, the free energy barrier is larger and the minimum less

pronounced than the corresponding values for potassium. As in the case of K^+, the energy minimum corresponds essentially to the maximum of the $g_{MO_w}(r)$ $R_0 = 2.43$ Å (see Figures A.1-A.7, Tables A.1-A.2).

A.4 Conclusions

Our calculations provide an energetic and structural description of water exchange at alkali ions in aqueous solution using the MSMD approach [12, 13, 103, 104].

The convergence of the free energy profiles depends on the steering velocity and on the initial sampling conditions. A proper choice of these key parameters leads to fast convergence by averaging only a limited number of simulations (≤ 10).

As expected, the solvation shell of sodium, and, even more so, the one of potassium is highly flexible. It includes dynamical structures involving a varying numbers of coordinated water molecules and different geometries within each coordination number. The water exchange reaction for potassium occurs with a direct substitution mechanism.

In contrast, for sodium, it takes places via an associative/dissociative mechanism, in which the coordination of the metal goes from trigonal-bipyramid to octahedral and vice-versa. Both processes are assisted by the formation and breaking of H-bonds between the incoming water molecule and the metal ligands. These results are in good agreement with recent ab-initio calculations [183].

In conclusion, the MSMD technique appears to be a very powerful, fast and reliable technique to study chemical processes. This approach holds great promise for free energy evaluations in complex systems with relatively long relaxation times, where molecular dynamics simulations cannot lead to the direct observation of the physically interesting events.

Notes

The works illustrated along this thesis have been collected in two published scientific papers:

- M. Cascella, L. Guidoni, U. Rothlisberger, A. Maritan, and P. Carloni. Multiple steering molecular dynamics applied to water exchange at alkali ions. *J. Phys. Chem. B*, 106:13027–13032, 2002.
- M. Cascella, S. Raugei and P. Carloni. Formamide Hydrolysis Investigated by Multiple Steering Ab-Initio Molecular Dynamics. *J. Phys. Chem. B*, 108:369-375, 2004.

And in two other manuscripts, at present submitted for publication:

- M. Cascella, C. Micheletti, U. Rothlisberger and P. Carloni. Evolutionarily conserved functional mechanics across pepsins and retropepsins.
- M. Neri, M. Cascella and C. Micheletti. Influence of conformational fluctuations on enzymatic activity: modelling the functional motion of beta secretase

During these years, I also collaborated to other scientific works on biological systems investigated in Paolo Carloni's group. These studies, which have not been presented in this thesis, concern the mechanism of permeation of water through the aquaporin channel, and the decarboxylation reaction of orotidine 5'-monophosphate by orotidine 5'-monophosphate decarboxylase, and have been collected in the following papers:

- P. Vidossich, M. Cascella and P. Carloni. Dynamics and Energetics of Water Permeation through the Aquaporin Channel. *Proteins*, 55:924-931, 2004.
- S. Raugei, M. Cascella and P. Carloni. A proficient enzyme: Insights on the mechanism of Orotidine Monophosphate Decarboxylase from computer simulations, *J. Am. Chem. Soc.*, in press.

Bibliography

[1] D. R. Davies. The structure and function of the aspartic proteinases. *Ann. Rev. Biophys. Biophys. Chem.*, 19:189–215, 1990.

[2] 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.

[3] M. N. G. James and A. R. Sielecki. Structure and refinement of penicillopepsin at 1.8 Å resolution. *J. Mol. Biol.*, 163:299–361, 1983.

[4] N. D. Rawlings and A. J. Barrett. Families of aspartic peptidases, and those of unknown catalytic mechanism. *Method. Enzymol.*, 248:105–120, 1995.

[5] 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.

[6] 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.

[7] S. Piana, D. Bucher, P. Carloni, and U. Rothlisberger. Reaction mechanism of HIV-1 protease by hybrid car-parrinello/classical md simulations. *J. Phys. Chem. B*, 108:11139–11149, 2004.

[8] H. Yang, G. Luo, P. Karnchanaphanurach, T. M. Louie, I. Rech, S. Cova, L. Xun, and X. S. Xie. Protein conformational dynamics probed by single-molecule electron transfer. *Science*, 302:262–266, 2003.

[9] T. H. Rod, J. L. Radkiewicz, and C. L. Brooks III. Correlated motion and the effect of distal mutations in dihydrofolate reductase. *Proc. Natl. Acad. Sci. USA*, 100:6980–6985, 2003.

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

[11] R. Car and M. Parrinello. Unified approach for molecular-dynamics and density functional theory. *Phys. Rev. Lett.*, 55:2471–2474, 1985.

[12] C. Jarzynski. Nonequilibrium equality for free energy differences. *Phys. Rev. Lett.*, 78:2690–2693, 1997.

[13] C. Jarzynski. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Phys. Rev. E*, 56:5018–5035, 1997.

[14] M. Cascella, L. Guidoni, U. Rothlisberger, A. Maritan, and P. Carloni. Multiple steering molecular dynamics applied to water exchange at alkali ions. *J. Phys. Chem. B*, 106:13027–13032, 2002.

[15] A. Laio, J. VandeVondele, and U. Rothlisberger. D-RESP: Dynamically generated electrostatic potential derived charges from quantum mechanics/molecular mechanics simulations. *J. Phys. Chem. B*, 106:7300–7307, 2002.

[16] D. J. Selkoe. Alzheimer's disease: Genes, proteins and therapy. *Physiol. Rev.*, 81:741–766, 2001.

[17] P. Carloni, U. Rothlisberger, and M. Parrinello. The role and perspective of a initio molecular dynamics in the study of biological systems. *Acc. Chem. Res.*, 35:455–464, 2002.

[18] M. Colombo, L. Guidoni, A. Laio, A. Magistrato, P. Maurer, S. Piana, U. Röhrig, K. Spiegel, M. Sulpizi, J. VandeVondele, M. Zumstein, and U. Rothlisberger. Hybrid QM/MM Car-Parrinello simulations of catalytic and enzymatic reactions. *CHIMIA*, 56:13–19, 2002.

[19] M. Sulpizi, A. Laio, J. VandeVondele, A. Cattaneo, U. Rothlisberger, and P. Carloni. Reaction mechanism of caspases: Insights from QM/MM Car-Parrinello simulations. *Proteins*, 52:212–224, 2003.

[20] P. Carloni, M. Sprik, and W. Andreoni. Key steps of the cis-platin-DNA interaction: Density functional theory-based molecular dynamics simulations. *J. Phys. Chem. B*, 104:823–835, 2000.

[21] K. Spiegel, U. Rothlisberger, and P. Carloni. Cisplatin binding to DNA oligomers from hybrid Car-Parrinello/molecular dynamics simulations. *J. Phys. Chem. B*, 108:2699–2707, 2004.

[22] 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.

[23] L. Stryer. *Biochemistry*. W. H. Freeman & Co, 1995.

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

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

[26] S. P. L. Sörensen. Enzymstudien II. Mitteilung. über die Messung und die Bedeutung der Wasserstoffionen-Konzentration bei enzymatischen Prozessen. *Biochem. Z.*, 21:201–304, 1909.

[27] B. M. Dunn. Structure and mechanism of the pepsin-like family of aspartic peptidases. *Chem. Rev.*, 102:4431–4458, 2002.

[28] R. C. Ogden and C. W. Flexner, editors. *Protease Inhibitors in AIDS therapy*. Marcel Dekker Inc. New York, 2001.

[29] M. Tourella, K. Gordon, and T. Hohn. Cauliflower mosaic-virus produces an aspartic proteinase to cleave its polyproteins. *EMBO J.*, 8:2819–2825, 1989.

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

[31] J. Tang, M. N. G. James, I. N. Hsu, J. A. Jenkins, and T. L. Blundell. Structural evidence for gene duplication in evolution of acid proteases. *Nature*, 271:618, 1978.

[32] X. Lin, Y. Lin, G. Koelsch, A. Gustchina, A. Wlodawer, and J. Tang. Enzymatic-activities of 2-chain pepsinogen, 2-chain pepsin, and the amino-terminal lobe of pepsinogen. *J. Biol. Chem.*, 267:17257, 1992.

[33] E. Ido, H. Han, F. J. Kezdy, and J. Tang. Kinetic studies of human immunodeficiency virus type 1 protease and its active-site hydrogen bond mutant A28S. *J. Biol. Chem.*, 266:24359, 1991.

[34] 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.

[35] M. N. G. James, A. R. Sielecki, and T. Hofmann. *Aspartic proteinases and their inhibitors*, page 165. de Gruyter, Berlin, 1985.

[36] 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.

[37] 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.

[38] 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.

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

[40] R. Bott, E. Subramanian, and D. Davies. 3-dimensional structure of the complex of the rhizopus-chinensis carboxyl proteinase and pepstatin at 2.5 Å resolution. *Biochemistry*, 21:6965, 1982.

[41] M. N. G. James, A. R. Sielecki, F. Salituro, D. H. Rich, and T. Hofmann. Conformational flexibility in the active sites of aspartyl proteinases revealed by a pepstatin fragment binding to penicillopepsin. *Proc. Natl. Acad. Sci. USA*, 79:6137, 1982.

[42] 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.

[43] 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.

[44] 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.

[45] T. D. Meek, E. J. Rodriguez, and T. S. Angeles. Use of steady-state kinetic methods to elucidate the kinetic and chemical mechanisms of retroviral proteases. *Methods Enzymol.*, 241:127–156, 1994.

[46] N. S. Andreeva and L. D. Rumsh. Analysis of crystal structures of aspartic proteinases: On the role of amino acid residues adjacent to the catalytic site of pepsin-like enzymes. *Protein Sci.*, 10:2439, 2001.

[47] 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.

[48] 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.

[49] 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.

[50] D. W. Dickson. The pathogenesis of senile plaques. *J. Neuropathol. Exp. Neurol.*, 56:321–339, 1997.

[51] C. Haass, A. Y. Hung, and D. J. Selkoe. Processing of β-amyloid precursor protein in microglia and astrocytes favors a localization in internal vesicles over constitutive secretion. *J. Neurosci.*, 11:3783–3793, 1991.

[52] C. Haass, M. Schlossmacher, A. Y. Hung, C. Vigo-Pelfrey, A. Mellon, B. Ostaszewski, I. Lieberburg, E. H. Koo, D. Schenk, D. Teplow, and D. J. Selkoe. Amyloid precursor protein is produced by cultured cells during normal metabolism. *Nature*, 359:322–325, 1992.

[53] P. Seubert, C. Vigo-Pelfrey, F. Esch, M. Lee, H. Dovey, D. Davis, S. Sinha, M. G. Schlossmacher, J. Whaley, C. Swindlehurst, R. McCormak, R. Wolfert, D. J. Selkoe, I. Lieberburg, and D. Schenk. Isolation and quantitation of soluble alzheimer's β-peptide from biological fluids. *Nature*, 359:325–327, 1992.

[54] M. Shoji, T. E. Golde, J. Ghiso, T. T. Cheung, S. Estus, L. M. Shaffer, X. Cai, D. M. MacKay, R. Tintner, B. Frangione, and S. G. Younkin. Production of the alzheimer amyloid β protein by normal proteolytic processing. *Science*, 258:126–129, 1992.

[55] S. Capsoni, S. Giannotta, and A. Cattaneo. beta-amyloid plaques in a model for sporadic Alzheimer's Disease based on transgenic anti-nerve growth factor antibodies. *Mol. Cell. Neurosci.*, 21:15–28, 2002.

[56] S. Capsoni, S. Giannotta, and A. Cattaneo. Nerve growth factor and galantamine ameliorate early signs of neurodegeneration in anti-nerve growth factor mice. *Proc. Natl. Acad. Sci. USA*, 99:12432–12437, 2002.

[57] J. T. Jarrett, E. P. Berger, and P. T. Lansbury Jr. The carboxy terminus of the beta amyloid protein is critical for the seeding of amyloid formation: implications for the pathogenesis of Alzheimer's Disease. *Biochemistry*, 32:4693–4697, 1993.

[58] S. F. Lichtenthaler, R. Wang, H. Grimm, S. Uljon, C. L. Masters, and K. Beyreuther. Mechanism of the cleavage specificity of alzheimer's disease gamma-secretase identified by phenylalanine-scanning mutagenesis of the transmembrane domain of the amyloid precursor protein. *Proc. Natl. Acad. Sci. USA*, 96:3053–3058, 1999.

[59] S. Sinha and I. Lieberburg. Cellular mechanism of β-amyloid production and secretion. *Proc. Natl. Acad. Sci. USA*, 96:11049–11052, 1999.

[60] 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.

[61] L. Hong, R. T. Turner, G. Koelsch, D. Shin, A. K. Ghosh, and J. Tang. Crystal structure of memapsin 2 (beta-secretase) in complex with inhibitor om00-3. *Biochemistry*, 41:10963, 2002.

[62] C. Cohen-Tannoudji, B. Diu, and F. Laloe. *Quantum mechanics*. J. Wiley & Sons, 1995.

[63] B. H. Bransden and C. J. Joachain. *Physics of atoms and molecules*. Longman, London, 1983.

[64] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Phys. Rev. B.*, 136:864–871, 1964.

[65] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Phys. Rev. A*, 140:1133–1138, 1965.

[66] A. D. Becke. Density-functional exchange-energy approximation with correct asymptotic-behavior. *Phys. Rev. A*, 38:3098–3100, 1988.

[67] C. Lee, W. Yang, and R. G. Parr. Development of the colle-salvetti correlation-energy formula into a functional of the electron-density. *Phys. Rev. B*, 37:785–589, 1988.

[68] G. J. Martyna and M. E. Tuckerman A reciprocal space based method for treating long-range interactions in *ab initio* and froce-field-based calculations in clusters. *J. Chem. Phys.*, 110:2810-2821, 1999.

[69] N. Troullier and J. L. Martins. Efficient pseudopotentials for plane-wave calculations. *Phys. Rev. B*, 43:1993–2006, 1991.

[70] L. Kleinman and D. M. Bylander. Efficacious form for model pseudopotentials. *Phys. Rev. Lett.*, 48:1425–1428, 1982.

[71] D. Marx and J. Hütter. *Modern Methods and Algorithms of Quantum Chemistry*, volume 1, chapter Ab initio molecular dynamics: Theory and implementation, pages 301–449. John von Neumann Institute for Computing, Jülich, NIC Series, 2000.

[72] S. J. Nosé. A unified formulation of the constant temperature molecular-dynamics method. *J. Chem. Phys.*, 81:511–519, 1984.

[73] W. G. Hoover. Canonical dynamics. equilibrium phase-space distributions. *Phys. Rev. A*, 31:1695–1697, 1985.

[74] G. Martyna, M. Tuckerman, and M. L. Klein. Nose-hoover chains. the canonical ensemble via continuous dynamics. *J. Chem. Phys.*, 97:2635–2643, 1992.

[75] J. Hutter, A. Alavi, T. Deutsch, M. Bernasconi, St. Goedecker, D. Marx, M. E. Tuckerman, and M. Parrinello. CPMD, version 3.5. MPI für Festkörperforschung and IBM Research Laboratory: Stuttgart and Zürich, 1995-2001.

[76] T. Darden, D. York, and L. G. Pedersen. The effect of long-range electrostatic interactions in simulations of macromolecular crystals - a comparison of the Ewald and truncated list method. *J. Chem. Phys.*, 98:10089–10092, 1993.

[77] U. Essman, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen. A smooth particle mesh Ewald method. *J. Chem. Phys.*, 103:8577–8593, 1995.

[78] W. D. Cornell, P. Cieplak, C. I. Bayly, I. R. Gould, J. W. Caldwell, and P. A. Kollman. A 2nd generation force-field for the simulation of proteins, nucleic-acids, and organic-molecules. *J. Am. Chem. Soc.*, 117:5179–5197, 1995.

[79] H. J. C. Berendsen, D. van der Spoel, and R. Vandrunen. Gromacs: a message-passing parallel molecular-dynamics implementation. *Comp. Phys. Comm.*, 91:43–56, 1995.

[80] 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.

[81] C. I. Bayly, P. Cieplak, W. D. Cornell, and P. A. Kollman. Electrostatic potential based method using charge restraints for determining atom-centered charges: the RESP method. *J. Phys. Chem.*, 97:10269, 1993.

[82] B. Hess, H. Bekker, H. J. C. Berendsen, and J. G. E. M. Fraije. Lincs: A linear constraint solver for molecular simulations. *J. Comp. Chem.*, 18:1463–1472, 1997.

[83] M. Parrinello and A. Rahman. Polymorphic transitions in single crystals. a new molecular dynamics method. *J. Appl. Phys.*, 52:7182–7190, 1981.

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

[85] B. Brooks and M. Karplus. Normal modes for specific motions of macromolecules: application to the hinge-bending mode of lysozyme. *Proc. Natl. Acad. Sci. USA*, 82:4995–3999, 1985.

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

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

[88] 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.

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

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

[91] A. Laio, J. VandeVondele, and U. Rothlisberger. A hamiltonian electrostatic coupling scheme for hybrid Car-Parrinello molecular dynamics simulations. *J. Chem. Phys.*, 116:6941–6947, 2002.

[92] M. Sulpizi and P. Carloni. Cation-π versus OH-π interactions in proteins: A density functional study. *J. Phys. Chem. B*, 104:10087–10091, 2000.

[93] S. F. Boys. . *Rev. Mod. Phys.*, 32:296, 1960.

[94] G. H. Wannier. The structure of electronic excitation levels in insulating crystals. *Phys. Rev.*, 52:191–197, 1937.

[95] N. Marzari and D. Vanderbilt. Maximally-localized wannier functions for composite energy bands. *Phys. Rev. B*, 56:12847–12865, 1997.

[96] P. L. Silvestrelli, N. Marzari, D. Vanderbilt, and M. Parrinello. Maximally-localized wannier function for disordered systems: application to amorphous silicon. *Solid State Commun.*, page 7, 1998.

[97] A. Laio and M. Parrinello. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA*, 99:12562–12566, 2002.

[98] D. Passerone and M. Parrinello. Action-derived molecular dynamics in the study of rare events. *Phys. Rev. Lett.*, 87:108302, 2001.

[99] R. Elber. *Recent developments in theoretical studies of proteins*. World Scientific, 1996.

[100] C. Dellago, P. G. Bolhuis, and D. Chandler. On the calculation of reaction rate constants in the transition path ensemble. *J. Chem. Phys.*, 110:6617–6624, 1999.

[101] M. Sprik and G. Ciccotti. Free energy from constrained molecular dynamics. *J. Chem. Phys.*, 109:7737–7744, 1998.

[102] J. Vandevondele and U. Rothlisberger. Accelerating rare reactive events by means of a finite electronic temperature. *J. Am. Chem. Soc.*, 124:8163–8171, 2002.

[103] G. E. Crooks. Path-ensemble averages in systems driven far from equilibrium. *Phys. Rev. E*, 61:2361–2366, 2000.

[104] G. Hummer and A. Szabo. Free energy reconstruction from nonequilibrium single-molecule pulling experiments. *Proc. Natl. Acad. Sci. USA*, 98:3658–3661, 2001.

[105] M. Ø. Jensen, S. Park, E. Tajkhorshid, and K. Schulten. Energetics of glycerol conduction through aquaglyceroporin GlpF. *Proc. Natl. Acad. Sci. USA*, 99:6731–6736, 2002.

[106] P. Vidossich, M. Cascella, and P. Carloni. Dynamics and energetics of water permeation through the aquaporin channel. *Proteins*, 55:924–931, 2004.

[107] G. Hummer. Fast-growth thermodynamic integration: Error and efficiency analysis. *J. Chem. Phys.*, 114:7330–7337, 2001.

[108] J. Liphardt, S. Dumont, S. B. Smith, I. Tinoco, and C. Bustamante. Equilibrium information from nonequilibrium measurements in an experimental test of Jarzynski's equality. *Science*, 296:1832–1835, 2002.

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

[110] M. M. Tirion. Large amplitude elastic motions in proteins form a single-parameter, atomic analysis. *Phys. Rev. Lett.*, 77:1905–1098, 1996.

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

[112] P. Doruker, A. 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.

[113] 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. *Bioph. J.*, 80:505–515, 2001.

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

[115] J. D. Watson and F. H. C. Crick. A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[116] C. Sander and R. Schneider. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, 9:56–68, 1991.

[117] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89:10915–10919, 1992.

[118] J. D. Thompson, D. G. Higgins, and T. J. Gibson. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Computer Applications in the Biosciences*, 1:19–29, 1994.

[119] C. Notredame, D. Higgins, and J. Heringa. T-COFFEE: a novel method for mutiple sequence alignments. *J. Mol. Biol.*, 302:205–217, 2000.

[120] D. Haussler, A. Krogh, I. S. Mian, and K. Sjölander. Protein modeling using Hidden Markov Models: Analysis of globins. In *Proceedings of the Hawaii International Conference on System Sciences*, volume 1, pages 792–802. IEEE Computer Society Press, Los Alamitos, CA, 1993.

[121] A. Krogh, M. Brown, I. Mian, K. Sjölander, and D. Haussler. Hidden Markov Models in computational biology: Applications to protein modeling. *J. Mol. Biol*, 235:1501–1531, 1994.

[122] K. Karplus, C. Barrett, and R. Hughey. Hidden Markov Models for detecting remote protein homologies. *Bioinformatics*, 14:846–856, 1998.

[123] P. Baldi, Y. Chauvin, T. Hunkapillar, and M. McClure. Hidden Markov Models of biologically primary sequence information. *Proc. Natl. Acad. Sci. USA*, 91:1059–1063, 1994.

[124] L. Hong and J. Tang. Flap position of free memapsin 2 (beta-secretase), a model for flap opening in aspartic protease catalysis. *Biochemistry*, 43:4689–4695, 2004.

[125] 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 Åresolution. *Science*, 246:1149–1152, 1989.

[126] M. Štrajbl, J. Florián, and A. Warshel. Ab initio evaluation of the potential surface for general base-catalyzed methanolysis of formamide: A reference solution reaction for studies of serine proteases. *J. Am. Chem. Soc.*, 122:5354–5366, 2000.

[127] J. Hine, R. S.-M. King, W. R. Midden, and A. Sinha. Hidrolysis of formamide at 80◦ celsius and pH 1-9. *J. Org. Chem.*, 46:3186–3189, 1981.

[128] B. A. Robinson and J. W. Tester. Kinetics of alkaline-hydrolysis of organic esters and amides in neutrally-buffered solution. *Int. J. Chem. Kinet.*, 22:431–448, 1990.

[129] H. Slebocka-Tilk, A. J. Bennet, H. J. Hogg, and R. S. Brown. Predominant O-18 exchange accompanying base hydrolysis of a tertiary toluamide: N-ethyl-n-(trifluoroethyl)toluamide. assessment of the factors that influence partitioning of anionic tetrahedral intermediates. *J. Am. Chem. Soc.*, 112:8507–8514, 1991.

[130] J. P. Guthrie. Hydration of carboxamides. evaluation of the free energy change for addition of water to acetamide and formamide derivatives. *J. Am. Chem. Soc.*, 96:3608–3615, 1974.

[131] T. Oie, G. H. Loew, S. K. Burt, J. S. Binkley, and R. D. MacElroy. Quantum chemical studies of a model for peptide-bond formation: formation of formamide and water from ammonia and formic-acid. *J. Am. Chem. Soc.*, 104:6169–6174, 1982.

[132] J. H. Jensen, K. K. Baldridge, and M. S. Gordon. Uncatalized peptide-bond formation in the gas-phase. *J. Phys. Chem.*, 96:8340–8351, 1992.

[133] B. Kallies and R. J. Mitzner. Models of water-assisted hydrolyses of methyl formate, formamide, and urea from combined DFT-SCRF calculations. *J. Mol. Model.*, 4:183–196, 1998.

[134] S. J. Weiner, U. C. Singh, and P. A. Kollman. Simulation of formamide hydrolysis by hydroxide ion in the gas-phase and in aqueous solution. *J. Am. Chem. Soc.*, 107:2219–2229, 1985.

[135] D. Bakowies and P. A. Kollman. Theoretical study of base-catalyzed amide hydrolysis: Gas- and aqueous-phase hydrolysis of formamide. *J. Am. Chem. Soc.*, 121:5712–5726, 1999.

[136] J. D. Madura and W. L. Jorgensen. Ab initio and montecarlo calculations for a nucleophilic-addition reaction in the gas-phase and in aqueous solution. *J. Am. Chem. Soc.*, 108:2517–2527, 1986.

[137] A. Warshel and S. Russell. Theoretical correlation of structure and energetics in the catalytic reaction of trypsin. *J. Am. Chem. Soc.*, 108:6569–6579, 1986.

[138] A. Warshel, F. Sussman, and J. K. Hwang. Evaluation of catalytic free-energies in genetically modified proteins. *J. Mol. Biol.*, 201:139–159, 1988.

[139] B. Chen, J. M. Park, I. Ivanov, G. Tabacchi, M. L. Klein, and M. Parrinello. First-principles study of aqueous hydroxide solutions. *J. Am. Chem. Soc.*, 124:8534–8535, 2002.

[140] S. Raugei and M. L. Klein. Dynamics of water molecules in the Br^- solvation shell: An ab initio molecular dynamics study. *J. Am. Chem. Soc.*, 123:9484–9485, 2001.

[141] J. Baker, J. Andzelm, M. Muir, and P. R. Taylor. OH + H_2 $\rightarrow$ H_2O + H. the importance of 'exact exchange' in density functional theory. *Chem. Phys. Lett.*, 237:53–60, 1995.

[142] S. Parthiban, G. de Oliveira, and J. M. L. Martin. Benchmark ab intio energy profiles for the gas-phase S_n2 reactions Y^-+CH_3X $\rightarrow$ CH_3Y + X^- (X,Y= F,Cl,Br). validation of hybrid DFT methods. *Phys. Rev. A*, 105:895–904, 2001.

[143] F. Alber, G. Folkers, and P. Carloni. Dimethyl phosphate: Stereoelectronic versus environmental effects. *J. Phys. Chem. B*, 103:6121–6126, 1999.

[144] M. Sprik. Computation of the pK of liquid water using coordination constraints. *Chem. Phys.*, 258:139–150, 2000.

[145] M. E. Tuckerman, K. Laasonen, M. Sprik, and M. Parrinello. Ab initio molecular dynamics simulation of the solvation and transport of H_3O^+ and OH^- ions in water. *J. Phys. Chem*, 99:5749–5752, 1995.

[146] M. E. Tuckerman, K. Laasonen, M. Sprik, and M. Parrinello. Ab initio molecular dynamics simulation of the solvation transport of hydronium and hydroxyl ions in water. *J. Chem. Phys*, 103:150–161, 1995.

[147] M. E. Tuckerman, D. Marx, and M. Parrinello. The nature and transport mechanism of hydrated hydroxide ions in aqueous solution. *Nature*, 417:925–929, 2002.

[148] Z. W. Zhu and M. E. Tuckerman. Ab initio molecular dynamics investigation of the concentration dependence of charged defect transport in basic solutions via calculation of the infrared spectrum. *J. Phys. Chem. B*, 106:8009–8018, 2002.

[149] A. Beveridge. A theoretical study of torsional flexibility in the active site of aspartic proteinases: Implications for catalysis. *Proteins*, 24:322–334, 1996.

[150] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.*, 79:926–935, 1983.

[151] O. A. Asojo, E. Afonina, S. V. Gulnik, B. Yu, J. W. Erickson, R. Randad, D. Mehadjed, and A. M. Silva. Structures of Ser205 mutant plasmepsin II from plasmodium falciparum at 1.8 angstrom in complex with the inhibitors rs367 and rs370. *Acta Crystallogr. D*, 58:2001, 2002.

[152] E. A. Carter, G. Ciccotti, J. T. Heynes, and R. Kapral. Constrained reaction coordinate dynamics for the simulation of rare events. *Chem. Phys. Lett.*, 156:472–477, 1989.

[153] M. Garcia-Viloca, J. Gao, M. Karplus, and D. G. Truhlar. How enzymes work: Analysis by modern rate theory and computer simulations. *Science*, 303:186–195, 2004.

[154] C. Alhambra, J. Corchado, M. L. Sanchez, M. Garcia-Viloca, J. Gao, and D. G. Truhlar. Canonical variational theory for enzyme kinetics with the protein mean force and multidimensional quantum mechanical tunneling dynamics. theory and application to liver alcohol dehydrogenase. *J. Phys. Chem. B*, 105:11326–11340, 2001.

[155] A. Radzicka and R. Wolfenden. A proficient enzyme. *Science*, 267:90–93, 1995.

[156] G. Winter, A. Fersht, A. J. Wilkinson, , M. Zoller, and M. Smith. Redesigning enzyme structure by site-directed mutagenesis - tyrosyl transfer-RNA synthetase and ATP binding. *Nature*, 299:756–758, 1982.

[157] M. Garcia-Viloca, C. Alhambra, D. G. Truhlar, and J. Gao. Quantum dynamics of hydride transfer catalyzed by bimetallic electrophilic catalysis: Synchronous motion of Mg^{2+} and H^{-} in xylose isomerase. *J. Am. Chem. Soc.*, 124:7268–7269, 2002.

[158] A. Suenaga, A. B. Kiyatkin, M. Hatakeyama, N. Futatsugi, N. Okimoto, Y. Hirano, and T. Narumi et al. Tyr-317 phosphorylation increases shc structural rigidity and reduces coupling of domain motions remote from the phosphorylation site as revealed by molecular dynamics simulations. *J. Biol. Chem.*, 278:4381–4384, 2002.

[159] J. Symersky, M. Monod, and S. I. Foundling. High-resolution structure of the extracellular aspartic proteinase from candida tropicalis yeast. *Biochemistry*, 36:12700, 1997.

[160] S. M. Cutfield, E. J. Dodson, B. F. Anderson, P. C. Moody, C. J. Marshall, P. A. Sullivan, and J. M. Cutfield. The crystal structure of a major secreted aspartic proteinase from candida-albicans in complexes with 2 inhibitors. *Structure*, 3:1261, 1995.

[161] E. T. Baldwin, T. N. Bhat, S. Gulnik, M. V. Hosur, R. C. Sowder II, R. E. Cachau, J. Collins, A. M. Silva, and J. W. Erickson. Crystal-structures of native and inhibited forms of human cathepsin-d: implications for lysosomal targeting and drug design. *Proc. Natl. Acad. Sci. USA*, 90:6796–6800, 1993.

[162] F. Canduri, L. G. V. L. Teodoro, V. Fadel, C. C. B. Lorenzi, V. Hial, R. A. S. Gomes, J. R. Neto, and W. F. De Azeve Jr. Structure of human uropepsin at 2.45 angstrom resolution. *Acta Crystallogr. D*, 57:1560–1570, 2001.

[163] R. B. Rose, C. S. Craik, and R. M. Stroud. Domain flexibility in retroviral proteases: Structural implications for drug resistant mutations. *Biochemistry*, 37:2607–2621, 1998.

[164] J. Kervinen, J. Lubrowski, A. Zdanov, D. Bhatt, B. M. Dunn, K. Y. Hui, D. J. Powell, J. Kay, A. Wlodawer, and A. Gustchina. Toward a universal inhibitor of retroviral proteases: Comparative analysis of the interactions of lp-130 complexed with proteases from HIV-1, FIV, and EIAV. *Prot. Sci.*, 7:2314–2323, 1998.

[165] G. S. Laco, C. Schalk-Hihi, J. Lubkowski, G. Morris, A. Zdanov, A. Olson, J. H. Elder, A. Wlodawer, and A. Gustchina. Crystal structures of the inactive D30N mutant of feline immunodeficiency virus protease complexed with a substrate and an inhibitor. *Biochemistry*, 36:10696–10708, 1997.

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

[167] Z. W. Zhu, D. I. Schuster, and M. E. Tuckerman. Molecular dynamics study of the connection between flap closing and binding of fullerene-based inhibitors of the HIV-1 protease. *Biochemistry*, 42:1326–1333, 2003.

[168] J. Rahuel, J. P. Priestle, and M. G. Gruetter. The crystal structures of recombinant glycosylated human renin alone and in complex with a transition-state analog inhibitor. *J. Struct. Biol.*, 107:227–236, 1991.

[169] T. L. Blundell, J. A. Jenkins, B. T. Sewell, L. H. Pearl, J. B. Cooper, I. J. Tickle, B. Veerpandian, and S. P. Wood. X-ray analyses of aspartic proteinases: the 3-dimensional structure at 2.1 Å of endothiapepsin. *J. Mol. Biol.*, 211:919–941, 1990.

[170] B. Veerpandian, J. B. Cooper, A. Sali, and T. L. Blundell. X-ray analyses of aspartic proteinases. 3. 3-dimensional structure of endothiapepsin complexed with a transition-state isostere inhibitor of renin ad 1.6 Å resolution. *J. Mol. Biol.*, 216:1017–1029, 1990.

[171] A. R. Sielecki, A. A. Fedorov, A. Boodhoo, N. S. Andreeva, and M. N. G. James. Molecular and crystal-structures of monoclinic porcine pepsin refined at 1.8 Å resolution. *J. Mol. Biol.*, 214:143–170, 1990.

[172] I. T. Weber. Structural alignment of retroviral protease sequences. *Gene*, 85:565–566, 1989.

[173] L. Helm and A. E. Merbach. Water exchange on metal ions: experiments and simulations. *Coord. Chem. Rev.*, 1999.

[174] J. Åqvist. Ion water interaction potentials derived from free-energy perturbation simulations. *J. Phys. Chem.*, 94:8021–8024, 1990.

[175] D. E. Smith and L. X. Dang. Computer-simulations of nacl association in polarizable water. *J. Chem. Phys.*, 100:3757–3766, 1994.

[176] H. J. C. Berendsen, J. P. Postma, A. Di Nola, W. F. Van Gunsteren, and J. R. Haak. Molecular-dynamics with coupling to an external bath. *J. Chem. Phys.*, 81:3684–3690, 1984.

[177] S. H. Lee and J. C. Rasaiah. Molecular-dynamics simulation of ionic mobility .1. alkali-metal cations in water at 25° C. *J. Chem. Phys.*, 101:6964–6974, 1994.

[178] N. T. Skipper and G. W. Neilson. X-ray and neutron-diffraction studies on concentrated aqueous-solutions of sodium-nitrate and silver-nitrate. *J. Phys.: Cond. Matt.*, 1:4141–4154, 1989.

[179] L. M. Ramaniah, M. Bernasconi, and M. Parrinello. Ab initio molecular-dynamics simulation of K^+ solvation in water. *J. Chem. Phys.*, 111:1587–1591, 1999.

[180] S. Obst and H. Bradaczek. Molecular dynamics study of the structure and dynamics of the hydration shell of alkaline and alkaline-earth metal cations. *J. Phys. Chem.*, 100:15677–15687, 1996.

[181] R. W. Impey, P. A. Madden, and I. R. MacDonald. Hydration and mobility of ions in solution. *J. Phys. Chem.*, 87:5071, 1983.

[182] N. T. Skipper and G. W. Neilson. K^+ coordination in aqueous-solution. *Chem. Phys. Lett.*, 114:35, 1985.

[183] J. A. White, E. Schwegler, G. Galli, and F. Gygi. The solvation of Na^+ in water: First-principles simulations. *J. Chem. Phys.*, 113:4668–4673, 2000.