SHORT THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (PHD)

THE THEORETICAL STUDY OF THE ENZYME MECHANISM OF PAPAIN AND THE EARLY EVENTS OF THE ACTIVATION OF HUMAN TRANSGLUTAMINASES

by Attila Fekete

Supervisor: Dr. István Komáromi

DEBRECENI EGYETEM

KÁLMÁN LAKI DOCTORAL SCHOOL

Debrecen, 2019

The theoretical study of the enzyme mechanism of papain and the early events of the activation of human transglutaminases

by Attila Fekete, Chemistry MSc

Supervisor: István Komáromi, PhD

Kálmán Laki Doctoral School, University of Debrecen

Head of the **Examination Committee**:	Prof. Zoltán Papp, MD, PhD, DSc
Members of the Examination Committee:	Zoltán Balajthy, PhD
	Ferenc Bogár, PhD

The Examination takes place at the Library of Division of Clinical Laboratory Science, Department of Laboratory Medicine, Faculty of Medicine, University of Debrecen, at 11:00 a.m., 9th of May, 2019.

Head of the **Defense Committee**:	Prof. Zoltán Papp, MD, PhD, DSc
Reviewers:	György Ferenczy, DSc
	Milán Szőri, PhD
Members of the Defense Committee:	Zoltán Balajthy, PhD
	Ferenc Bogár, PhD

The PhD Defense takes place at the Lecture Hall of Bldg. A, Department of Internal Medicine, Faculty of Medicine, University of Debrecen, at 1:00 p.m., 9th of May, 2019.

INTRODUCTION

Two distinct topics have been choosen for the basis of my dissertation. The first one is the theoretical study of the enzyme mechanism of papain and the second one is the study of the human transglutaminases through the example of tissue translgutaminase and the A subunits of coagulation FXIII by the means of molecular dynamics simulations.

In the last few decades the density functional chemistry went through a significant development. The reason of that is the results of density functional computations are comparable to the results of post-Hartree Fock calculations. Moreover the computational time needed for density functional theory is singificantly shorter. The papain enzyme served as an ideal target for the comparison of different density functional methods since it has been studied widely in the last four decades with theoretical and experimental tools. Using density functionals and systematically increasing size of the applied basis sets we presented a group of geometry optimized stationary points on the potential energy surfaces of the chemical reactions of peptide bond hydrolysis and these structures were proved by frequency analysis. Our hybrid quantum mechanical / molecular mechanical ONIOM-EE type calculations embraced the resting state of the papain, the formation of acyl-enzyme intermediate and its hydrolysis as well.

We could not be unconcerned regarding the dynamical behaviour of the biological processes. The classical mechanics can serve as a good help in the study of dynamical events which could not be studied with other experimental methods. Practically such methods can be applied in the events between the picosecond to milisecond timescale of biomacromolecules. Nowadays the upper limit of these simulations are determined by the size of the studied system, the scalability of computational problems and the presence of computational resources. These factors together represent the event window what can serve informations about the studied systems.

At the study of the A subunits of coagulation FXIII (FXIII-A) and the tissue transglutaminase (TG2) those phenomena have been investigated deeply which primarily contribute to their large-scale conformational activation and are in connection with the binding of calcium ions. Similar computational approach has never been applied for the examination of these transglutaminases by this time.

Background of theoretical studies of papain

The papain enzyme (EC 3.4.22.2) belongs to the C1 superfamily of cysteine endopeptidases. In 1970, partly based on the newly revealed crystallographic information, a mechnism has been proposed by Lowe regarding the cleavage of peptide bond on the analogy of serine proteases. By this time, the published results suggest that a generally acceptable mechanism concerning the enzymatic reaction cannot be drawn.

Both the serine and the cysteine proteases cleave the peptid bond via an acyl-enzyme intermediate, what is the corresponding esther or thioesther, respectively. Notable difference between the chymotrypsin and papain is that in the first one the His57, Asp102 and Ser195 form the catalytic triad while in the cysteine proteases the His159, Asn175 and Cys25 residues take place in the formation of active site. Moreover the clearly visible difference of the nucleophile centre, notable point that in the papain an asparagine can be found in the catalytic cavity instead of aspartate. This asparagine residue contributes to the stabilization of the thiolate – imidazolium ion pair in the resting state of papain, however its effect is not so significant as the corresponding aspartate in the serine proteases. A second interesting finding is that the sulphur atom of cysteine can be found in deprotonated state in the pH range of 4 to 8, what is essential regarding the appropriate nucleophile attack.

Assuming that in the resting state of papain-like cysteine proteases the side chains of Cys25 and His159 residues of active site can be found in ion pair, the role of oxyanionic intermediate is not so important. In this particular case the formal charge centres are blurred by contrast the serine proteases where ion pair is formed. The Gln19 residue plays important role in the stabilization of the tetrahedral intermediate and its loss effects slightly the first acylation step of the peptidase reaction.

In the last four decades, several theoretical studies were published in order to clarify the details of the cysteine protease mechanism. Some of the results suggest different mechanisms, where the protonation from the histidine ring to the amide nitrogen is either happens concertic with the attack of thiolate nucleophile or precede that. It is also notable, that in some case the cystein of the active site was found to be protonated. Suhai et al. was found that the tetrahedral intermediate is stable only when the amide nitrogen has become protonated from the Nδ atom of the imidazolium ring.

It was investigated in different protein environments that in case of cysteine-histidine diad where has been found the moving hydrogens. An obvious idea was to investigate the role of water molecules and substrates to the transition of hydrogens. With the aid of benchmark QM/MM

calculations the effect of continuum dielectric solvent was also studied and compared their results to those obtained by gas phase calculations. In the condensed phase examination of simple modell systems it was found that the hemi-thioacetate anion corresponds to the tetrahedral intermediate, although in the case of serine proteases it was not found as a stable intermediate. Ma et al. pointed out that on the potential surface of acylation of cathepsin K the tetrahedral intermediate represent a local minima, while they were not found similar intermediate for the deacylation step. Wei et al suggested a one step reaction for the acylation step and they found that the acyl-enzyme hydrolysis was going through the formation of an other intermediate.

Regarding the level of theories applied in the studies of cysteine proteases, the highest level of computations were carried out with B3LYP and B-LYP density functional theories (DFT) and they were supplemented with double or triple zeta basis functions with polarized and diffuse functions. For optimization of stationary points usually an even lower level theories were applied. In hybrid QM/MM calculations for the QM layer also a lower level of theories were used.

The choice of DFT methods seems ideal as a QM component of hybrid calculations, since regarding the relative and activation energies and also the geometrical properties the achievable performance can be compared to costly post-HF methods. Although the density funcionals have a serious deficiency, namely the description of long range exchange and dispersion interactions are not correct.

Recently several quantum chemistry softwares contains DFT methods in which the London dispersion energy was taken into account. This energy term can calculated easily through the C_6/r^{-6} pair-terms. The ωB97X-D and B97D functionals now contains such a terms. The former contains also a term for the long range exchange of electrons, what can explain its good performance in van der Waals complexes. Another solution can be the use of funcionals which do not contain explicit term for the dispersion correction but use built-in parameters which were optimized on large data sets. The M06-2X functional uses 35 fine-tuned parameters, hence it serves as a typical example for the latter functionals and also performes well in the calculations of activation barriers and transition states. Generally the simple pair-like dispersion correction significantly improves the calculation of activation barriers in biochemical QM/MM calculations.

It was pointed out the the level of DFT theory and also the quality of basis functions significantly affect the activation barriers, geometries of transition states and the assinchronicity of reactions.

Goals of papain study

At the study of the enzyme mechanism of papain we appointed multiple goals. The main goal was to compare the performance of the used density functionals together as a QM component of hybrid ONIOM type calculations. We wanted to present a comparison of dispersion corrected functionals to the „gold standard" B3LYP method. The existance of diverse reaction paths depending the applied theories was also in the centre of our work. Such a major indicator can be the existence of oxyanionic or zwitterionic intermediates, since this question was not clarify by this time.

In order to verify that the found stationary points represent really a local minimum or first order saddle points on the QM/MM (ONIOM) potential energy surface, the second geometrical derivatives were calculated in every case. All of the calculations were performed with the „Electronic Embedding" (EE) ONIOM what was implemented into the Gaussian quantum chemistry suite.

Background of the studies of transglutaminases

Both blood coagulation factor XIII subunit A (FXIII-A) and tissue transglutaminase (TG2) are protein-glutamine gamma-glutamyltransferase enzymes (EC 2.3.2.13). In the plasma, FXIII (pFXIII) is circulating as a heterotetramer with two carrier B subunits (FXIII-A_2B_2), while its cellular form (cFXIII) exists as a homodimer (FXIII-A_2) in the cytoplasm of platelets and monocytes. The pFXIII requires proteolytic activation by thrombin which cleaves the peptide bond between Arg37 and Gly38 of FXIII-A_2, which then provides the activation peptides (here denoted as: APs) and the FXIII-A_2'. Crystallographic studies suggest that the cleaved N-terminal peptides remains attached to the surface of the bovine α-thrombin activated homodimer form, blocking the entrance to the cavities where the catalytic residues are located (Cys314, His373, Asp396). It has been previously noted that the pFXIII can be activated without proteolytic cleavage under non-physiological conditions using high Ca^{2+} concentration (≥ 50 mM). Whereas the cellular FXIII (cFXIII), derived from placenta can be activated either proteolitically by thrombin and low Ca^{2+} level or with high sodium-chloride (NaCl) or potassium-chloride (KCl) concentration in the presence of 2 mM Ca^{2+}. Even in the absence of any Ca^{2+} a slow progressive activation can still be observed. Based on resolved active transglutaminase 2 structure large conformation changes were predicted in order to obtain the enzymatically active FXIII-A_2* form. Such major domain motions were published by Stieler et al. where the recombinant cellular FXIII (rcFXIII) was activated

without proteolytic cleavage (FXIII-A^o) using 300 mM sodium-chloride and 3 mM Ca^{2+} concentrations. The displacement of β-barrel 1 and β-barrel 2 domains caused no significant change in the structure of the other two domains, however the β-barrels were significantly displaced providing an accessible catalytic cavity for substrate molecules. Nevertheless, domain movements in this structure differs from those that were suggested based on the analogy with TG2. In summary, calcium ions play a role in the dissociation of FXIII-B_2 subunits from FXIII-A_2 and also to induce conformational activation of FXIII-A_2. It was suggested that FXIII-A dimers are responsible for Ca^{2+} binding (K_m= 0.12 mM, based on fluorescence measurement) and were estimated to bind 1.2-1.5 calcium ions based on equilibrium dialysis experiments. The hypothesis also proposed the existence of two strong binding sites per symmetrical dimer, along with some other weaker binding sites. Crystallographic information can be found in the literature for the location of Ca^{2+} binding sites and even for binding Sr^{2+} or Yb^{3+}. Two symmetrical Ca^{2+} and Sr^{2+} binding sites can be found near residues Ala457, Asn436, Asp438, Glu485 and Glu490. It is worth noting, that in the crystallographic structure published by Fox et al., there are eight Yb^{3+} ions present with a total of four Yb^{3+} bound between the Asp270 and Asp271 residues of both monomer units. Interestingly, this is not one of the known calcium binding sites, however a recently published active FXIII-A structure also seems to show a single Ca^{2+} bound to this site. Site-directed mutagenesis studies were carried out in order to validate the importance of Glu485 and Glu490 of main binding sites in the activation process and whether residues Asp476 and Asp479 play a role in the conformational changes indirectly via a switch between the protease-sensitive and protease-resistant form. Surface polarity analysis and ^{43}Ca NMR experiments also suggest that weaker binding sites exist and the dissociation constant of calciums for zymogen FXIII-A_2 is 0.51 mM, while in the presence of thrombin the Kd increased significantly (5.9 mM), based on NMR experiments. Due to claims regarding the limitations of NMR techniques hydrogen/deuterium exchange experiments were conducted in order to examine the percentage of deuteraion of four different FXIII-A_2 systems (zymogen, thrombin-cleaved, thrombin cleaved + 1 mM Ca^{2+}, and the zymogen activated by 50 mM Ca^{2+}) using MALDI-TOF mass spectrometry, which resulted in the identification of several regions with elevated deuteration. The *in vitro* activated zymogen FXIII-A^o structure has been published by Stieler et al. in the presence of a covalent inhibitor and using high ionic strength. Within this study they have found that β-barrel 1 and β-barrel 2 undergoes remarkably large movements and exposes its active site. More recently a detailed structure/function study has been conducted in which the generated pathway between the active and inactive conformations were studied and the interaction between the A and B subunits was also modelled. More detailed information can be found about the structure and function of FXIII in numerous other relevant publications.

Tissue translgutaminase (TG2 or Gh protein), another member of the TG group of enzymes, is present in different cellular compartments including the plasma membrane, cytosol and the nucleus although as a monomer. It is common for both TG2 and FXIII-A that activation only occurs in the presence of Ca^{2+}, however structural evidence is available in the absence of Ca^{2+} and presence of GDP, GTP or ATP. The same lack of Ca^{2+} presence has been reported even with the active conformation too. TG2 possesses GTPase activity and the binding of guanosine-phosphates regulates the transamidation activity in a negative way. Interestingly, based on equilibrium dialysis experiments, TG2 can bind up to six Ca^{2+} which can decrease with the binding of GTP. Modest sequence homology can be observed between the eight known human TGs (Stieler et al., 2013), but they share a highly conserved secondary and tertiary structure. Using the known structure of FXIII-A_2, as a base template for homology modelling, 50 ps unconstrained MD simulations were carried out in the absence of explicit solvent molecules both with and without Ca^{2+} (3 pcs.). In this, it was suggested that the short simulation time and low stoichiometric ratio were the main cause for the lack of large scale movements, although both simulations predicted that the added calciums would increase the gyration radius. When digesting the enzyme in limited proteolysis experiments, it was found that only the loop region between the catalytic and the first β-barrel domain has been cleaved. Results of ^{43}Ca NMR investigations suggested the presence of some low affinity sites, more interestingly the Kd was seemed to be equal to that of thrombin activated FXIII-A_2 (6.0 mM versus 5.9 mM), hence these two enzymes can follow similar activation mechanism. In a detailed *in vitro* experiment five calcium binding sites were identified following systematic mutations of certain residues within the putative binding sites. These sites were established based on computer modelling and structural analogy with FXIII-A and epidermal translgutaminase (TG3). It should be noted that the 1KV3 structure contains a glycine residue in the position of 224 while its native form contains valine in the very same position (https://www.uniprot.org/uniprot/P21980). As of now, there are no known structures available of the native closed TG2 except 4PYG (Jang et al., 2014), which has the native Val224, but contains three other point mutations, and the reasons behind these were not discussed in the original article. Kanchan et al. pointed out, that there are several differences between the calcium affinity, GTPase activity and other measures of these different TG2 forms. The open conformation of TG2 was resolved by Pinkas et al. via incubating the enzyme in 150 mM NaCl, 10 mM Ca^{2+} and in the presence of an irreversible, modified-pentapeptide inhibitor (Pinkas et al., 2007). A recent review concerning the protein-protein interactions of the TG2 can be found in the literature.

In summary, coagulation factor XIII and TG2 represent well studied systems, however, several details of their calcium induced activation process still remains unclear. In this present work

a series of microsecond long all-atom MD simulation were conducted on both the zymogen and on the computationally cleaved FXIII-A_2', both in the presence and absence of Ca^{2+} cofactors and also using activation peptide free proteins. Individual and combined effects of Ca^{2+} and GDP were also considered in the study of TG2, then the outcomes of these MD simulations were compared to existing *in vitro* experimental results. The effect of the Gly224Val mutation on calcium binding affinity and overall protein dynamics has also been investigated within this study.

Goals of transgluatminase study

Regarding the large number of factors which may play role in the calcium induced activation and in the large scale conformation rearrangement of FXIII-A_2 and TG2 we set the goal, namely the detailed investigation of calcium binding and its impact to dynamics of the proteins are necessary in order to shed light onto the early phase of the activation. In the case of FXIII-A_2, we wanted to clarify the individual details of zymogen, proteolytically cleaved forms and also the FXIII-A_2 alone without activation peptides. With similar attention we also would like to examine the calcium binding properties of TG2 and compare our results to the existing experimental ones which were based on site-directed mutagenesis. Based on covariance matrices and principal component analysis we wish to identify those structural elements which may have emphasized importance through the activation mechnism. As a supplement to our extensive *in silico* studies we would like to check out the relevance of the formed binding sites via the usage of „multi-site" calcium modell.

APPLIED METHODS

Study of the enzyme mechanism of papain

Starting from a structure available in the protein data bank (PDB ID: 1PPN), short (250 ns) molecular dynamics simulations on papain were carried out. Both the neutral and ion pair forms of its catalytic dyad were considered in order to reveal how the geometries of the catalytic sites change during simulations. In the constant particle number, constant pressure and constant temperature (NPT) simulation dodecahedral periodic box, a TIP3P explicit water model and 150 mM ionic strength (set by Na^+ and Cl^- ions) were used. The FF99SB force field with a Berendsen barostat and a v-rescale thermostat were applied in these calculations, which were carried out by using GROMACS software. For the short range electrostatic and van der Waals energy terms 10 Å cut-off distances were used, while the particle mesh Ewald (PME) method was applied for long-range electrostatic energy corrections. The initial QM/MM model was also constructed from the 1PPN X-

ray structure of papain. Because no significant deviation from the starting (optimized) catalytic center geometry was observed in dynamics simulations with the exception of thermal fluctuation, the AMBER 12 sander optimized structure was used as the starting geometry in all the ONIOM calculations instead of a selected frame from the dynamics trajectory. At first, the whole structure was truncated that way so that only the catalytic center and its neighborhood (Val16-Ala30, Ser60-Trp69, Val130-Ala136, Lys156-Ala162, Ile173-Gly178, Thr204-Phe207) were retained, which comprised all the amino acid residues known to have considerable influence on the enzyme reaction. During truncation, whole amino acids were always retained. The runcated N- and C-terminal peptide bonds were closed by acetyl or N-methyl caps, respectively. The orientation of the N-H and C=O bonds in these ‘‘caps’’ corresponded to those peptide bonds that were cut previously. This way, the first model system with an N-methylacetamide (NMA) substrate consists of 780 atoms. The NMA substrate has been placed at the active site by means of the xleap module of the AMBER package. In the new structure constructed this way, the QM (‘‘model’’) subsystem is comprised of the side chains and C a atoms of His159 and Cys25 residues as well as the substrate model NMA when it was considered. During the optimizations and potential energy scans the Gln19-Gly-Ser-Cys-Gly-Ser-Cys25, Gly62-Cys-Asn-Gly-Gly66, Asp158-His159, Asn175-Ser-Trp177 residues and the NMA substrate were allowed to move, while all the other atoms were fixed.

First, calculations were carried out on this truncated system. When all the transition states and local minima were identified, their geometries were fitted back to the whole protein by applying the VMD package, and the calculations were repeated. In this paper, only the results corresponding to the whole protein and the NMA substrate will be discussed.

For the QM (‘‘high level’’) methods the B97D, M06-2X, ωB97XD and B3LYP functional forms and the standard Pople style 6-31G(d,p), 6-31+G(d,p), 6-311+G(d,p) and 6-311++G(d,p) basis sets were applied. For the MM (‘‘low level’’) method the AMBER force-field implemented in the Gaussian packages was used. Electronic embedding approximation was applied throughout the ONIOM QM/MM calculations, i.e. the QM methods were polarized by the surrounding MM partial charges through the modified $1e^-$ Hamiltonian. The standard link atom approach using hydrogen atoms to saturate dangling bonds was applied for covalent bonds in the QM-MM borderline region. At all local minima and transition states we obtained the second geometrical derivatives calculated at the corresponding level of theory. Here, we report only ONIOM type QM/MM calculations. Therefore, mentioning the QM method should be regarded as the abridgement of the full QM/MM calculations.

It is generally accepted that for polar solutes (like proteins) in a polar solvent (e.g., water which contains even charged particles), the electrostatic part of solvation energy is the dominant one, which can be satisfactorily estimated using the Poisson–Boltzmann method. Therefore, the electrostatic solvation free energies were calculated for the protein taking into account both the neutral and the ion-pair forms of the catalytic triad. A thousand frames from the last (equilibrated) 100 ns simulations were selected equidistantly and, for each structure, the electrostatic solvation free energy was calculated using the DelPhi software. For the calculations, the Amber FF99SB force field partial charges and the standard particle sizes provided with the DelPhi package were applied. The averaged solvation free energies for both the neutral and the ion-pair forms can be considered as a further correction to the zero point corrected ONIOM energy values obtained at different levels of theory.

The conversion between the PDB and Gaussian input file formats was performed by using the TAO package. The ONIOM type QM/MM calculations were carried out by using the Gaussian 03 and Gaussian 09 software packages. Chimera software was used for molecular graphics representations. The potential energy surfaces (PESs) for the first step of acyl–enzyme formation were generated along the (Cys)Sγ...C(NMA carbonyl) vs. the (HisNδ)Hδ...N(NMA amide) bonds as well as along the NMA amide N-C vs. both the (NMA)N–Hδ(HisNδ) and (Cys)Sγ...C(NMA carbonyl) bonds using 0.05 Å adjoining grid point distances. PESs for the deacylation steps were also scanned along the (water)O...C(NMA carbonyl) vs. the (HisNδ)...H(water) bonds using the same grid point distances. The surface plots were generated by using the DPlot software. Only the B3LYP and M06-2X methods with 6-31G(d,p) and 6-31+G(d,p) basis sets in ONIOM were used for surface mapping.

Study of the early events of the transglutaminase activation

The simulation systems

In order to examine the effect of Ca^{2+} on the dynamics of FXIII-A_2, the following simulation sets were assembled. The first simulation set (i) was used to study the zymogen enzyme via the calcium bound form of FXIII-A_2 (i/a) (PDB ID: 1GGU) and in addition to the two bound calcium ions, the total Ca^{2+} concentration was set to 14 mM. A second zymogen model was created from the 1F13 PDB structure (i/b) and a total of 50 mM Ca^{2+} was added into the simulation cell. To clarify the important structural features of the thrombin treated FXIII-A_2', the peptide bond between Arg37 and Gly38 was cleaved using computational methods and from these four different model systems were constructed. The ii/a and ii/b model systems were analogous of i/a and i/b, respectively. The ii/

c model was similar to ii/b but contained only 14 mM Ca^{2+} (practically 20 calcium ions) while the final model of this set of four contained only the neutralized protein with a total of 150 mM NaCl (ii/d) and was used as a reference. We have suspected that after the proteolysis of the APs, the next step in the activation process should be the relocation or at least partial dissociation of the APs in the presence of calcium. The simulations in the absence of APs (FXIII-A_2') were carried out under different conditions; with no calciums present at all (iii/a), with 14 mM Ca^{2+} (iii/b) and with 1000 mM Ca^{2+} (iii/c), in order to examine the protein dynamics under extreme ionic strength.

In the fourth simulation set, the careful investigation of the individual and combined effects of calciums and guanosin di- or triphosphates were studied on TG2 systems. In the reference simulation the bound GDP was removed and enzyme solvation was achieved in the presence of 150 mM NaCl (iv/a), and also containing 8 mM Ca^{2+} (iv/b), which correspond to 10 calcium ions, then we also repeated these two simulations in the presence of GDP (iv/c,d) as well. Due to the suspected importance of Gly224 the final model of this set (iv/e) was based on GTP bound TG2, in the presence of 6 mM Ca^{2+} (practically 8 calcium ions). All simulations of iv/a-d were based on the 1KV3 crystallographic structure, however iv/e was based on the 4PYG X-ray structure. The final simulation set (v) contained the open conformation of the FXIII-A^O monomer (PDB ID: 4KTY) with three bound calcium ions (v/a) and the calcium free open conformation of TG2 (v/b). Both of these model systems contained 14 mM Ca^{2+} concentration and were considered to be interesting cases, but they did not couple closely to any part of this present study. Throughout the work we have numbered the residues according to existing literature (*id est* the first methionine was omitted in the case of FXIII-A but not in the case of TG2) and the recommended abbreviations of Muszbek et al. were used for factor XIII, where it was possible.

Molecular dynamics simulations

The protein models were solvated using explicit TIP3P water molecules in octahedral boxes such a way that the closest distance between the box and the protein was 12 Å. In the next step all systems were neutralized and the NaCl concentration was set to 150 mM/dm^3. After a short energy minimization step, a 2 ns long simulated annealing equilibration was carried out, where the minimized systems were heated up to 310 K and the protein heavy atoms were kept restrained by a force constant of 1000 $kJmol^{-1}nm^{-2}$. Following the equilibration procedure, model system dependent, one or two microsecond long isobaric-isothermal (NpT) production runs were performed using periodic boundary conditions with the aid of virtual sites that allowed a 4 fs step size for

integrating the Newtonian laws of motions. All bonds were constrained using the LINCS algorithm. In simulations where guanine di- or triphosphate was present, a 2 fs step size was applied and these simulations lasted 1 μs. A cut-off of 10.0 Å was used for the Lennard-Jones and short-range electrostatic interactions and a force-switch was applied to smoothly switch forces between 7.0 Å and 10.0 Å. The long-range electrostatics interactions were calculated by Particle Mesh Ewald summation. For the coupling of thermal bath the velocity-rescaling method was used, while pressure was regulated with the isotropic Parrinello-Rahman method. The Amber99SB-ILDN-NMR force field was used throughout the simulations since it was successfully used in the study of calcium and magnesium binding features of small antifungal proteins. In systems where TG2 was complexed with GDP or GTP, the AMBER GAFF parameters were used for the substrates. A total of 21 μs all-atom MD simulations were carried out in this study to be able to examine the details of the early events of calcium induced activation of human TGs.

In two cases (ii/c and iv/d) an additional 50 ns NpT MD simulation was carried out with the Amber ff14SB force field and using multi-site Ca^{2+} models in order to simulate more precisely the already established binding sites. The starting structures were extracted from the very end of the corresponding simulations and all bound calcium ions were kept as is. If any virtual interaction sites were present, they were removed. The XZ (central) atom of the multi-site Ca^{2+} model was aligned to all of the bound calciums. The reconstructed systems were solvated in an octahedral box filled with TIP3P water molecules and the final NaCl concentration was set to 150 mmol/dm^3. A 10.0 Å cut-off was used over the initial energy minimization step, the 1.7 ns long (3-stage) equilibration protocol and the 50 ns long production runs.

The MD simulations were performed using GROMACS 5.1.4 and with the pmemd software of the Amber16 package. The trajectories were analysed with the software tools of GROMACS and the cpptraj program. The dynamic cross-correlation matrices (DCCM) were calculated over 12.500 frames with the aid of the Bio3D v2.3 R package. R was used to perform all matrix operations as well. For visualization of the protein structures either UCSF Chimera 1.11.2 software or Visual Molecular Dynamics (VMD) 1.9.1 was used. All unresolved segments of the protein structures were reconstructed with MODELLER 9.10 and only those models were used which had the best DOPE score via the graphical unit interface of UCSF Chimera. It should be noted that the conformation of the missing loop region (between Thr508 and Ser516) of 1GGU was found to have an overall importance, therefore, it was remodelled as it can be found in the 1F13 structure with the loop of the A chain of 1F13 placed to the B chain of 1GGU and vice versa. All of the data that can be seen in this present work was plotted by various in-house written Python scripts using matplotlib and numpy. The protonation states of titratable residues were predicted using the H++ webserver at

pH=7.4, however, it should be note that all of the aspartate and glutamate side-chains of TG2 systems were predicted to be deprotonated (i.e. negatively charged).

Analysis of molecular dynamics simulations

Frames of the dynamic trajectories were saved every 80 ps, and contained the corresponding protein plus Ca^{2+} ions and any GDP or GTP if any were present. For the analysis of root mean square deviation (RMSD), root mean square fluctuations (RMSF) and radius of gyration (r_{gyr}), the heavy atoms within the protein were used. In order to ensure that the N-terminal activation peptide does not interfere our conclusions, the RMSD and r_{gyr} values were calculated for residues between Gly38 and Met731 in the case of FXIII-A_2 simulations. For dimers, the per-residue averaged values of RMSFs were calculated per individual chain and then averaged. For the calculation of DCCM matrices the coordinates of Cα atoms were used which yielded symmetrical N x N matrices. Since the interpretation of correlation matrices can be quite difficult in such a large proteins, we have tried to focus only on the difference DCCM matrices ((c_{ij}(2)-c_{ij}(1), where c_{ij}(1) was taken as a reference)), however all reference matrices are also included in the SI. The Ca^{2+} binding properties were based on the calculation of distance between the Cγ atom of aspartate or the Cδ atom of glutamate and any Ca^{2+}. The large amplitude low frequency motions were extracted from the Cα trajectory for principal component analysis and the first 10 eigenvectors were calculated via the ProDy software. For helping the interpretation of the domain motions of monomers relative to each other, the centre of masses of main structural domains were used to calculate a pseudo torsion angle and the standard deviations were also indicated in each case.

NEW RESULTS

The enzyme mechanism of papain

The resting state of papain

The trajectories of the 250 ns molecular dynamics simulations on free papain with both the ion-pair (zwitterionic) and the neutral forms of the Cys–His catalytic dyad indicate that these systems are equilibrated during the first 100–150 ns time frame. The simulations also show that despite the significant fluctuation in the (Cys25)Hγ and Nδ(His159) distance in the neutral form of papain, the shortest values (1.8–2.0 Å) that correspond to a usual H-bond distance are still significantly populated. This suggests that the proton transfer from the (Cys25)Sγ to the Nδ(His159) can occur even without any explicitly contributing water molecule, therefore, such a water

molecule(s) was(were) not considered in our calculations. On the other hand, the already transferred proton also resides with high probability in a position which corresponds to the (His159Nδ)Hδ...Sγ(Cys25) hydrogen bonded distance.

When analyzing the trajectories, significant differences can be noticed between them regarding the number of water molecules in contact with the Cys25Sγ atom. While the neutral Cys25Sγ can have close contact typically with 0–3 surrounding water molecules, the ion pair form Cys25Sγ can make at least 2–4 such close contacts. This suggests the possibility that the solvent contributes to the stability of these forms in different degrees. Interestingly, apart from the thermal fluctuation, the geometry of the Cys–His dyad and its proximity, either in their (His159)Nδ...HγSγ(Cys25) or in their (His159)NδHδ...Sγ(Cys25) hydrogen bonded forms, have considerable similarity to the starting X-ray structure.

The existence of the (His159)Nδ...Hγ-Sγ(Cys25) H-bond connections during the whole simulation leads to the plausible assumption that in static ONIOM calculations we can start from the optimized H-bonded structure. In all cases, the ONIOM QM/MM calculations without zero point energy correction resulted in approximately the same energy values for the neutral and ion pair (thiolate and imidazolium) forms of the Cys25–His159 side chains. The differences between the calculated energies are not greater than 0.8 kcal mol -1. The ωB97XD method predicts the ion pair form to be the most stable one, while the other methods showed mixed results. The M06-2X values are closer to those obtained by the ωB97XD method.

Interestingly, the B97D and B3LYP methods resulted in very similar energy differences between these states with the same basis-set dependence for the basis sets we used in these calculations. It is worth noting here that some earlier calculations also showed the ion pair form to be more stable than the neutral one. Applying the zero point energy correction, the neutral form becomes the most stable one in all kinds of calculations we carried out. Nevertheless, the corrected energy difference between the two characteristic states remains small, less than 2 kcal mol^{-1}, for all cases. For the ion pair - neutral cysteine–histidine side chain conversion, the methods we applied predicted a lower ZPVE corrected energy for the transition state than the corresponding ion-pair endpoint. It formally means that the ion-pair spontaneously turns into its neutral form even at 0 K and without any extra (activation) energy. These zero point corrected theoretical energy values are in conflict with previous experiments which demonstrated that the ion pair state is more populated. However, from these experiments certain (albeit significantly less than 50%) probability could be assigned to the co-existence of the neutral form or, it is supported that, at least, such a neutral form cannot be excluded. From this observation one can conclude that the (free) energies of the ionic and neutral forms should be close to each other.

The importance of solvent molecules on the relative stability of such an ion pair is demonstrated by sophisticated QM/MM computations that were carried out on cathepsin, which is homologous to papain. It was previously shown that the ion pair form was the most stable one only when its interactions with the surrounding water molecules were also considered. Therefore, we have performed electrostatic solvation free energy calculations on both the ion-pair and neutral cysteine–histidine side chains as well in order to estimate the effect of solvation. Based on the averaged values from the last 100 ns simulations, the PB calculations predicted a larger electrostatic solvation energy term for the ion pair form than for the neutral one by about 20 kT (i.e., about 11.9 kcal mol^{-1} at 300 K). These solvation (free) energy values from the PB computations can be used as a further correction to the corresponding zero point corrected ONIOM energies. The correction is large enough to reverse the relative stability of the ion pair and neutral forms and predicts the previous one to be the most stable one again in all our cases. While the method we used for correction differs substantially from the one used by Mladenovich et al., our results also underline the importance of the solvent in the stabilization of the ion pair form.

The data show that all the density functional methods applied in ONIOM resulted in very similar values for the S-H and imidazol neutral side chain pair. Note, however, that the B97D method predicted a significantly shorter Nδ...Hγ distance and a little longer Sγ...Hγ distance, i.e., a stronger Nδ...Hγ-Sγ hydrogen bond. The importance of the size of the selected basis was negligible. It is noteworthy that the B97D method predicted the longest S–H distances. The values obtained for the transition states spread over a slightly larger range. The d1 and d2 distances in the neutral- and ion pair forms, as well as in the transition state associated with their conversion, are in good agreement with the published theoretical values for the papain and N-acetyl-Phe-Gly-4-nitroanilide enzyme substrate complex. It should be emphasized that the d1 and d2 distances for the ion-pair are in good agreement with the averaged values obtained using sophisticated QM/MM dynamics calculations, while the corresponding values for the neutral form are significantly different from those reported. The latter discrepancy can be explained by the more flexible nature of the neutral form resulting in larger deviations, i.e., larger averaged values compared to the static equilibrium ones. The larger neutral (His159)Nδ...Hγ-Sγ(Cys25) H-bond distance fluctuation can be observed in our simulations as well.

Formation of acyl-enzyme intermediate

According to our calculations a (zwitterionic) tetrahedral intermediate exists, therefore, the transition state energies compared to the energy of the Michaelis complex, the tetrahedral

intermediate, as well as to the energy of the products. Four characteristic distances are listed for this particular acylation process. These are shown at each of the stationary points (Michaelis complex, tetrahedral intermediate, product and the two transition states between them). In contrast to these results, carrying out B3LYP/6-31G(d) reaction path modeling which was augmented with B3LYP/6-31++G(d,p) energy calculations and pseudo-bond free energy estimation in the framework of the QM/MM method, Wei et al. proposed a one elementary step mechanism for the crucial acyl enzyme formation without any stable (tetrahedral) intermediate state. With regard to the Michaelis complex geometry, the bond length parameters demonstrated only marginal dependence on the levels of theory at which they were derived. On the other hand, the non-bonding distances represented more characteristic method dependence. The longest values for both the Sγ...C(peptide carbonyl) and (His Nδ)Hδ...N(peptide amide) distances were derived using the B3LYP method. The reason for the elongated distances is probably that the B3LYP method does not include appropriate terms for long range interactions, while the other methods do. The most glaring non-bonding distance differences can be observed using the B3LYP and M06-2X methods. Comparing our characteristic distances with those that were published recently by Wei et al. for the Michaelis complex of papain and N-acetyl-Phe-Gly-4-nitroanylide, the chemical bond lengths are similar to each other, although their reported amid bond length (1.37 Å) is a little longer than ours. Similarities between the previously published and our present non-bonding distances can also be observed with the exception of the significantly longer (3.7 Å) Sγ...C(peptide carbonyl) separation calculated by Wei et al.

The papain-NMA Michaelis complex is stabilized by the H-bonds between the NMA carbonyl oxygen and the Gln19 side chain amid group as well as the Cys25 backbone amid H. In the appropriate orientation of the Gln19 side chain the Trp177 and Gly23 residues play a significant role. It should be mentioned that the imidazole ring of His159 and the (Cys25)S atom are approximately in a common plane. It is immediately apparent that the tetrahedral intermediate corresponds to a formal zwitterionic structure, i.e. a single O atom is connected to the tetrahedral carbonyl C atom and at the same time the amide group is protonated. This is in contrast to the anionic TI structure obtained from calculations carried out for serine proteases. This implies that not only the sulfur attack on peptide carbon but also the proton transfer from the histidine to the peptide (amide) nitrogen take place in the same elementary step. However, comparing the d1 (Sγ...C(carbonyl)), d3 (amide(N)...Hδ(His)) and d4 (H...Nδ(His)) distances at the transition state it is evident that the first elementary step of the acylation process is not fully synchronous.

Interestingly, the B3LYP method using polarized Pople-type split shell double- and triple zeta quality basis sets augmented with diffuse functions either on heavy or on both heavy and hydrogen atoms predict a different type of asynchronicity than all the other methods we applied. In

the former case, the d1 (S...C) distance is close to the value that should exist in the anionic (serine proteinase analogue) TI structure and the d3 and d4 distances (i.e., the proton transfer) feature intermediate (transition state like) values. This means that the transition state resembles the structure which connects an already formed serine-protease-like tetrahedral intermediate to the zwitterionic one. By all the other methods, the transition state is mainly determined by the S...C bond formation and the H...N(imidazolium) distance is close to the value which was observed in the Michaelis complex. It should be underlined that the proton transfer in neither of these cases precedes the S...C bond formation. The B97D method predicts significantly longer S...C distances for the transition state structure than all the other levels of theory we applied (independently from the quality of the basis sets we used). The same trend was observed, although to a less extent, for the tetrahedral intermediate structure as well.

The transition state(s) and the TI are also stabilized dominantly by the same H-bonds which were observed at the papain-NMA Michaelis complex. A significant difference can be observed, however, in the orientation of the imidazole ring of the His159 residue. In contrast to the roughly parallel orientation of the NMA amid bond and the imidazole ring in the Michaelis complex, the imidazole ring turned away from its original position and now the Nδ...H bond points to the N atom of NMA in the TS. The imidazole ring of His159 and the (Cys25)S atom are no longer in a common plane.

The B3LYP/6-31G(d,p) method has an exceptional feature among the methods we used in the ONIOM calculations. At this level of theory, an additional transition state can be localized in the pathway which connects the Michaelis complex and the zwitterionic tetrahedral intermediate. TS1a is similar to those transition states which were found at the dispersion corrected density functionals. However, in this case it leads to an intermediate which resembles the anionic (serine protease-like) tetrahedral intermediate. The next (separate) step on the potential energy surface is the formation of the zwitterionic tetrahedral intermediate from the anionic structure. Interestingly, the sole TS on the Michael complex - zwitterionic tetrahedral intermediate pathway found by the B3LYP method using larger basis sets roughly corresponds to the anionic intermediate - zwitterionic intermediate transition.

With regard to the energy values, it is immediately apparent that on the pure (i.e., not ZPVE corrected) potential energy surface the two transition states (TS1a and TS1b) and the (oxy)anionic tetrahedral intermediate between them have almost the same energies. The zwitterionic intermediate is considerably more stable than the anionic intermediate. Applying ZPVE corrections the barrier which separates the anionic intermediate from the zwitterionic tetrahedral intermediate disappears. This means that in our case the barrier has only theoretical importance. One can also notice that the

amide(peptide) bond breaking energy barrier (TS2), which results in the acyl-enzyme product, is predicted to be significantly lower by the B3LYP method than by all the other functional–basis set combinations we used. On the other hand, it implies that depending on the chemical structure of the substrate, the chemical environment and even the optimization methods applied, these tetrahedral/anionic intermediates simply do not exist or can be easily missed. This may explain why Wei et al. found the acylation step to be a single elementary step when using the B3LYP functional in the QM/MM calculations, in contrast to our present results. The potential energy surface (PES) we have derived can explain why the B3LYP/6-31G(d,p) and M06-2X/6-31G(d,p) methods predict different TS structures and also why it was extraordinarily difficult to find the transition state on the B3LYP/6-31G(d,p) (and all the other B3LYP) surfaces. On the B3LYP/6-31G(d,p) PES there are two transition states, TS1a at d1 = 2.7 Å and d3 = 1.8 Å, and TS1b at d1 = 2.26 Å and d3 = 1.47 Å.

The first one connects the Michaelis complex to the ‘‘classic’’ anionic intermediate, while the TS1b connects the anionic intermediate to the zwitterionic intermediate. In the extremely flat B3LYP PES region, the position and even the existence of transition state(s) can be influenced by small geometry perturbations or changes in the basis sets. In contrast, the M06-2X PES has a much more characteristic saddle point. Because geometry optimization is performed at each grid point with the exception of the grid variables, only the transition states and their proximity were mapped and, e.g., the Michaelis complexes can be found outside of the map. Nevertheless, unconstrained geometry optimization from the grid points leads either to the Michaelis complex or to the product with the exception of the B3LYP/6-31G(d,p) PES as expounded above. It should be stressed that there was no spontaneous C-N (amide) bond breaking in the substrate during the PES scan. However, during the potential energy scan (see the details above) only the (amid)N...Hδ(His159) and the (Cys25)Sγ...C(amid carbonyl) distances were scanned and only the C-N(amid) bond was optimized. In order to examine how other choices of variables can influence the potential energy, similar scans along the C-N(amid) and the (Cys25)Sγ...C(amid carbonyl), as well as the (amid)N...Hδ(His159) distances were also carried out. Interestingly, during the (amid)N...Hδ(His159) and C-N(amid) scan the (Cys25)Sγ...C(amid carbonyl) bond spontaneously formed at short (amid)N...Hδ(His159) distances and the transition state (amid)N...Hδ(His159) geometry parameters estimated from these surfaces are approximately the same as we obtained from the transition state search based on the surface scan expounded previously. While this is not an exact and unambiguous proof of the acylation reaction mechanism(s) we proposed above, it can be regarded as further support for them. It should also be mentioned that the C-N(amid) and (Cys25)Sγ...C(amid carbonyl) as well as the (amid)N...Hδ(His159) and C-N distance scans showed substantially weakened C-N(amid) bonds at the B3LYP/6-31G(d,p) level compared to those we

obtained at the M06-2X/6-31G(d,p) level. This may also explain the single elementary step for acyl enzyme formation observed by Wei et al. from a QM/MM molecular dynamics simulation. Hybrid QM/MM calculations carried out by Harrison et al. showed almost perfect synchronicity between the Sγ attack and the proton transfer. Moreover, Wei et al. obtained a TS where the proton transfer, the Sγ...C(peptide carbonyl) bond formation and the peptide bond breaking of the substrates are approximately synchronous processes. The discrepancy between their results and the results presented here can be related to the different levels of theory, the different computational models and the different QM/MM partitioning schemes. As explained above, all these factors can essentially influence the position and even the existence of the transition state at the B3LYP level of theory Wei et al. used in the QM/MM computations. The second transition state structure, which connects the tetrahedral intermediate state to the product (i.e., an acyl-enzyme structure with an already cleaved amide bond), features a considerably elongated C...N distance (with a bond length of ~2 Å or even longer) for the scissile amide bond. Notable exceptions can be observed using the B97D method with all but the smallest (6-31G(d,p)) basis set. It should also be highlighted that a significantly smaller activation energy barrier was predicted for the zwitterionic tetrahedral intermediate – acyl-enzyme reaction step when using the B3LYP method than when using all the other methods applied in our ONIOM calculations. For the product (i.e., the acyl-enzyme plus the methyl–amine cleaved model ‘‘C-terminal part of the peptide’’) all the methods we applied predicted very similar results for the bond lengths. The notable exception is the slightly longer S–C distances predicted by the B97D method (independently from the basis sets we applied) compared to all the other levels of theory. In general, the B3LYP method predicted the largest values for the interatomic separations examined here, while the other DFT methods, especially the M06-2X method, predicted significantly smaller values. The reason for this is probably that, as mentioned above, the exchange–correlation functional applied in the B3LYP method is not an appropriate one for the long-range electron–electron exchange interaction, and it does not have an appropriate long range correlation term. Comparing the values to those reported by Wei et al. a remarkable similarity in bond lengths is observed. Nevertheless, Wei et al. obtained significantly shorter d4 distances (2.09 Å) which indicates a stronger H-bond between the N-terminal amide hydrogen and the histidine Nδ atom.

With regard to the activation energies, the general conclusion derived from our results is that for the acyl enzyme formation, the first elementary step (i.e., the formation of the zwitterionic tetrahedral intermediate) is the rate determining step. The calculated activation (ZPVE-corrected) energies are in the range of ~10-13 kcal mol^{-1}. All methods predicted the zwitterionic TI to be less stable than the Michaelis complex. From the tetrahedral intermediate state, the second transition

state (leading to the acyl–enzyme product) can be much more easily accessed than the first one leading back to the Michaelis complex. The calculated total energy of the system, consisting of the cleaved product and papain, was found to be higher than the papain Michaelis complex by comparing the (TS2-TI)-(TS2-AE) vs. (TS1-TI)-(TS1-MC) energy differences. This suggests that the reaction is shifted toward the reactant. It should be considered, however, that in the real reaction, the amine group (i.e., the N-terminal-end of the released peptide/protein sequence in real cases) can dissociate from papain and can be protonated, which should shift the reaction equilibrium in the product direction. Owing to the quite extended set of parameters used to parameterize the exchange and correlation functionals of the B97D, ωB97XD and M06-2X methods it is difficult to explain the origin of the differences observed between the calculated energies and geometries. However, the general observation is that for this particular reaction the B3LYP method predicts a remarkably longer non-bonding interatomic separation which is probably caused by the missing long-range terms in the exchange–correlation functional of the B3LYP method. Nevertheless, all the corresponding calculated distance values, especially those that represent ‘‘chemical bonds’’ have remarkable similarities.

Hydrolysis of acyl-enzyme intermadiate

The next process in the amide (peptide) hydrolysis is the acyl hydrolysis reaction. We modeled this reaction by adding a simple water molecule to the acyl–enzyme structure. The specific distances that can be used to follow the reaction are the S-C (d1), O(water)-C (d2), H(water)-O(water) (d3) and H(water)-N(imidazol) (d4) distances. In the acyl–enzyme water complex each method resulted in distance parameters that were very similar to each other. Only the slightly longer S-C and C-O distances calculated using the B3LYP and B97D methods are worth mentioning. Note that longer S-C distances were already observed at the product site of the acylation step. None of these parameters showed any significant basis set dependence in the applied 6-31G(d,p) to 6-311++G(d,p) range.

In the transition state, which connects the acyl-enzyme water complex to the acetic acid and papain complex, the results are essentially the same: all levels of theory we applied predicted very similar transition state geometries. For the calculated transition state structures, the d2 distances were predicted to be slightly longer when using the B3LYP and B97D methods. With regard to the breaking d3 and the newly forming d4 bond lengths, they are almost the same in the transition state. Only the ωB97X-D method predicted the breaking bond to be a little shorter than the newly forming

one. The PESs calculated at the B3LYP/6-31+G(d,p) and M06-2X/6-31+G(d,p) levels of theory showed rather synchronous reactions regarding the new d2 and d4 bond distances.

However, at the TS geometries the S-C (d1) distances were only marginally longer than those in the reactants. The S-C (d1) bond breaks spontaneously at shorter d2 and d4 values, and geometry optimizations starting from the pre- and post TS regions lead to the reactant (acyl-enzyme plus water) and product (enzyme plus acetic acid) geometries, respectively. Based on these data, an asynchronous but still a single elementary step reaction mechanism can be proposed for the deacylation process in which the O(water)-C(carbonyl) and H(water)-Nδ(histidine) bond formation and O(water)–H(water) bond breaking precede the S–C bond breaking. The product geometries were also very similar to each other with the exceptions of slightly longer S-C and remarkably longer O(carboxylate)-Hδ(histidine) formerly O-H (water) non-bonding distances predicted using the B3LYP method.

Interestingly, Wei et al. proposed a two distinct step mechanism for this particular step of papain hydrolysis. However, the activation (free) energy required for the second elementary reaction step was extremely small. Comparing the transition state structures for the first elementary step that Wei et al. obtained for the sole transition state that we calculated, a remarkable similarity can be recognized between them. In their paper, the transition state Ow-C(carbonyl), Ow-Hw and Nδ-Hw distances were 1.68 Å, 1.16 Å and 1.32, respectively, which are in good agreement with our d2, d3 and d4 values. The second elementary step that Wei et al. proposed corresponds mainly to S-C bond breaking, which is peculiar to the second phase in the single elementary reaction step in our calculations.

With regard to the transition state energies, all the density functional methods predicted values which were in the ~10-14 kcal mol^{-1} range. In general, the more complete the basis set, the higher the activation energy calculated. It should also be recognized that the energy level of the product site is always higher than that of the corresponding acyl–enzyme reactant. As emphasized in the case of acyl enzyme formation, neither the solvation of the product site nor the leaving of the acetic acid (model peptide fragment) from the papain site has been considered.

Comparing the two distinct chemical processes (acyl-enzyme formation and acyl-enzyme hydrolysis), the rate limiting steps have comparable activation energies. Nevertheless, the second (deacylation) reaction step can be featured with slightly higher activation energies in most of the methods we applied. However, it is generally accepted that the rate determining step for amide bond hydrolysis is acylation, in contrast to ester hydrolysis in which the deacylation step is the rate determining step. The few available theoretical calculations on the whole cysteine protease reaction also demonstrate comparable reaction barriers for these distinct steps, with a slightly larger barrier

for the acylation step. It should be remembered, however, that the most simple peptide model was used in our calculations and different reference points for the acyl-enzyme formation and the deacylation processes (the papain-NMA and the acyl-enzyme - (single molecule) water complexes, respectively) were applied.

Applying the Eyring-Polanyi equation for the k cat values (which are typically in the range of 1-50 s^{-1}, depending on the substrate, temperature and other conditions) at 310 K and taking the transmission coefficient to be unity, the activation Gibbs free energies could be calculated to be in the range of 15.8-18.1 kcal mol^{-1}. Assuming that neither the entropy nor the pΔV term of enthalpy has a crucial contribution, these values should be close to the activation energies. The available activation energies or Gibbs free energies derived from the reaction kinetics parameters are in the range of ~5 kcal mol^{-1} to ~18 kcal mol^{-1}. These values can also be derived from the k_{cat} rate constants using the classic Arrhenius equation and the equation given for conventional transition state theory, respectively. Therefore, in spite of the obvious simplifications in our model calculations, they still predict reliable activation energy values.

Activation of FXIII-A_2 and TG2

Results of molecular dynamics simulations

Simulations of FXIII-A_2

Due to the complex nature of the activation mechanism of FXIII-A_2, we followed a step-by-step approach to clarify the key events which may contribute to an enzymatically active protein conformation and can also be related to the binding of calcium. This meant starting from zymogen structures (*i/a,b*) through proteolitically activated ones (*ii/a-d*) and finally simulating the AP free proteins (*iii/a-c*). The root mean square deviations of zymogen and proteolitically activated FXIII-A_2 models suggest that the presence of calcium ions in general increases the heavy atom RMSDs in every case, compared to the reference simulation (*ii/d*). More importantly, on the timescale studied, the high level of Ca^{2+} concentration (50 mM) does not appear to influence the dynamics significantly, although it should be noted that these simulations converge much slower towards the ~2.5 Å final RMSD value, than any of the other ones. Therefore, we cannot exclude the importance of high calcium concentration on a much longer timescale since the increased fluctuations of activation peptides are clearly demonstrated (see later). The reason why dynamics of simulation sets *i* and *ii* are quite similar, lies in the presence of the APs. Independently from the fact whether the APs are cleaved or not, it seems that their main role is to keep the monomer subunits together. Since the fluctuations of the first 37 residues are significantly smaller in the absence of calcium ions, it is

highly probable that calciums play an essential role in releasing the APs. Two arginine residues (Arg11, Arg12) play a critical role in holding the APs in their place as it was found in earlier studies. These residues are contained in a relatively deep cavity and several acidic residues (Asp343, Asp345, Asp367 and Glu401) hold them in place via salt-bridge contacts. We assume that full dissociations of the APs are rare events on the MD timescale, if they are possible at all, but in the ii/c simulation we can observe a slight relocation of an arginine contained within one AP, however it does not dissociate completely. The added calcium ions increase the fluctuations of N-terminal peptides significantly, contributing directly to the release of APs, however other additional requirements cannot be excluded at this stage. The 2 μs long simulations without APs (*iii/a-c*) show that the RMSDs increase in every scenario even in the absence of calcium ions. The detailed effect of extremely high calcium concentration will be discussed in different section. In the context of gyration radius, all of the studied zymogen and cleaved models represent almost identical final values, with the average radius of gyration of these simulations being 36.90 ± 0.1 Å. The RMSD and r_{gyr} were also calculated for each monomer chain and can be found in the SI of correspoding paper. Notable regions with high local RMSF could be identified and generally speaking, these regions are loop segments or located in close proximity to the terminals of the AP. Nevertheless we would like to highlight the possible importance of four regions which may have distinguished significance. These regions can be found between the residues of Val274-Asp280, Met350-Trp370, Ile440-Ile460 and Pro505-Arg515. Other than the first segment, all of the others are situated in close proximity to each other. The loop regions between Val274-Asp280 shows moderate fluctuations except for the third simulation set, where the RMSF values are significantly higher. The second region is also in contact with sixteen residues of the other subunit N-terminal and a helix can be found at the N-terminal end of the last region. This helix contain the residues of the main Ca^{2+}-binding site, hence an allosteric pathway can be drawn from the main binding site to the AP. Since the residues between Pro505-Arg515 are usually poorly or non-resolved at all in crystallographic structures, even in the 1F13 structure the subunits show different conformations in these regions, one can obviously associate a regulating function to these segments. The Met350-Trp370, Ile440-Ile460 regions are organized in three adjacent anti-parallel β-strands and one disordered loop. Despite contradictions in the details of the proposed activation mechanisms the above mentioned segments should be equally important in the movement of β-barrel 1. Simulation *iii/c* corroborate this theory, since the local fluctuations are higher than in simulation *i* and *ii* and also a special monomer movement can be observed within this system as well.

Simulations of TG2

Following the same step-by-step approach in the case of TG2, not only the individual effects of Ca^{2+} ions and guanosine-phosphates, but their simultaneous effects have also been studied. Here we cannot observe notable differences in RMSD and r_{gyr} values between *iv/a* and *iv/b* simulations (both without GDP), however in the presence of GDP we can see clear differences in RMSDs which follow the *iv/a ~ iv/b > iv/d > iv/c* tendency). These results can be elucidated by the GDP-bound forms being *in vivo* inactives and the smaller RMSDs can also be the sign of the conservative behaviour of the inactive form. It should be noted here that the added calcium ions obviously increase the RMSD indicating their significance in creating an, eventually, active (open) TGase conformation. The r_{gyr} values in this case are much higher suggesting that the parallel occurrence of the Ca^{2+} ions and the GDP leads to different dynamics. The simulation using the 4PYG structure as a starting point (*iv/e*) also supports the hypothesis that has been stated in the case of *iv/d*. It is notable that after 650 ns the RMSD values of *iv/e* show an elevation, probably since the initial structure contains three point mutation (Gln186Glu, Thr533Asn, Val655Leu) and these mutations influence the dynamics in condensed phase. The r_{gyr} values do not reflect serious changes, i.e. the protein remains stable and there are no signs of unfolding). Due to the highly conserved secondary and tertiary structure of the inactive (closed) FXIII-A_2 and TG2 we can conclude similarities in RMSF values. The earlier highlighted regions of FXIII-A_2 correspond to the following residues in the TG2 structure: Ile313-Asn333, Val401-Val422 and Asn460-Thr471 (according to the TG2 residue numbering). Indeed, these regions seem to share many common features with the FXIII-A, thus we focus on the fine details between the TG simulations by looking at the RMSF differences between the corresponding sets.

As it can be seen, the largest differences affect the β-sandwich domain, however these discrepancies can be attributed to regions with no known importance, such as Leu12-Lys30, Gly64-Gly72, Leu79-Asp87. The region between Arg240-Tyr245 shows high fluctuation in the presence of Ca^{2+}), although importance of this segment is not fully understood. More interestingly the active site Cys277 is concealed by this loop and a binding site can also be found near Asp232 and Asp233 (Asp270 and Asp271 in FXIII-A_2). Since this loop most likely adapts an ordered conformation in the active TG2 and FXIII-A_2 as well, it is possible that it can control the opening of the active site during activation not only in TG2 but in FXIII-A_2 too. The loop between Thr343-Glu352 also shows high fluctuation but only in the absence of calcium and GDP. In the presence of calcium ions this loop is fixed by a bound calcium, which seems like an acceptable interpretation of the observed high fluctuation (*iv/a*).

Concerning the loop region between Asn460-Thr471, we can provide the following plausible explanation. Despite the residues of the main binding site in FXIII-A_2 (Asp438, Gly457, Glu485, Glu490) being almost perfectly conserved in TG2 (Asp400, Ser419, Glu447, Glu452), we can only assume, based on the above mentioned identities, that this site would actually bind any calcium. Since the β-strand in FXIII-A (Thr449-Ile460) is partially disordered in TG2 (Ser411-Val422), this assumption seems highly plausible indeed. On the other hand, the space available in TG2 is much wider than in FXIII-A_2. Another possibility could be the formation of a binding site under the helix which pulls it downward, hence strengthening the Asn460-Thr471 loop, just like bending a note on a guitar string. The fluctuations only increased when there were calciums added to the simulation box, meaning that the C-terminal end of this loop is connected to the β-barrel 1, hence on a much longer timescale it may initiate more significant rearrangements, such as pulling down the first β-barrel and turning both barrels upside down. This mechanism would be completely in line with the "swiss army knife" theory for the activation of TGs.

Results of calcium binding

FXIII-A_2

It was known that FXIII-A_2 can bind one calcium per A subunit and several structures can be found in the Protein Data Bank that contain di- or trivalent cations bound to FXIII-A_2, eg. 1GGU, 1GGY, 1QRK, 1EVU. Based on these studies, factor XIII-A_2 can bind only one Ca^{2+} or Sr^{2+} per subunit in the very same binding site (Asp438, Gly457, Glu485, Glu490) and can accommodate up to eight Yb^{3+} in three binding sites (4 Yb^{3+} - chain A/B Asp270, Asp271, Glu272; 2 Yb^{3+} - chain B: Asp438, Glu485, Glu490; 1 Yb^{3+} - chain A: Asp438, Glu485, Glu490; 1 Yb^{3+} - chain A: Asp574, Glu585). It is worth noting that the 4KTY X-ray structure of the open FXIII-A^O also contains three calcium ions. One of them is located in the main binding site (Asp438, Gly457, Glu485, Glu490), the second bound by Asp270 and Asp271, and the third bound calcium was accommodated by the side chains of Asp343, Asp345, Asp351 and Asp367. The last one is buried in the closed (inactive) FXIII-A_2 by the APs, which could explain why it was unoccupied in the case of ytterbium(III) ions. On the other hand, these three occupied sites seem to be well conserved features, since they are also present in the TG3 structure (PDB ID: 1NUD). As the radius of the previously applied di- or trivalent cations are very close to each other, one can assume that beside the main site, the interfacial site of the homodimer (Asp270, Asp271 of both subunits) can also bind Ca^{2+} not only Yb^{3+}, and other sites can also exist.

Indeed, based on our simple MD simulations several possible binding sites have been identified, with most of them being formed by either spatially close or consecutive negatively charged side-chains. We have applied a selection criteria in order to reduce the number of possible binding sites and to get rid of temporary or weak binding modes. This meant, that only those residues with a carboxylate carbon were collected which had a calcium ion within 4.5 Å and the ion can be found within that threshold in more than 5 % of the total simulation time. Here we discuss mostly the results of ii/a (FXIII-A_2', 1GGU + 14 mM Ca^{2+}) and ii/c (FXIII-A_2', 1F13 + 14 mM Ca^{2+}) simulations and preferably underline those sites which are most likely to be conserved in FXIII-A, TG2, TG3 enzymes.

Both ii/a and ii/c systems are equally important for understanding calcium binding. It is notable, despite the initial structure of *ii/c* not containing any bound calciums, that after a few tenths ns of simulation time one of its main sites becomes occupied and the Ca^{2+} ion preserves its position throughout the whole simulation, while the other monomer does not bind any calciums at all in that place.

In a preliminary simulation, what was conducted under the very same conditions as in *ii/c*, both main sites were filled. In this particular case, the major site of chain A has been filled after 500 ns and after 900 ns both sites contained bound calciums. The protonation states of titratable residues were not predicted in this case, which ultimately lead to the binding of multiple calciums at the interfacial site (Asp270, Asp271 and Glu272 of both subunits).

The main sites of *ii/a* appear to bind the calciums quite tightly, even though the Pro505-Arg515 loops were remodelled based on how they were found in the 1F13 structure. Thus, we can assume that the loop conformation does not influence the calcium binding at all and there is no clear connection between the secondary structures of these loops and the calcium binding events. Comparing the geometries of main sites in the *ii/a* and *ii/c* simulations, we can see an almost identical arrangement in their cluster representations. Throughout the simulations the remodelled loop regions preserved their initial structure excluding some minor changes, as one of them became almost completely disordered while the other one still contained a one turn helix element. Based on this information it is unclear whether the loop regions have any other role over the *in vivo* activation or they are just in a fine equilibrium between the disordered and the one-turn helix arrangements.

In the *i/b* and *ii/b* simulations those major sites were occupied, which belong to chain B, while in the simulation system of the AP free protein model (*iii/b*) the main site of chain A was found to bind a calcium ion. The extreme high calcium load (*iii/c*) caused the binding of at least one calcium ion per binding site in both A subunits. Thus, a straight conclusion cannot be drawn

concerning these, seemingly, random preferences based on these relatively long equilibrium MD simulations.

In order to study this “blind spot” in great detail, 17 simulations were performed with different initial seeds, starting from the last frame of *ii/c* by keeping the bound calciums and resolvating the system in the presence of 100 mM Ca^{2+}, each one lasting 25 ns. In only one out of the 17 simulations the second site has become filled after 11 ns of simulation time. We assume that this latent preference comes from the stochastic nature of the simulations and the random ion placement.

The second binding site which became occupied quite frequently in our simulations, formed by Asp270 and Asp271 of the monomers and mentioned earlier as an interfacial site, is known to be able to accommodate Yb^{3+} ions while these aspartates also bind the Ca^{2+} ions in the 4KTY (non-proteolitically activated) structure of factor XIII-A and in the transglutaminase 3 (TG3) (PDB ID: 1NUD). Our results suggest that in this site at least one bound calcium ion can always be found. The prediction of protonation states suggests that the sidechain of Glu272 is possibly protonated and, depending on its protonation state, is able to bind up to two calcium ions. Despite our significant efforts, the role of this site is still not fully understood, although these loop regions are quite flexible and therefore can easily adapt a conformation that can be seen in the 4KTY or 1NUD structures. It is important to note that this region is situated very close to the central salt-bridge network, so we cannot exclude its regulatory role over the whole activation process. Also, these sites are directly connected with the loops between Ser278-Gly294 of the A subunits, which serve as the boundary between the active site and the partially neighbouring β-barrel-2.

The third calcium binding site, which is also present in 4KTY and in 1NUD, can be found under the activation peptide and is formed by the side-chains of Asp343, Asp345, Asp351 and Asp367. This site was not easily accessible while the AP was in its place and the Asp343, Asp367 residues were in direct contact with the guanidium group of the side-chains of Arg11 and Arg12, therefore binding was not possible at this site. However, in the *iii/b* simulation a calcium can be found near the Asp343 and Asp345 of the A chain, with the dimer structure remaining intact and the other two residues being further away, so additional changes would be required for binding to occur at this site. As mentioned above, the extreme high calcium concentration (*iii/c*) induces notable changes in the overall position of the A subunits compared to each other. It seems that in both main chains the site is formed after 250 ns of simulation and additionally almost all of the possible anionic surface sites became occupied). Sites with reduced affinity (Glu355, Glu356, Asp357; Asp521, Glu523, Glu525; Asp574, Glu585, Glu614; Asp472, Glu631, Glu720, Asp722) may also be important in the non-proteolytic activation. We suggest that the reason why extreme high ionic

strength is able to cause large scale monomer rotation is probably due to the breakdown of the electrostatic interactions between the A subunits. On the other hand, it is remain unclear what causes the separation of the β-barrels from the catalytic core domain. In fact, the site formed by Asp472, Glu631, Glu720 and Asp722 represents an interdomain contact between the β-barrel 2 and the core domain.

An important question could be whether the simplified molecular mechanics parameters of divalent cations are able to predict the geometries of binding sites appropriately? To try and answer this question, we have decided to continue the *ii/c* simulation for a further 50 ns NpT MD by replacing the bound calciums to multi-site calcium models in order to validate the geometries of the already established binding sites. Our results suggests that all sites preserve their potency perfectly.

It is worth mentioning that at least five binding sites were identified in the case of *v/a* (4KTY, 14 mM Ca^{2+}) and each one binds its calcium ion continuously, while in the zymogen or cleaved systems some of these sites seem to have somewhat weaker potency (eg. Asp138, Glu139 or Glu355, Glu356, Asp357).

TG2

Based on site-directed mutagenesis studies, five out of six binding sites of TG2 were identified, partly by structural homology and partly by studying the electrostatic charge distribution on the protein surface. Despite the significant efforts of this work, the static snapshot used as starting point, provided by X-ray crystallography, was lacking the true dynamic behaviour of proteins *in vivo,* nevertheless, MD simulations provide a good insight to study these time-resolved atomic-level events. Since such a detailed mutagenesis studies are not available in the case of FXIII-A_2, it is important to compare the results of our *in silico* work to *in vitro* experiments and TG2 embodies a perfect target for such a purpose. Instead of plotting binding events alone, we have concentrated on the representation of distances between the carboxylate carbon atom of acidic residues and the calcium ions. In the following we use the site labelling suggested by Király et al.

Both *iv/b* and *iv/d* simulations share many common features, however we can also identify some minor differences between them, which can be related to the longer simulation time of *ii/b* and the absence of GDP. Although it is worth noting that both the 1 and 2 μs long simulations were sufficient to identify almost all sites. Compared to *in vitro* results and static modelling, an important difference is that at least two sites exist on the β-sandwich domain. In the case of the GTP bound form of TG2 (*iv/e*) these sites appear much less significant, however we cannot completely diminish the importance of them. During the simulation of the active (open) TG2 (*v/b*) these sites were also

identified, but probably do not play an important role in the whole regulatory process. In agreement with previous experiments, the S1 site was also found in all of the simulations except in *iv/b* (GDP-free). It should be noted that the S1 site is formed via the contribution of the Asp232 and Asp233 residues. These residues correspond to Asp270 and Asp271 in the FXIII-A sequence and was identified as a binding site in the case of TG3 as well. Concerning the acidic residues of the S2A site, we can see that the sidechain of Glu396 orientates to the inner side of the core domain and therefore is buried within. Besides Glu396, there is another negatively charged residue located at this site, Asp400. In the case of *iv/a-d* the distance between the Cγ of Asp400 and the Nζ of Lys464 is 5.55 ± 1.25 Å on average, which does not deviate considerably during the production runs. Comparing this value to those which can be calculated in the case of *ii/a-d* (FXIII-A_2'), the corresponding residues are significantly further apart (6.25 ± 1.30 Å), thus the shorter distance can be attributed to the salt-bridge blocking of the Asp400 sidechain. Instead of the S2A and S2B sites a new site has been identified with a high rate of occupancy. By the residues of Glu381, Glu451 and Glu454 a new site can be formed and the occupancy of this site is almost continuous in every case. A second new site has also been formed by the residues of Glu319, Asp408 and Asp409. It is worth noting that in the open conformation the Asp581 takes part in the formation of this site and also the relocation of β-barrel 1 and β-barrel 2 domains is required, since in the closed conformation the Asp581 is in ca. 30 Å distance from the other three residues. The Glu467, Glu469, Glu470 residues of the 460s loop possess weak binding affinity and the simulation with multi-site calcium models also confirm that the Ca^{2+} remains close to Glu467 and Glu470. Further two binding sites, namely the S3A and S3B were proposed with one-one acidic residue and indeed, both two amino acids bind the same calcium ion and, in the minority of the simulation time, the Glu363 also contributes to the binding. The S4 site is composed from the consecutive five acidic residues between Asp151 and Glu158. This site is located very close to S5 and out simulations are in good agreement with experimental results, meaning that Asp151 and Glu155 seem to be an appropriate diad for binding calcium ions. On top of the already known sites a new one has also been identified, formed by Asp640, Glu643 and Glu646 in every case. Similar to the FXIII-A_2 simulation, *iv/d* was continued for a further 50 ns using multi-site calcium models and the overwhelming majority of the ion binding contacts remained unchanged.

Based on our results, the following general conclusions can be made on the calcium binding sites of TG2 compared to the experimental results by Király et al. The S1, S3A/S3B and S4 sites are found to be suitable for binding calciums in our simulations, however opposing results have been found in regards to the S2A/S2B and S5 sites. As we can see in the case of FXIII-A_2, at least one of its main binding sites was occupied, according to our criteria mentioned before, for 1 or 2 μs of

simulation time. Nevertheless, the very same site of TG2 (S2A/S2B) does not bind calcium at all, despite that the pocket seems to open up and provides a much wider cavity to than in FXIII-A_2. We cannot rule out that in a later phase of the activation process the significance of this site will increase, but starting from crystallographic structures even during these relatively long MD simulations we were unable to observe binding events at those sites. In respect to the S5 site our results suggest that this site has no significance in this early stage of the activation either and does not show any binding in the case of the open (active) TG2 (*v/b*) which can make its overall relevance questionable.

Analysis of correlated motions

General conclusions can be made on the correlated motions within the FXIII-A_2 systems. By taking the differences between the zymogen (*i*) and the cleaved (*ii*) simulation sets, the main common feature that clearly appears is the strong correlation of the β-sandwich and β-barrel 1 domains within the subunits. This positive correlation indicates the movement of the domains in the same direction, thus these motions are practically the same as represented by the principal component analysis in the next section. Taking the comparison of *i/a* and *ii/a* and also the *i/b* and *ii/b* simulations several important features can be observed. As mentioned above, the region between the residues of Met350-Trp370 and Ile440-Ile460 are assumed to play an important role regardless of the activation mechanism, meaning that these regions should be involved in both suggested. Supporting this hypothesis, these segments show increased correlation with the β-barrel 1 and also with the very same segments of the other subunit in the cleaved structures, while in the zymogen structures they seem much less correlated. Another interesting observation is that these regions also correlate significantly with the residues between Gly210-Trp225. This region is located next to the C-terminal end of the β-barrel 1 and the N-terminal of β-barrel 2 and also shows somewhat weaker correlation to these in the zymogen models. The effects of bound calcium ions become evident when comparing *ii/c* and *ii/d*, where their correlation to the β-sandwich domains can mainly be attributed to the presence of these bound ions. Last but not least, all of the notable correlations are present in the third simulation set as well and the extreme ionic strength causes even higher correlation between the β-sandwich domains and strong anti-correlation between the β-sandwich and the β-barrel domains. These findings can predict the observed monomer rotation and are in good agreement with the results of the PCAs.

Despite the well-preserved tertiary structures of FXIII-A_2 and TG2, it is important to note that in the latter the β-barrel 1 is correlated with the β-sandwich and the β-barrel 2 seems to

correlate rather with the core domain, making it a unique feature of TG2. The second notable difference can be found in the β-strand regions between the residues of Ile313-Asn333 and Val401-Val422 (correspond to the Met350-Trp370 and Ile440-Ile460 in FXIII-A_2), where these segments correlate with β-barrel 1 and also with the core domain, while in the case of FXIII-A_2 we cannot observe this latter correlation.

Concerning the simulation of open transglutaminases (*v/a,b*) the cross-correlation matrices suggest a clear difference. Namely, in v/b (TG2, 2Q3Z) the β-barrel 2 is correlated with the β-sandwich domain and is strongly anti-correlated with the core and β-barrel 1 domains. In the *v/a* (FXIII-A^o, 4KTY) we can observe exactly the opposite phenomena. To our current knowledge, we emphasize that either these enzymes follow two completely different activation mechanisms or one of the open conformations was kinetically trapped in a transient state during the large scale motions.

Principal component analysis and the role of central salt-bridge network

The large amplitude motions are usually difficult to visually detect even over longer MD simulations, therefore we have decided to examine these possible motions in another way. Most of these dihedrals are distributed closely around -40° with one exception. In the case of *iii/c* this value is approximately -70°, what means a further 30 degrees of rotation. Interestingly, the principal component analyses of *iii/a-c* systems shows that the motion along the first eigenvector is in perfect agreement with our findings in the case of iii/c and with the domain correlations discussed before. The high Ca^{2+} concentration induced monomer rotation from the MD trajectory also can be observed in the PCA analysis as the first eigenvector in each simulation. It is important to point out that in the presence of the AP, these eigenvectors are hardly present at all or are significantly smaller, which provides further evidence on the protective roles of APs. Moreover, these findings support the experimental results which suggest that a slow progressive activation occurs and also that added calciums increase the speed of activation for the zymogen FXIII-A_2. Since these motions are not visible in the presence of APs, it can be assumed that the relocation of the N-terminal part of the AP is necessary for the activation, which is in agreement with the enhanced fluctuations of these regions. The rotation takes place as a contrary motion of the subunits involved, where the axis perpendicular to the plane of the paper goes through an interdomain salt-bridge network. This salt-bridge network has been never discussed previously. The Asp404, Asp427 and the Arg260, Arg408 residues of both subunits play a role in the formation of this salt-bridge network and despite the significant movement these contacts remain stable, excluding minor differences in the case of *iii/c*.

Moreover, it was found earlier that the mutation of Arg260 causes serious deficiencies, therefore the effect of this mutation was also investigated by computational means and then compared to the energies of optimized structures. Our results suggest that after 1 μs simulation time the subunits do not suffer significant changes, despite the obvious fact that Cys260 was unable to establish any native contacts. It is also possible that this mutations causes structural changes within the A subunit, hence after folding the native dimer could not be formed. Last but not least, an interesting difference can be observed between the sequences of FXIII-A and TG2. It has been identified that Arg408 is not conserved in TG2, but Cys370 can be found in this exact position, which can lead to the conclusion that the Arg260Cys can cause serious malfunction and the subunits are unable to assemble. Based on this analogy, the Cys370 could then be the reason why TG2 preserves its monomer structure *in vivo*. Structurally, Cys370 and Cys371 residues are important and are known as a disulphide switch which helps to stabilize the active (open) conformation of TG2.

DISCUSSION

The full enzyme mechanism of papain

The proteolytic reaction of papain has been modeled using ONIOM type QM/MM methods using a simple peptide model substrate, N-methylacetamide. The applicability of a hybrid GGA (the popular B3LYP) method to a few of the more recent DFT methods, with a long range correction term or suitable parameters for such an interaction, as the QM component in ONIOM QM/MM calculations, has been evaluated.

Our calculations show that in the resting state of papain the ion pair and neutral forms of the His-Cys side chains of the catalytic dyad have approximately the same energy levels that are separated by a small barrier. Zero point correction shifts this equilibrium slightly in the direction of the neutral form, while the implicit solvent model correcting the ONIOM method with PB computations predicts the ion pair form to be the most populated one, in good agreement with the available experimental data.

With regard to the enzyme mechanism, all the dispersion corrected DFT methods applied as well as the B3LYP method using the larger (6-31+G(d,p) to 6-311++G(d,p)) basis sets, predict two elementary steps for the acylation phase and a single elementary deacylation step. According to these calculations, in the acylation phase a zwitterionic tetrahedral intermediate exists where the carbonyl carbon becomes tetrahedral and holds a negatively charged oxygen and, simultaneously, the amide nitrogen is protonated as well. The activation energies we derived are in the range that

can be found in the literature or can be derived from the available kinetic constants. Although all the density functional methods applied in this work predict an asynchronous rate determining first step for the acylation reaction, it should be emphasized that ONIOM type QM/MM computations using long range corrected DFT methods resulted in a significantly different transition state compared to those that can be obtained using the B3LYP method using larger (6-31+G(d,p) to 6-311++G(d,p)) basis sets. Nonetheless, the proton transfer lags behind (or at least does not precede) the S-C bond formation. Interestingly, the B3LYP/6-31G(d,p) method in ONIOM predicts three elementary steps for the first (acylation) step through an anionic and a subsequent zwitterionic intermediate. It should also be emphasized that the activation energy between these two intermediates was found to be extremely small. Since the B3LYP functional in ONIOM predicted a very small barrier for the amid bond breaking (independent of the basis sets), it increases the possibility that, depending on the chemical environment and the method one applies for the transition search, different, one to three elementary step mechanisms are possible for acyl enzyme formation.

An additional conclusion, which can also be drawn from our present work, is that for the cysteine protease reaction, the use of dispersion corrected DFT methods is strongly advised both in pure QM and QM/MM calculations, because it might qualitatively influence the computationally derived reaction mechanism, in contrast to the serine protease mechanism.

Activation of human transglutaminases

A series of microsecond long all atom MD simulation were carried out in order to clarify the important details which may initiate the activation of the studied TGs and are primarily linked to the binding of calcium ions to blood coagulation factor XIII-A_2 and TG2.

In the case of the dimer A subunit of blood coagulation factor XIII we have found that the calcium concentration increases the flexibility of the N-terminals of the APs and three calcium binding sites were identified with possible significance in the activation mechanism. The main binding site of FXIII-A_2 was known and based on our work it was found that binding events can be observed in at least one of the main sites on the studied 1 or 2 μs long timescale. In those models which were based on 1GGU (Ca^{2+} bound crystallographic structure), the calcium ions preserved their position throughout all of the simulations, thus these sites can be considered strong binding sites. Suggestions have been made for the possible significance of the binding site near Asp270 and Asp271 of both subunits during the activation procedure. A third binding site was identified, which has possible importance only in the absence of APs and the rotation of A subunits are needed to form this site (Asp343, Asp345, Asp351 and Asp367). It should be noted that this site was known in

the case of TG3 already, and simulation with the open FXIII-A also confirmed the existence of this site.

It was also pointed out that in the absence of APs the high Ca^{2+} concentration (1000 mM) was able to cause large scale monomer movements, probably by breaking down the intermolecular electrostatic interactions. The calculated PCA eigenvectors predict this rotation even in the absence of any calcium ions and this finding is supported directly by *in vivo* experiments in the case of zymogen structures, if we assume that the high ionic strength causes immediate dissociation of APs.

The calcium binding ability of TG2, a structurally very similar protein although with other distinct functions, was also investigated using *in silico* modelling methods, based on the profound work of Király and co-workers. Our current work provides further evidence on most of these proposed sites, however in a few cases results from our simulation studies seem to point to different conclusions. According to the labelling of the binding sites of Király et al., the S1, S3A/S3B and S4 sites are in good agreement with the results of previously conducted *in vitro* studies, however instead of S2A/S2B we would like to propose another site which can be formed by the contribution of Glu381, Glu451 and Glu454 in this early phase of activation. Based on our work, we could not confirm whether the S5 site really appropriate to binding calcium ion or not. Among other possible and earlier unknown binding sites, another site was also identified which can be formed by the Asp640, Glu643 and Glu646 residues and found to be frequently populated by calcium ions.

NEW RESULTS OF THE PHD THESIS

Hereby I declare that all of the published results have been performed by myself.

• During the study of enzyme mechanism of papain it was showed that in its resting state the transition of proton between the neutral and the ion pair diads was spontaneous.
• With the zero-point energy correction the transtition of proton was also found to be spontaneous.
• The solvation free energy correction predicted the ion pair form to be the most stable one.
• The formation of acyl-enzyme intermediate is a two step process thus it was proved that the zwitterionic intermediate exists and it is formed through a concertic transition state in which the nucleophilic attack of thiolate sulphur atom and the proton transtition happen in an asynchronous step.
• Th hydrolysis of acyl-enzyme intermediate was going through one transtition state in the presence of one explicite water molecule. This reaction produce the acetic acid and the active form of the enzyme.

• The presence of calcium ions effected inevitably the condensed phase dynamics of studied translgutaminases.
• We shown that there is no preference between the main calcium binding sites of FXIII-A_2, buti t is possible to draw an allosteric regulation pathway between the main binding site and the Arg11 and Arg12 of the activation peptide.
• A new calcium binding site was found which formed by the very same aspartate residues (Asp270, Asp271) of both A subunits. However it can be found in the crystallographic structure of active FXIII-A but int he case of closed structures it was unknown. Moreover several other weaker sites have been suggested.
• The extremely high calcium (1 M Ca^{2+}) concentration caused the rotation of A subunits which was proved by principal component analysis and correlation matrices.
• Three previously proposed binding sites (S1, S3, S4) of the tissue transglutaminase were found to be potentially active sites.
• We were unable to detect binding events in the S2A/S2B (proposed by analogy with FXIII-A) and the S5 sites ont he studied timescales.
• Three new binding sites have been identified in the TG2 and probably one of these sites is the missing sixth binding site (S6) of the protein.

SUMMARY

The enzyme mechanism of cysteine proteases was invastigated several times with both theoretical and experimental methods on the papain. However by this time the full catalytic cycle was not studied in one place with the means of verified stationary points. In present work we wish to use ONIOM-EE type hybrid calculations for the calculation of the deep details of mechanism and for the QM layer, recently popular density functional methods were used. We also published a comparison concerning the performance of these density functional theories.

Based on our results it was found, that the transition between the neutral and the thiolate-imidazolium ion pair form of the catalytic diad is spontaneous in gas phase and the PB solvation correction the equilibrium is pushed toward the ion pair form. We found the first acylation step of the mechanism to a concertic reaction in which the attack of thiolate nucleophile and the transition of hydrogen from the imidazolium ring to the substrates amide nitrogen is nearly parallel but not fully synchronous process. The cleavage of peptide bond possesses only a small energy barrier and its importance is only theoretical. The activation barrier regarding the hydrolysis of acyl-enzyme intermediate is comparable to the energy of the acylation step. This final step was found to be also concertic process which yield the corresponding carboxylic acid product and the regenerated active centre.

The activation of human transglutaminases with papain-like catalytic core domain is clearly in connection with the presence of calcium ions, we presented the study of two human translgutaminases and the effects of the presence of calcium ions were deeply investigated. These two enzymes were the coagulation FXIII-A_2 and the tissue transglutaminase. Both ezymes share highly conserved secondary and tertiary structures despite the relatively low sequence identities, the latter fact predicts the diverse biological functions of these transglutaminases.

During our extensive in silico molecular dynamics simulations we attempted to study these proteins on the microsecond timescale, in particular the calcium binding properties of these transgluatminases were carefully examined. Based on our results we found that in the case of FXIII-A_2 at least two different binding sites exist beside the known main site of A subunits. The importance of the binding site formed by the Asp270 and Asp271 residues of both main chains is highlighted. At the study of tissue transglutaminase we wanted to compare out results to the existing in vitro experiments which based on site-directed mutagenesis. Beside several suggestions the most imporant finding was a strong, sixth calcium binding site of TG2 which location remained hidden by this time.

LIST OF PUBLICATIONS

UNIVERSITY AND NATIONAL LIBRAY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

Registry number: DEENK/51/2019.PL
Subject: PhD Publikációs Lista

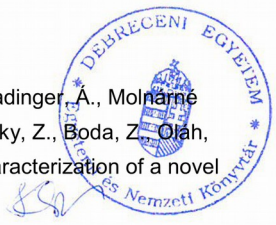
Candidate: Attila Fekete
Neptun ID: L0TV9S
Doctoral School: Kálmán Laki Doctoral School
MTMT ID: 10062749

List of publications related to the dissertation

1. **Fekete, A.**, Komáromi, I., Mucs, D.: On the early events of the calcium induced activation of coagulation factor XIII-A2 and tissue transglutaminase: an in silico study.
   J. Biomol. Struct. Dyn. [Epub ahead of print], 1-16, 2019.
   IF: 3.107 (2017)
2. **Fekete, A.**, Komáromi, I.: Modeling the archetype cysteine protease reaction using dispersion corrected density functional methods in ONIOM-type hybrid QM/MM calculations: the proteolytic reaction of papain.
   Phys. Chem. Chem. Phys. 18 (48), 32847-32861, 2016.
   DOI: http://dx.doi.org/10.1039/C6CP06869C
   IF: 4.123

List of other publications

3. Fizil, Á., Sonderegger, C., Czajlik, A., **Fekete, A.**, Komáromi, I., Hajdu, D., Marx, F., Batta, G.: Calcium binding of the antifungal protein PAF: Structure, dynamics and function aspects by NMR and MD simulations.
   PLoS One. 13 (10), 1-19, 2018.
   DOI: http://dx.doi.org/10.1371/journal.pone.0204825
   IF: 2.766 (2017)
4. Selmeczi, A., Gindele, R., Ilonczai, P., **Fekete, A.**, Komáromi, I., Schlammadinger, Á., Molnárné Rázsó, K., Kovács, K. B., Bárdos, H., Ádány, R., Muszbek, L., Bereczky, Z., Boda, Z., Oláh, Z.: Antithrombin Debrecen (p.Leu205Pro) - Clinical and molecular characterization of a novel mutation associated with severe thrombotic tendency.
   Thromb. Res. 158, 1-7, 2017.
   DOI: http://dx.doi.org/10.1016/j.thromres.2017.07.023
   IF: 2.779



UNIVERSITY AND NATIONAL LIBRAY
UNIVERSITY OF DEBRECEN
H-4002 Egyetem tér 1, Debrecen
Phone: +3652/410-443, email: publikaciok@lib.unideb.hu

5. Tóth, L., **Fekete, A.**, Balogh, G., Bereczky, Z., Komáromi, I.: Dynamic properties of the native free antithrombin from molecular dynamics simulations: computational evidence for solvent-exposed Arg393 side chain.
J. Biomol. Struct. Dyn. 33 (9), 2023-2036, 2015.
DOI: http://dx.doi.org/10.1080/07391102.2014.986525
IF: 2.3

6. Kovács, K. B., Pataki, I., Bárdos, H., **Fekete, A.**, Pfliegler, G., Haramura, G., Gindele, R., Komáromi, I., Balla, G., Ádány, R., Muszbek, L., Bereczky, Z.: Molecular characterization of p.Asp77Gly and the novel p.Ala163Val and p.Ala163Glu mutations causing protein C deficiency.
Thromb. Res. 135 (4), 718-726, 2015.
DOI: http://dx.doi.org/10.1016/j.thromres.2015.01.011
IF: 2.32

7. Bokor, É., **Fekete, A.**, Varga, G., Szőcs, B., Czifrák, K., Komáromi, I., Somsák, L.: C-(β-d-Glucopyranosyl)formamidrazones, formic acid hydrazides and their transformations into 3-(β-d-glucopyranosyl)-5-substituted-1,2,4-triazoles: a synthetic and computational study.
Tetrahedron. 69 (48), 10391-10404, 2013.
DOI: http://dx.doi.org/10.1016/j.tet.2013.09.099
IF: 2.817

Total IF of journals (all publications): 20,212
Total IF of journals (publications related to the dissertation): 7,23

The Candidate's publication data submitted to the iDEa Tudóstér have been validated by DEENK on the basis of the Journal Citation Report (Impact Factor) database.

05 March, 2019