UNIVERSITÀ DEGLI STUDI DI FOGGIA

DOTTORATO DI RICERCA IN
"HEALTH FOOD INNOVATION AND MANAGEMENT- I.M.A.E.V." (XXXI CICLO)

COORDINATORE: PROF. MATTEO A. DEL NOBILE

TRANSGLUTAMINASE, NUTRITION AND HUMAN HEALTH

Dottorando:	GIORDANO DEBORAH
Tutor:	SABATO D'AURIA
Co-tutor:	ANGELO FACCHIANO

ANNO ACCADEMICO 2017/2018

Abstract

Background: transglutaminases (TGase) are a class of enzymes widely spread in eukaryotic and prokaryotic organisms. Enzymes of this family catalyze post-translational modifications in many proteins by acyl transfer reactions, deamidation and crosslinking (polymerisation) between protein intra- or inter-chain glutamine (acyl donor) and lysine (acyl acceptor) peptide residues. Due to its facility of expression and purification, the only TGase enzyme widely used for industrial applications is the microbial TGase extracted from *Streptomyces mobaraensis* (MTGase). Nowadays the MTGase is commercially available and widely used in biopolymers industry, in cosmetics, in clinical applications, in wool textiles, and above all in the food processing industry. Its ability to catalyze crosslinks on many different protein substrates is increasingly used not only for sausage, ham and cheese production but, very recently, also for flour detoxification, as a possible alternative therapy to the gluten free diet.

It follows that nowadays the industrial applications of MTGase have increased, covering more and more fields producing a very active scientific research about this topic aimed at attempt to meet specific industrial needs, as the implementation of more efficient system for MTGase production, the research of alternative sources of microbial TGase, and safe source of recombinant enzymes.

Aims of the doctorate project: the main aim of the project is the identification of novel forms of microbial TGases that could become an alternative to that in use. A depth screening of known sequences has been performed, with the aim of obtaining a classification of microbial TGases for their similarity to known forms. To select the best candidates to be active forms under appropriate conditions, molecular modelling and molecular simulations have been performed on selected sequences. To test the enzymatic activity, experimental assays have been performed with a novel form, and another novel form has been expressed.

Results: the present work proposes at first an analysis, lacking so far, of the wide microbial transglutaminase world, developing the first classification of the microbial TGase based on their sequence features and their specific predicted secondary structures.

In order to classify and analyze the structural features of all the sequences annotated as having a TGase core computational techniques involving sequence analyses, comparative studies, building of phylogenetic trees, homology models and molecular dynamic simulations have been used. From this approach, a preliminary classification of these sequences was done by dividing them in five main groups. Each group has been investigated from the sequence point of view to analyze the presence of specific motifs. For three of this five groups, also the secondary structures have been investigated and, from this analysis, features specific for each group have

been detected. Moreover, two novel forms of microbial TGase (mTGase) have been investigated in the detail: *K. albida* mTGase and the hypothetical mTGase from *SaNDy* (organism not disclosed for patent opportunity). Molecular dynamics simulations and active site pocket analyses have been performed for the first, in comparison with MTGase. For the second, instead, experimental technique has been used to purify the hypothetical enzyme in order to test it on food related substrates. Experimental assays on both the proteins are still ongoing, to find the best enzymatic activity conditions and the best substrates of reaction. The molecular dynamic simulations performed on *K. albida* mTGase have suggested some explanations to the higher specificity of this enzyme than MTGase, experimentally demonstrated by Steffen et colleague, and several indications to change the activity conditions used to test it. Moreover, the substrates screening has allowed to find novel possible substrates, on which this enzyme could be employed for the allergenicity reduction. On the other hand, the enzyme extracted from *SaNDy*, showing a higher similarity with MTGase, could be less selective than *K. albida* mTGase for specific substrates, so it could be possible its application also on the gliadin substrate, but to prove it further experiments are necessary.

Note: the present PhD work has been mainly performed in the Bioinformatics Laboratory at the CNR of Avellino under Dr. Facchiano's supervision, however all the MD simulations have been performed at the Biochemistry Department of the University of Zurich, in the computational and structural biology laboratory under the supervision of Prof. A. Caflisch and his research group (compulsory abroad training period). Experimental activity assays on gliadin substrate have been performed by the spectrometry mass CeSMA-ProBio lab at the CNR of Avellino; and the hypothetical mTGase from *SaNDy* was instead cloned, expressed and purified in collaboration with the Laboratory for Molecular Sensing at the CNR of Avellino.

Index

Index of figures

Index of tables

LIST OF ABBREVIATIONS

AOECS	Association Of European Coeliac Societies
DH	Dermatitis Herpetiformis
EFSA	European Food Safety Authority
EmA	Anti-Endomysial Antibody
ESPAGHAN	European Society for Pediatric Gastroenterology Hepatology and Nutrition
FXIII	Plasma Human Transglutaminase/Factor XIII
GDP	Guanosine diphosphate
Gln	Glutamine
hTGase1	Human Keratinocyte Transglutaminase/Transglutaminase type 1
hTGase2	Human Tissue Transglutaminase/Transglutaminase type 2
hTGase3	Human Epidermal Transglutaminase/transglutaminase type 3
hTGase3	Human Prostate Transglutaminase/transglutaminase type 4
hTGase5	Human Transglutaminase type 5
hTGase6	Human Transglutaminase type 6
hTGase7	Human Transglutaminase type 7/ hTGase Z
hTGases	Human Transglutaminases
IPTG	Isopropyl β-D-1-thiogalactopyranoside
iTCLs	Gliadin-Specific Intestinal T-Cell Lines
KalbTGase	*Kutzneria Albida* Transglutaminase
K-C_2H_5	Lysine Ethyl Ester
K-CH_3	Lysine Methyl Ester
K-gliadins	Insoluble-Transamidated gliadin
LGG	*Lactobacillus Rhamnosus* GG
Lys	Lysine
MD	Molecular Dynamic
mTGase	Microbial Transglutaminase
MTGase	*Streptomyces mobaraensis* Transglutaminase
NCGS	Non-Celiac Gluten Sensitivity
OD	Optical Density
PDB	Protein Data Bank
QMEAN	Quality Model Energy Analysis
QPS	Qualified Presumption of Safety
SandyTGase	hypothetical mTGase sequence from *SaNDy* (organism not disclosed for patent opportunity)
SPF	Soluble Protein Fraction
SPI	Soy Protein Isolate
T1D	Type 1 Diabetes
TAP	Transglutaminase Activating Proteases
TAPI	Transglutaminase Activating Proteases Inhibitor
TGase	Transglutaminase
TGase3	Mammalian Epidermal Transglutaminase
TGase4	Mammalian Prostate Transglutaminase
TGl	*Bacillus Subtilis* Transglutaminase

INTRODUCTION

1.Transglutaminase enzymes

Transglutaminases, also named protein-glutamine gamma-glutamyltransferase (TGases; EC 2.3.2.13) are a class of enzyme widely spread in nature, being found in mammalian, vertebrates, invertebrates, mollusks, plants, and microorganisms (Martins et al., 2014).

TGases are enzymes that catalyze post-translational modification of proteins by acyl transfer reactions, deamidation and cross-link formation:

Acyl transfer reactions: occur through the transfer of γ-carboxamide groups of glutamine residues (acyl donor) in proteins/peptides to a variety of primary amines (acyl acceptor) (*Fig.1a*).

Deamidation: when lysine residues, free lysine or primary amines are absent from the reaction system, water becomes the acyl acceptor and hydrolytic deamidation of the glutamine residues occurs, transforming them into glutamic acid (*Fig.1b*). This reaction alters the protein charge, which also leads to changes in protein solubility (Kuraishi et al., 2001).

Inter- or intramolecular cross-linking (polymerisation): occurs through acyl-transfer between γ-carboxyamide groups of glutamine residues (acyl donor) and Ɛ-amino groups of lysine peptide residues (acyl acceptor) (*Fig.1c*) (Carvajal et al., 2011). These isopeptide bonds form a stable protein network that affects its solubility and therefore other functional properties, such as gelation, emulsification, foaming formation and viscosity, resulting from the production of changes in the hydrophobicity of the protein surface. Obtaining a stable protein network is very important, for example, in the formation of gels (Calmolezi Gaspar et al., 2014).

Among the three possible reactions, the hydrolytic deamidation is much slower than the linkage to primary amines and the formation of cross-links in the presence of accessible ε-amino side-chains from protein-bound lysine, respectively, moreover, in protein-containing food systems, the crosslinking reaction proceeds prior to the other reactions (Dube et al., 2007; Kuraishi et al., 2001).

The results of TGase activity are the modification of the protein conformation and other more extensive conformation changes due to bonding between the same protein and between different proteins, forming high molecular weight conjugates (Carvajal et al., 2011).

There are many examples of the TGase deamidation activity in nature. Among the most studied is the human tissue transglutaminase (hTGase2) deamidation activity on the gliadin peptides, which induces the trigger of the coeliac disease. Noteworthy is also the selective deamidation of Gln53 and Gln61 of the Rho family GTPase RhoA and Ras, respectively, by *Escherichia coli* cytotoxic factor and *Bordetella bronchiseptica* necrotoxin, which are considered functional relatives of TGase (Lorand and Graham, 2010).

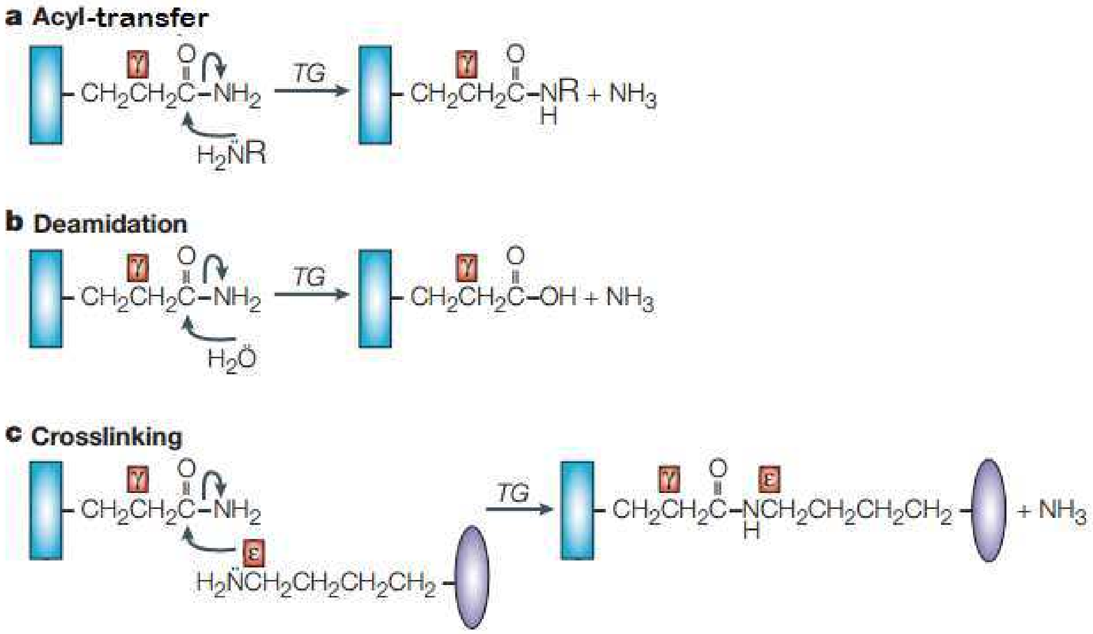


Fig1: **Transglutaminase catalyze post-translational reaction**. TGase can catalyze **a**: Acyl transfer reactions: incorporation of an amine (H_2NR) into the Gln residue of a protein (blue rectangle) **b**: Deamidation **c**: Protein cross-link by forming a N^{ε}(γ-glutamyl)lysine isopeptide bridge between the deprotonated Lys residue of one protein (purple ellipse) and the Gln residue of another (blue rectangle) (image from Lorand and Graham, 2010).

With regard to TGase-catalyzed amide incorporation and the cross-links formation, there is a wide range of examples: processes of crosslinking are involved in the clotting reactions, in the dimerization of interleukin-2, in the inactivation of the hepatitis C virus core protein, which can be modified by both amine incorporation and crosslinking by hTGase2. Further example of TGase amide incorporation could be represented by the conjugation to polyamines by γ-glutamyl linkages of proteins in the chloroplast of *Helianthus tuberosus* and the soybean storage protein glycinin (Lorand and Graham, 2010).

However, among these reactions, from the industrial point of view, only the cross-linking is of interest in modification of the techno-functional properties of proteins.

Furthermore, the reaction between the γ-carboxylamide groups and primary amines is a promising tool to improve the nutritional value of vegetable proteins by fortification with amino acids (Dube et al., 2007).

More in general, it is possible to say that the functional roles of the TGase are very different.

In multicellular organisms TGase reactions involved manly the transamidation of glutamine residues to lysine residues and the resulting side-chain to side-chain isopeptide bonds add strength to tissues and increase their resistance to degradation (Griffin et al, 2002).

In higher organisms, transglutaminases play important roles in diverse biological functions by selectively cross-linking proteins. Among the members are factor XIIIa, which stabilizes fibrin clots; keratinocyte transglutaminase and epidermal transglutaminase, which cross-link proteins on the outer surface of the squamous epithelium; and transglutaminase 2 (hTGase2), which shapes the extracellular matrix and promotes cell adhesion and motility (Santhi et al., 2017). In sea urchin and fish, TGases are necessary to the assembly and the elevation of the fertilization envelope (Ha and Iuchi, 1998)

In bacteria the functions of this enzyme are still unknown, even if some studies suggest that microbial transglutaminase (mTGase) could have a role in sporulation, being required for the cross-linking of a specific coat protein (Zilhão et al., 2005) or in the inhibition of several proteins during the development of aerial hyphae (Santhi et al., 2017).

In plants, TGases and their functionality have been less studied than in humans and animals (Carvajal et al. 2011). However, their physiological role in plants appears to be related to photosynthesis, fertilization, response to abiotic and biotic stresses, senescence, and programmed cell death (Martins et al. 2014).

In the following paragraphs the human and microbial TGases already known will be analyzed more in the details.

1.1 The human transglutaminases (hTGases)

Mammalian transglutaminase can be classified into immunogenically and genetically distinct types. Nine hTGase enzymes are present in humans; among them, eight are catalytically active and one, named the erythrocyte membrane protein band 4.2, is inactive. As mentioned in the previous paragraph, these proteins serve as scaffolds, maintain membrane integrity, regulate cell adhesion, and modulate signal transduction (Eckert. et al., 2014). Although the primary sequence of these enzyme shows differences, all share an identical not contiguous amino acid triad at the active site, i.e. Cysteine, Histidine, Aspartate.

Therefore, according to their physiological role and/or their organ-specific location hTGases are described as hTGase type 1, type 2, type 3, type 4, type 5, type 6, type 7, XIII factor and band 4.2. (Eckert et al., 2014).

HTGase type1 (hTGase1): it is the keratinocyte hTGase; it is expressed in the stratified squamous epithelia of the skin and upper digestive tract and in the lower female genital tract, is activated by proteolytic cleavage, increased Ca^{2+} level, and interaction with tazarotene-induced gene 3 (Eckert et al., 2014). HTGase1 together with hTGase 3 plays essential roles in epidermal keratinization. TGase 1 predominantly serves as a membrane-bound TGase isozyme (Thacher and Rice 1985), whose role is associated mainly with generation of the cross-linked cell envelope in epidermal keratinocytes. Mutations of the gene encoding membrane-bound TGase 1 elicit an autosomal recessive skin disorder known as lamellar ichthyosis, which results from an aberrant stratum corneum with the lipid and cornified envelopes being seriously injured (Terazawa et al., 2015).

HTGase type2 (hTGase2): it is the tissue transglutaminase; it is a multifunctional, Ca^{2+} dependent enzyme, which has been involved in the regulation of cell growth, differentiation, and apoptosis (Peng et al., 1999). In addition to the transamidation reaction, this enzyme displays GTPase, ATPase, protein kinase, and protein disulfide isomerase activity. It interacts with phospholipase $C\delta_1$, β-integrins, fibronectin, osteonectin, RhoA, multilineage kinases, retinoblastoma protein, PTEN, and IκBα (Eckert et al., 2014).

HTGase2 has been localized mainly in the cytosol, however, a detectable hTGase2 expression, both as a cross-linking enzyme and a G-protein, has been reported in the nucleus where it may be actively transported into via binding to an importin-α3/Qip-1 family protein (Peng et al., 1999). Moreover, hTGase2 is also localized in plasma (10-15%) and the nuclear membranes (5%); it is secreted by unknown mechanism from the cell, where it localizes to the cell surface and the extracellular matrix (Lorand and Graham, 2010).

Retinoic acid, vitamin D, TGF-β1, IL-6, tumor necrosis factor (TNF), NF-κB, epidermal growth factor (EGF), phorbol ester, oxidative stress, and Hox-A7 induce hTGase2 expression. HTGase2 dysfunction contributes to celiac disease, neurodegenerative disorders, and cataract formation (Eckert et al., 2014).

HTGase type3 (hTGase3): it is the epidermal hTGase; it is present in hair follicles, epidermis, and brain. Like TG2, TG3 binds to and hydrolyzes GTP. It catalyzes the crosslinking of trichohyalin and keratin intermediate filaments to harden the inner root sheath of a hair follicle, which is critical for hair fiber morphogenesis (Eckert et al., 2014).

Moreover, the enzyme is thought to be critically involved in the cross-linking of structural proteins and in the formation of the cornified cell envelope, thereby contributing to rigid structures that play vital roles in shape determination and/or barrier functions (Lee et al., 1996).

TGase3 knockout mice show impaired hair development and increased fragility of isolated corneocytes of the skin barrier (Bognar et al., 2014).

HTGase type4 (hTGase4): it is the prostate hTGase; it is present in the prostate gland, prostatic fluids, and seminal plasma (Eckert et al., 2014). TGase4 knockout mice show defects in copulatory plug formation resulting in a fertility reduction (Dean 2013). In rats, TGase4 participates in masking the antigenicity of the male gamete via incorporating seminal protein, such as uteroglobin, or polyamines into sperm cell surfaces (Cho et al., 2010).

In contrast to rodent, the exact function of TGase4 in humans is not known, but some recent reports suggest that this enzyme showed prostate-restricted expression pattern (Dubbink et al., 1999). These results raise a possibility to use this enzyme as a novel target for prostate-related diseases, particularly in prostate cancer, due to the observation of a link between increased expression of hTGase4 and promotion of an aggressive prostate cancer phenotype (Cho et al., 2010; Eckert et al., 2014)

HTGase type5 (hTGase5): transglutaminase 5 is mainly expressed in foreskin keratinocytes, epithelial barrier lining, and skeletal muscle. hTGase5 crosslinks loricrin, involucrin, and SPR3 in epidermis and contributes to hyperkeratosis in ichthyosis and psoriasis patients. hTGase5 inactivating mutations result in skin peeling syndrome (Eckert et al., 2014). Studies observed that hTGase 5 contributes, as a secondary effect, to the hyperkeratotic phenotype in ichthyosis (both vulgaris and lamellar) and in psoriasis. Moreover, in patients affected by Darier's disease, an autosomal dominant skin disorder characterized by loss of adhesion between epidermal cells (acantholysis) and abnormal keratinization, hTGase5 expression, as well as hTGase3, is completely missregulated, being overexpressed or totally absent in different areas of the same lesion (Candi et al., 2002).

HTGase type6 (hTGase6): hTGase6 expression is localized in the human testes and lungs, and in the brain of mice (Eckert et al., 2014). Analysis of its temporal and spatial pattern of induction in mouse development indicates an association with neurogenesis. Autoantibodies to hTGase6 were identified in immune-mediated ataxia in patients with gluten sensitivity, suggesting a critical role for hTGase6 in cortical and cerebellar neurons. Moreover, hTGase6 is also expressed in human carcinoma cells with neuronal characteristics. Molecular modelling of hTGase6 indicates that this enzyme could have Ca^{2+} and GDP-binding sites related to those of hTGase3 and hTGase2, respectively (Thomas et al., 2011).

HTGase type7 (hTGase7): hTGase 7, also known as hTGase Z, is widespread expressed in different tissues (Grenard et al., 2001). One report suggested a correlation between hTGase7 transcript levels and cancer cells (Eckert et al., 2014). Significantly increased levels of transcripts of hTGases7, together with higher levels of hTGase4 and significantly lower levels of FXIII, were seen in tumor tissues; instead, lowest levels of hTGases-7 and 3 were seen in patients with metastatic disease. Moreover, breast cancer displays an aberrant expression of TGases, wherein the levels of TGases 7, 2 and 3 have a relationship with node involvement and patient outcome (Jiang et al., 2003).

XIII Factor (FXIII): Plasma hTGase (FXIII) is found in cells of osteoblast lineage, in plasma, platelets, macrophages, astrocytes, dermal dendritic cells, chondrocytes, synovial fluid, the placenta, the eyes, and the heart. It is an important component of the blood coagulation cascade because it is the last zymogen activated in the blood coagulation cascade. In the presence of Ca^{2+}, the enzyme catalyzes crosslinking of fibrin molecules to stabilize fibrin clots. FXIII also plays a role in inflammation and bone synthesis (Eckert et al., 2014). Moreover, this enzyme catalyzed also the crosslinking of the Angiotensin II receptor type 1, resulting in enhanced monocyte adhesiveness of hypertensive patients and thereby may sustain the process of atherogenesis by chronic sensitization of circulating monocytes (AbdAlla et al, 2004).

FXIII exists as the A_2 homodimer in many cells (such as platelets, macrophages, astrocytes and chondrocytes) and as A_2B_2 heterotetramer in blood plasma. The active site is located in the A subunits, and the B subunits serve as a carrier protein protecting the A subunits in the circulation and contributing to the binding of the complex to fibrinogen; moreover B subunits act as a brake on FXIII activation (Lorand and Graham, 2010). FXIII deficiency is an extremely rare autosomal recessive disorder characterized by a lifelong bleeding tendency, including recurrent pregnancy loss, umbilical cord bleeding, and intracranial hemorrhage. Life-threatening bleeding episodes can be reduced significantly with timely diagnosis of FXIII deficiency and appropriate prophylaxis utilizing blood-derived components such as fresh frozen plasma and FXIII concentrate, or recombinant FXIII (Dorgalaleh et al. 2016).

Band 4.2: Band 4.2 is a unique hTGase that lacks catalytic activity. The Cysteine-Alanine substitution in band 4.2 of the cysteine corresponding to the active site in the other forms is responsible for the lack of enzymatic activity.

Band 4.2 is mainly present in erythrocytes, bone marrow, fetal liver, and spleen. Band 4.2 is a major component of the erythrocyte membrane cytoskeleton and plays an important role in maintenance of membrane integrity and regulation of cell stability (Eckert et al., 2014). Actually, Band 4.2 is a major membrane-associated protein of 72kDa and is present at≈200,000 copies per cell. It is clearly important for normal erythrocyte function since patients whose erythrocytes are deficient in or lack band 4.2 are anemic due to accelerated erythrocyte destruction and have abnormally shaped possibly fragile, erythrocytes (Korsgren and Cohen, 1991). Band 4.2 null mice show alterations in red blood cell function, including spherocytosis and altered ion transport.

From this brief overview is possible to conclude that most tissues express multiple TGase forms and share common substrates. This may explain why TGase family members can compensate for the loss of an individual enzyme. Studies on TGase2 gene knockout mouse indicated tissue-specific mechanisms of compensation for the loss of TG2, including transcriptional compensation in heart and liver versus functional compensation in aorta, kidney and skeletal/cartilagenous tissues. On the contrary, no compensation has been detected in skeletal muscle, suggesting a limited role for the TGase2-mediated transamidation in normal development of this tissue (Deasey and Nurminskaya, 2013).

Of the eight catalytic hTGases, hTGase2 has been the most comprehensively studied due to its ubiquitous expression in multiple cell types, that has allowed to obtain a wider knowledge about the hTGase world and their regulation and activities.

1.1.a The human tissue transglutaminase: structure, activity and allosteric regulation.

The multifunctionality of hTGase2 is dependent on its structural features. HTGase2 is composed of four domains: a N-terminal β-sandwich domain (aa 1–140), the catalytic core (aa 141–460) and two C-terminal β-barrel domains (aa 461–586 and 587–687) (*Fig.2a*). The site of transamidating activity is composed of the catalytic triad of cysteine proteases (*Fig.2b*): cysteine 277 (C277), histidine 335 (H335) and aspartate 358 (D358).

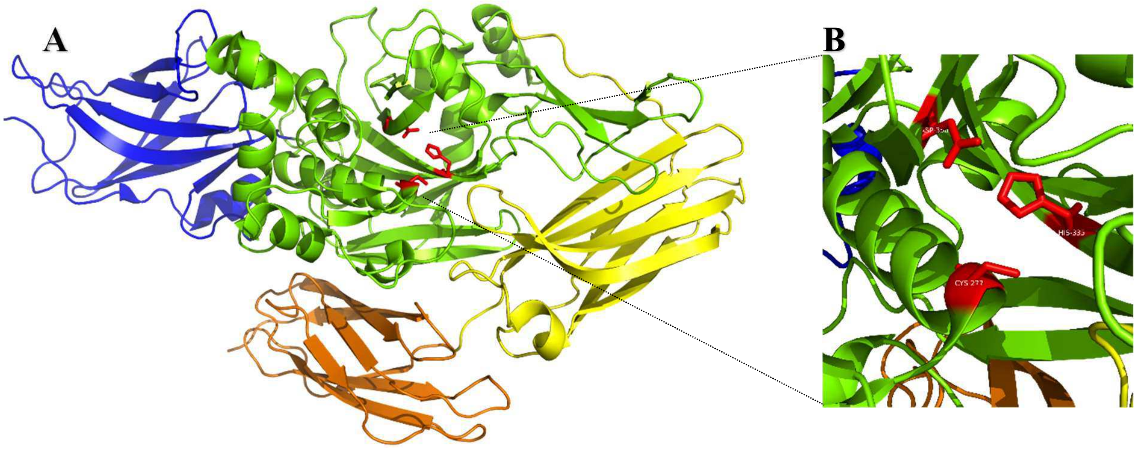


Fig.2: **Structure of the human tissue transglutaminase.** HTGase2 PDB code 4PYG shows **A:** 4 domains: N-terminal β-sandwich domain in blue, the catalytic core in green (with the catalytic triad highlighted in red) and the C-terminal β-barrel-1 and β-barrel-2 domains in yellow and orange, respectively; **B**: the catalytic triad is composed of C277, H335 and D358.

The enzyme is regulated by redox potential, Ca^{2+} and GDP, particularly Ca^{2+} induces enzyme activation, instead GDP induces enzyme inhibition.

The mechanism for hTGase2-catalysed transamidation (*Fig.3*) is proposed on the basis of papain-reaction mechanism; due to their similarities in the catalytic triad and reaction mechanism. The reaction primarily involves the exchange of primary amines for ammonia at the γ-carboxamide group of glutamine residues, in the presence of Ca^{2+}. The binding of Ca^{2+} is vital to the cross-link formation because it initiates a conformational change that exposes a cysteine residue in the active site domain; the cysteine reacts with the glutamine substrate, resulting to the formation of an acyl-enzyme intermediate and release of ammonia. The subsequent reaction between the acyl-enzyme complex and a primary amine results to the formation of γ-glutamyl-amino cross-link, and concomitant release of the enzyme (Onyekachi and Coussons, 2014).

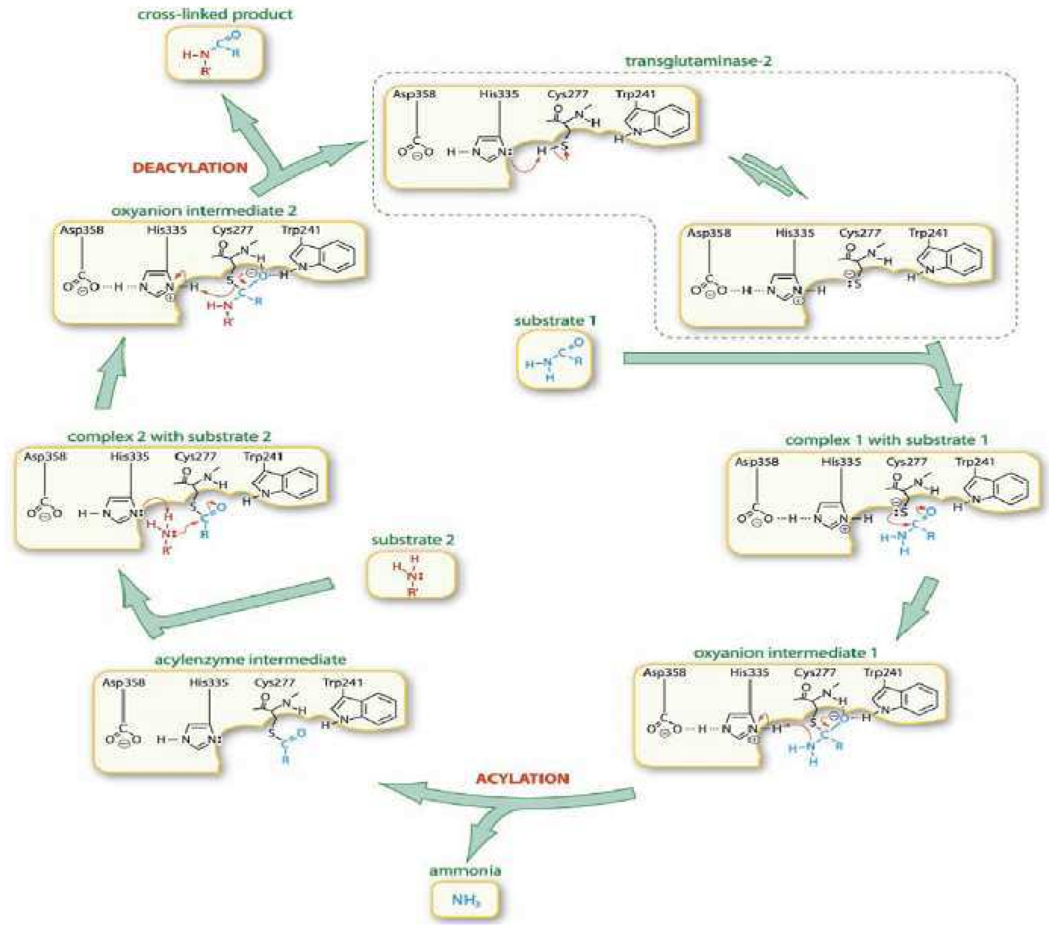


***Fig. 3*: The reaction mechanism of hTGase2** (image from Onyekachi and Coussons, 2014)

If the Ca^{2+} is important to induce the conformational change that exposes the catalytic cysteine residue, the GDP molecule, bounded at the hTGase2, blocks the access to the active site (*Fig.4*).

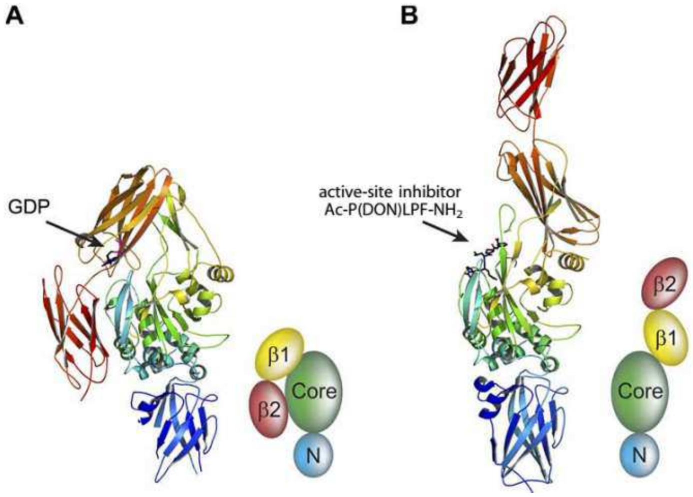


Fig.4: **Open and closed conformation of the hTGase2**. The N-terminal β-sandwich is shown in blue ribbons (N), the catalytic domain (Core) in green ribbons, and the C-terminal β-barrels (β1 and β2) in yellow and red ribbons, respectively. **A**: Closed conformation: GDP-bound hTGase2. **B**: Open conformation: hTGase2 inhibited with the active-site inhibitor Ac-P(DON)LPF-NH2, which mimics the acyl-donor substrate. Simplified cartoons are included for clarity (Pinkas et al., 2007).

In the GDP-bound form of hTGase2, access to the transamidation active site is blocked by a loop connecting the third and fourth β-strands, as well as by a loop connecting the fifth and sixth β-strands of the first β-barrel domain (*Fig.5*). Tyr-516, which is conserved in hTGases and located in the first loop, forms a hydrogen bond with Cys-277 (*Fig.5*). Transamidation activity requires an accessible Cys-277, and Tyr-516 (Y516) with its associated loop from the first β-barrel domain must move to make the active site accessible to substrates (Liu et al., 2002).

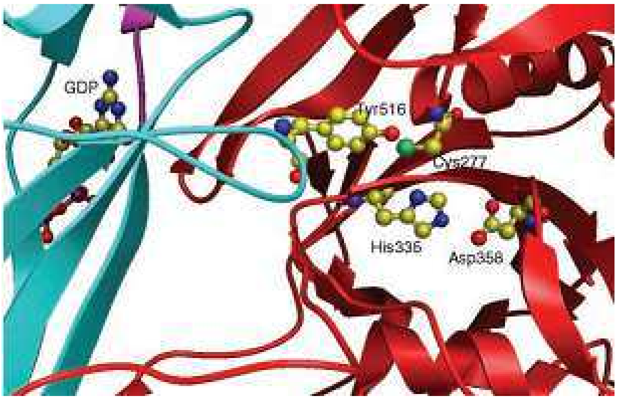


Fig.5: **Transamidation active site of hTGase2 in the closed conformation.** Highlighted by red ribbons the catalytic core with the catalytic triad in ball and stick (C277–H335–D358) by the blue ribbons the first β-barrel domain with Y516 (showed in ball and stick) relative to the guanine nucleotide-binding site. Y516 points toward C277, the catalytic nucleophile, in the active site.

The GDP molecule engages both the first and last β-strands of the first β-barrel domain, which should maintain the inactive state by stabilizing the loops that block the access to the catalytic domain.

ATP and GDP were found to bind the same nucleotide-binding pocket. It is a perfect example of a tertiary pocket, with contributing residues from the catalytic domain β-barrel-1 and β-barrel-2 domains. It is composed of at least ten residues, many of which are involved in both adenine and guanine nucleotide binding; however, Ser482 and Arg580 were found to be involved only in guanine, not adenine, nucleotide binding (Gundemir et al., 2012).

If the GDP binding pocket has been resolved, the Ca^{2+} binding sites are still unknown, so they have been proposed comparing hTGase2 structure with FXIII structure, whose Ca^{2+} pockets have been demonstrated. A putative Ca^{2+} binding site, homologous to the one of FXIII, is distorted in hTGase2, with the largest difference occurring in the proximity of Ser-419 (equivalent to Ala-457 in FXIII), due to the bound nucleotide. Ca^{2+} binding, by altering the position of the Ile-416–Ser-419 peptide, would eliminate the stabilizing effects of these peptides on the nucleotide-binding site and could thereby weaken nucleotide binding, as has been observed experimentally. In consequence of this weakening nucleotide binding; the protein has a conformational change, which involves the substrate binding and the related displacement of the hydrogen bonded Tyrosine, making the active site accessible (Liu et al., 2002; Fesus and Piacentini 2002).

Actually, in addition to the catalytic triad, two conserved tryptophan residues (W241 and W332), located at the opposite sides of the "catalytic tunnel", are critical for the transamidating activity, since these residues stabilize the enzyme-thiol intermediate that forms during catalysis. A threonine residue (T360) at the entrance of the catalytic tunnel controls the entry of the acyl-acceptors for the second step of the catalysis. The mutation of this residue increases the preference for deamination over transamidation. Another important residue in the catalytic site is the tyrosine residue at the position 516 (Y516). Actually, a hydrogen bond forms between C277 and Y516 in the closed conformation of hTGase2, which is believed to further stabilize the closed conformation and keep the enzyme inactive (Gundemir et al., 2012).

1.2 Microbial transglutaminase

TGases have been found in prokaryotes and eukaryotes, including guinea pig, and were first extracted from the liver of this animal in 1973 (Folk and Chung 1973). This TGase was the only form arriving at the market by the end of the 80s, not arousing much interest from the industrial point of view, since it was very expensive. In addition, as the other mammalian TGases, it was a Ca^{2+}-dependent enzyme, which led to the precipitation of proteins of some foods containing casein, soybean globulin, or myosin (Martins et al., 2014).

However, TGases are present also in bacteria, and peculiar forms of this enzyme have been widely characterized. In 1989, a Ca^{2+}-independent microbial TGase from *Streptomyces mobaraensis* (MTGase) was extracted by Ando et al. Purification was rather easy, and the production of MTG has been established commercially and applied in research and biotechnology fields (Dube et al., 2007; Calmolezi Gaspar and de Góes-Favoni, 2015).

This enzyme is characterized by the same triad that appears in the mammalian TGase with a different order, i.e. Cysteine, Aspartate, Histidine. Another relevant difference concerns the structural organization as a single domain in the bacterium enzyme, while in eukaryotes, four structural domains are observed. Nowadays mTGase by *Streptomyces mobaraensis* is commercially available and widely used in biopolymers industry, cosmetics, clinical applications, wool textiles, and food processing industry (Carvajal et al. 2011). Since the early 1990s, many mTGase-producing strains have been found, and production processes have been optimized. In the meantime, novel bacterial forms of the enzyme have been investigated. From a screening of microorganisms that produce MTGase, an actinomycetes strain, T-2, has been isolated from soil and identified as *Actinomadura sp.* (Kim et al., 2000). A protein with the deamidation activity typical of TGases has been isolated from *Chryseobacterium sp*. (Yamaguchi et al., 2001), but it lacks other typical activities of TGases and has been defined as a protein-glutaminase, a different enzyme class (EC 3.5.1). Transglutaminase activity in vitro has been demonstrated for the periplasmic portion of a protein from *Pseudomonas aeruginosa*, named TgpA (Milani et al., 2012). A small microbial transglutaminase from *Bacillus Subtilis* works through a unique partially redundant catalytic dyad formed by Cysteine and Glutamine residues, with a Histidine residue that also plays a role in the function of the enzyme, that is reduced but not eliminated by its mutation (Fernandes et al., 2015). TGase from *Kutzneria albida* (KalbTGase) has been characterized for 3D structure and specificity of substrate recognition, for potential applications in highly site-specific labelling (Steffen et al., 2017). Therefore, investigations on microbial TGases during the recent years suggest evolutive differences in specificity and catalytic mechanisms within the world of microbial TGases, with many aspects to be deciphered, too. The large application of genome sequencing in the last years increased the availability of sequences and make it possible now a large-scale comparison, in order to implement more efficient system for mTGase production and to obtain mTGase from other microorganisms for an industrial utilization.

1.2.a *Streptomyces mobaraensis* Transglutaminase (MTGase)

Ando et al. in 1989 isolated about 5,000 strains from soil, collected from a variety of locations, and investigated hydroxamate-forming activity, found strong enzyme activity in an actinomycete strain that seemed to be *Streptoverticillium S-SI12*, later classified as *Streptomcyes mobaraensis*.

According to Ando et al. reports, MTG is a calcium independent enzyme with a molecular weight of 40,000 Da, isoelectric point about 8.9 and optimal pH is from 6 to 7 with the reaction time of 10 min at 37°C. This was an important finding for the utilization of MTG, a low cost and easy to purify enzyme, in food processing. Actually, this enzyme is able to cross-link most food proteins including the meat proteins through Glutamine-Lysine bond (Santhi et al., 2017) The amino acid sequence is very different from that of mammalian TGases, therefore, no sequence homology is detectable. Moreover, MTGase exhibits a unique 3D structure compared to the mammalian TGase. Actually, despite the hTGase2 composed of four domains, this enzyme consists in a unique catalytic domain, containing a central 8-stranded β-sheet surrounded by 11 α-helices (*Fig.6A*). Both the human transglutaminases and mTGase contain a Cysteine-Histidine-Aspartate catalytic triad (*Fig.6B*); however, the structural orientation differs. Relative to the active site cysteine, the position of Histidine and Aspartate are reversed in the two enzymes. The active site is a 16 Å deep cleft created by two protruding loops, with the catalytic cysteine located at the bottom of the cleft (Strop, 2014).

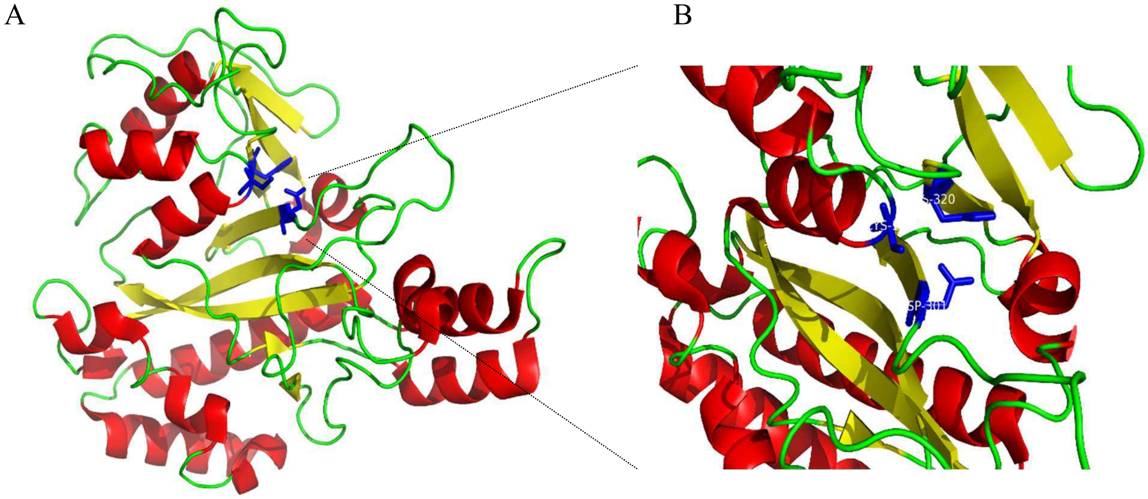


Fig6: **Structure of the *Streptomyces mobaraensis* transglutaminase.** MTGase PDB code 3IU0 shows **A**| the catalytic core composed of 8 central β-strands (highlighted in yellow ribbons) surrounded by 11 α-helices (highlighted in red ribbons), with the catalytic triad highlighted in blue sticks **B**| the catalytic triad is composed of C110, D301 and H320.

The mechanism of human transglutaminases is thought to be similar to a number of cysteine proteases where the first step consists of deprotonation of the active site cysteine thiol by nearby histidine. In MTGase, it was proposed that catalytic Aspartate plays a similar role as the catalytic histidine in hTGase2. The critical catalytic role of Aspartate in MTGase was also confirmed by alanine mutagenesis where the activity of mutant was reduced to background levels. Furthermore, mutagenesis of MTGase catalytic Histidine reduced the catalytic activity only by 50%, suggesting that this residue does not play a critical role (Kashiwagi et al., 2002). The enzyme is produced as a zymogen, where the N-terminus folds into a helical structure that occupies the active site and blocks substrate access (*Fig.7*). The zymogen is activated by the cleavage and dissociation of the N-terminal helical structure by endogenous metalloprotease and tripeptidyl aminopeptidase (Strop, 2014).

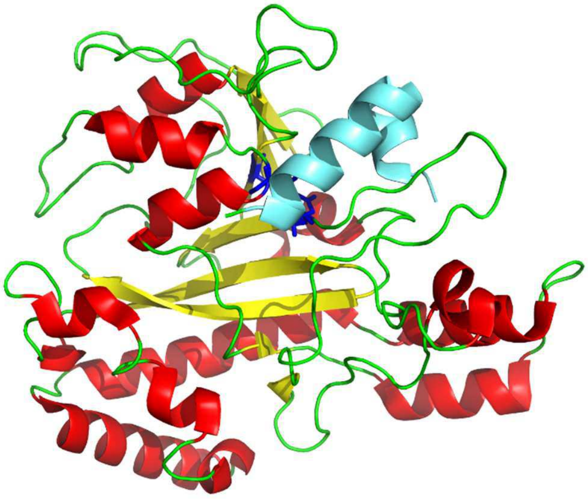


Fig.7**: Structure of the *Streptomyces mobaraensis* zymogen transglutaminase.** The N-terminus folds into a helical structure (cyan ribbons) that occupies the active site and blocks substrate access.

The discovery of transglutaminase zymogen (pro-transglutaminase) has revealed the activation mechanism of transglutaminase from *Streptomyces*. In 1998, Pasternack and colleagues found that MTGase was secreted as a pro-transglutaminase and could be activated by several exogenous proteases, such as bovine trypsin, intestinal chymotrypsin.

Subsequently a metalloprotease was isolated from *Streptomyces mobaraensis* as an endogenous transglutaminase-activating protease (Zotzel et al., 2003). Recent studies suggest that *Streptomyces* pro-transglutaminases have a conserved amino acid sequence preceding the N-terminal domain of transglutaminase, which contains cleavage sites for both serine protease and metalloprotease, indicating that activation of pro-transglutaminase is not a specific process (Zhang et al., 2008).

Further researches have proven that transglutaminase activation process is inhibited by a transglutaminase activating protease inhibitor (TAPI), that it is a member of the *Streptomyces* subtilisin inhibitor family. TAPI possess surface activity, therefore these molecules are distributed mostly at the air-liquid interface, allowing the existence of enough free transglutaminase-activating proteases (TAP) molecules in the submerged liquid. This is what makes TAP able to perform their function to activate pro-transglutaminase (Zhang et al., 2009). Metalloprotease, serine protease and *Streptomyces* subtilisin inhibitor protein, the key factors involved in the activation process of transglutaminase, are all under regulation by the *A-factor,* a microbial hormone controlling the differentiation of *Streptomyces* (Kato et al., 2002). Actually, MTGase results secreted and activated during differentiation rather than during nutrition growth, suggesting a strategical role in this phase (Zhang et al., 2009).

A better comprehension of the activation mechanisms and the biological role of the MTGase is crucial from the industrial point of view. So far, based on the activation mechanism of pro-transglutaminase in *Streptomyces*, novel strategies have been developed to improve the production: i.e. the use of molecules as the cetyltrimethyl ammonium bromide to remove the inhibition of TAPI or the addition of protease in the prophase of fermentation. Moreover, because transglutaminase secretion is associated with differentiation of *Streptomcyes*, appropriate feeding strategy are under assessment in order to enhance transglutaminase production (Zhang et al., 2009).

1.2.b *Bacillus subtilis* Transglutaminase (TGl)

TGl is a transglutaminase extracted from the bacterium *Bacillus subtilis*. It is produced during sporulation and cross-links the surface of the highly resilient spore.

TGl is the smallest transglutaminase characterized to date, with a molecular weight of 38KDa. It is a single-domain protein, is produced in active form and no factor as Ca^{2+} or GTP is known to control its activity (Fernandes et al, 2015).

Bacillus spores are known to be resistant to a variety of environmental stresses including heat, organic solvents, ultraviolet radiation, X-rays, hydrogen peroxide and lysozyme. This resistance is due to the presence of a coat, which consists of various polypeptides. The structure of the spore coat is similar to that of keratin and some of these proteins are insoluble even under extreme conditions (Kobayashi et al, 1996). Since often, in a large variety of animals, insoluble structural proteins have Ɛ-(γ-Glu) Lys crosslinks, these bonds have been researched also in the *Bacillus subtilis* spore coat. The results of these studies showed the detection of Ɛ-(γ-Glu) Lys crosslinks in both spore coat fraction and spore coat proteins, isolated from disrupted spores, and a significant TGase activity in sporulating cells at 6 or 10h after the beginning of the stationary phase (Kobayashi et al, 1996; Zilhão et al., 2005).

Activity assays performed on *Bacillus subtilis* TGase demonstrated that this enzyme presents optimal temperature and pH values of respectively 60 °C and 8.2 (Soares de Barros et al, 2003) and is able to catalyze cross-links; in particular, the gel forming reaction of α_s-casein and BSA by the formation of Ɛ-(γ-Glu)Lys isopeptides (Suzuki et al., 2000).

Crystalized in 2015 by Fernandes C. G. et al., the enzyme exhibits the NlpC/P60 fold at its catalytic core, suggesting to be structurally related to a group of bacterial cell wall endopeptidase. NlpC/P60 domain is the current closest representative of the minimal ancestral structural unit of the thiol protease fold, it takes its name from the bacterial NlpC/P60 cell wall endopeptidases, which show a catalytic triad composed of a Cysteine, Histidine, and a third polar residue.

Moreover, the detected structure of TGl shows a unique, partially redundant, catalytic dyad, where the catalytic Cys116 is insulated within a hydrophobic tunnel that traverses the molecule from side to side (Fernandes et al., 2015).

In general, the molecule comprises three β-sheets (two-, four-, and six-β-stranded) and 10 helices, of which three are 3_{10}-helices (*Fig.8A*). Cys116, essential for the activity of TGl both in vivo and in vitro, is located at the N-terminus of a long helix, α6.

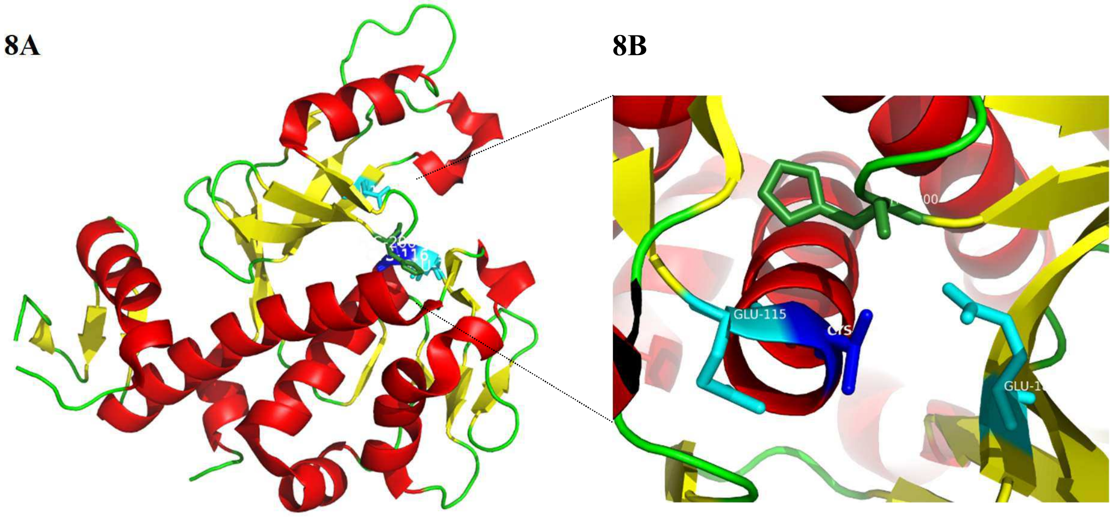


Fig.8: **Structure of the *Bacillus subtilis* transglutaminase.** TGl PDB code 4PA5 shows **A**: the catalytic core composed of 12 β-stranded (highlighted in yellow ribbons) and 10 α-helices (highlighted in red ribbons), with the catalytic residues highlighted in sticks; **B**: the catalytic residues C116 in cyan, E115 and E187 in blue; H200 is highlighted in green sticks.

The superimposition of the catalytic Cysteine in TGl (Cys116), MTGase (Cys101), and TGase 3 (Cys272) places His200 of TGl in the same relative position of His320 in MTGase, raising the possibility that His200 is also not essential for catalysis. Furthermore, it also suggests that Glu187 in TGl could be a catalytic residue: not only because it occupies the position equivalent to His330 in TGase 3 or Asp301 in MTGase, but also because the nearest side-chain atoms of Glu187 are close (~4.3 Å) to the Sγ atom of the catalytic Cys116 residue. However, despite of that, from the purification and the activity, testing in parallel with the wild-type enzyme, of the enzyme bearing single Alanine substitutions of residues Cys116 (TglC116A), Glu187 (TglE187A), and His200 (TglH200A), was detected that unlike Cys116, Glu187 and His200 are not essential for catalysis. TglH200A and TglE187A, in fact, still retains considerable activity in both amine incorporation and crosslinking assays, raising the possibility that another residue, near the catalytic Cys116, can compensate for the absence of Glu187, which, for its interactions, appears to be a catalytic residue serving the role of the proton acceptor for Cys116. Actually, it was found a second acidic residue, Glu115, in the close vicinity of Cys116, which can substitute for Glu187 and whose substitution (E115A) abrogates enzyme activity.

However, if the role of Glu 187 seems to be the primary proton acceptor for Cys116, His200, located at the back entrance of the catalytic tunnel, may be involved in substrate recognition, contributing in that way to the overall activity of the enzyme. In particular, strong positive correlations between the His200-Glu187 and Cys116-Glu115 interactions have been found. The interaction between His200 and Glu187 weakens the Glu187-Cys116 interaction, so that Cys116 turns to the free Glu115. Similarly, the Cys116-Glu187 and His200-Glu115 pairs are positively correlated. It follows that the His200-Glu115 interaction directs the catalytic Cys116 to the alternative Glu187. These observations suggest that TGl uses a catalytic dyad formed by either Cys116 and Glu187 or Cys116 and Glu115. (Fernandes et al, 2015).

1.2.c *Chryseobacterium sp.* Transglutaminase

In 2001 Yamaguchi and colleagues published the discovery of a novel protein-deamidating enzyme, purified to homogeneity from nonpathogenic bacterium *Chryseobacterium proteolyticum* and the cloning of the gene encoding it in *E. coli*.

The enzyme is a monomer with a pI of 10.0, a measured M_r of ≈ 20000 and a calculated M_r of 19860. It is an extracellular enzyme, expressed as a prepro-protein with a putative signal peptide of 21 amino acids and a pro-sequence of 114 amino acids. The amino-acid sequence has no obvious homology to any sequence included in public database.

In their studies, the researchers performed extensive comparison with MTGase in order to examine the putative transglutaminase activity of this enzyme. In particular, the purified enzyme was compared with MTGase in terms of hydroxamate-formation between benzyloxycarbonyl-Gln-Gly and hydroxylamine, deamidation of benzyloxycarbonyl-Gln-Gly, and amine-incorporation into casein. The results showed that the protein-deamidating enzyme lacked transglutaminase activity in terms of hydroxamate formation between benzyloxycarbonyl-Gln-Gly and hydroxylamine, or monodansylcadaverine (a fluorescent primary amine) incorporation into casein, however showed also an initial rate for deamidation of the purified enzyme equal to 25.01 μmol·min^{-1}· mg^{-1}.

To determine which amino acids in protein are deamidated by the enzyme, Yamaguchi and collegues incubated oxidized insulin A and B chains with the enzyme, monitoring the released ammonia during the reactions and determining the complete amino-acid sequence of the deamidated A chain. The results showed that the enzyme deamidates the two glutaminyl residues (Gln5 and Gln15) in the oxidized insulin A chain; moreover, it was also observed that it deamidates both casein and the oxidized insulin B chain (long chain peptide) with higher catalytic efficiencies (kcat /Km) than with short peptides.

The enzyme is active against several proteins as milk caseins and insoluble wheat gluten, whereas the activity against bovine serum albumin and hen egg ovalbumin is very poor. It does not deamidate asparaginyl residues in peptides, free glutamine or other amides.

Due to these results and the absence of any proof of its cross-links activity, the enzyme was named protein-glutaminase (EC 3.5.1).

Additionally, Yamagouchi studies demonstrate that the enzyme shows the highest activity at pH 5.0 and more than 90% of the remaining activity at pH 5± 8.7 after incubation of the enzyme in the buffers at various pHs for 18 h. The optimal temperature for the activity is between 50 and 60°C when the activity is assayed in sodium phosphate buffers for 10 min at various temperatures. The heat stability studies indicate the protein-glutaminase retains more than 93% of its activity after incubation at up to 55°C for 60 min. The enzyme loses 21% and 90% of its activity after incubation at 60°C for 10 min and 60 min, respectively (Yamaguchi et al, 2001).

So far, the structure of this enzyme is still unresolved and, due to its particular amino acids composition, very different from the others whose structure is already known, no models are available.

1.2.d *Kutzneria albida* Transglutaminase (KalbTGase)

In 2017, it was published by W. Steffen et al., from Roche Diagnostics GmbH, the functional and structural characterization of a novel microbial transglutaminase extracted from *Kutzneria Albida*, (KalbTG), which exhibited no cross-reactivity with known MTG substrates or commonly used target proteins, such as antibodies. KalbTGase was produced in *Escherichia coli* as soluble and active enzyme in the presence of its natural inhibitor ammonium to prevent potentially toxic cross-linking activity. The crystal structure of KalbTG revealed a conserved core similar to other mTGases but very short surface loops, making it one of the smallest mTGases characterized to date. KalbTGase and MTGase show 30% of sequence similarity with a distinct conservation of the active site residues Cysteine-Aspartate-Histidine. The *K.albida* gene product is significantly smaller than MTGase, amounting to a calculated molecular mass of 30.1kDa and a molecular mass of 26.4kDa in the active form, even smaller by 2kDa than the structurally unrelated TGl. Because MTGase is produced as inactive proenzyme and processed by extracellular proteases to yield the 38kDa active form, it has been assumed a similar activation mechanism also for KalbTGase. From the resolved structure, it is possible to observe the catalytic Cys82-Asp211-His224 triad located at the bottom of the active site groove (*Fig.9B*). The groove is wide enough to be covered by a kinked helical pro-peptide in the unprocessed enzyme, similar to what has been observed with the MTGase zymogen. As in all other TGase structures, the catalytic Cys82 in KalbTGase is located at the N terminus of an α-helix, which reduces its pKa and increases its nucleophilicity for attack on the substrate glutamine. Of note, the catalytic Cys82 in KalbTGase is embedded 1.7Å deeper in the active cleft than its MTGase counterpart. More in general, from the superimposition of KalbTGase structure with the MTGase structure is possible to see a similar disc-shaped core structure (root mean square deviation of 1.5Å) of a central β-sheet with flanking α-helices and a surface depression forming the active site cleft (*Fig.9A*). However, KalbTGase is more compact, having much shorter surface loops.

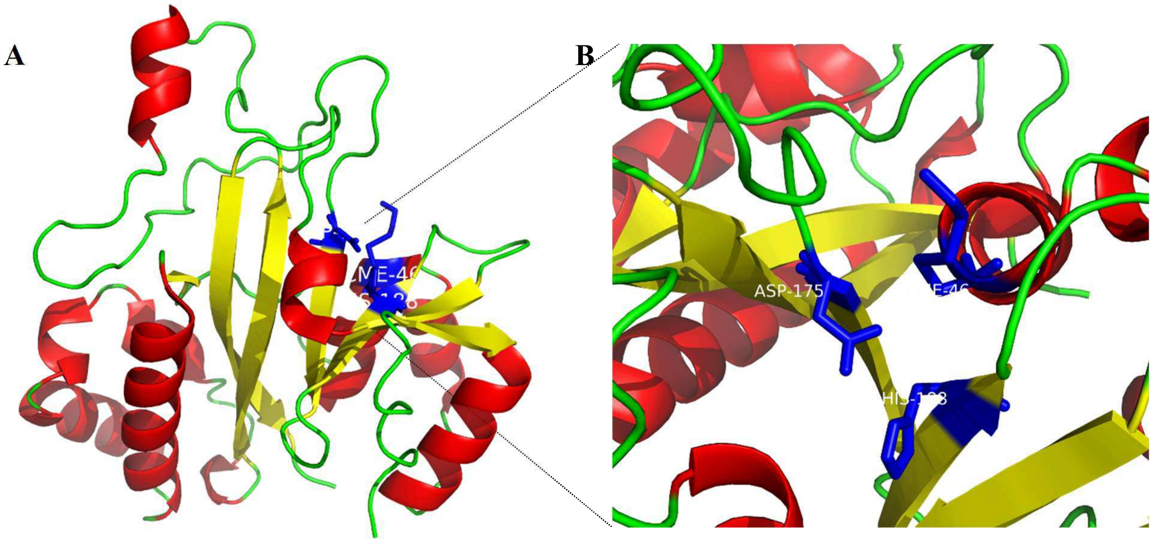


Fig.9: **Structure of the *Kutzneria albida* transglutaminase.** KalbTGase PDB code 5M6Q shows **A**: the catalytic core composed of a central β-sheet(highlighted in yellow ribbons) with flanking α-helices (highlighted in red ribbons), and a surface depression forming the active site cleft with the catalytic residues highlighted in blue sticks **B**: the catalytic residues Cys82-Asp211-His224 highlighted in blue sticks are numerated with a difference of 36, corresponding to the first 36 amino acids not crystalized because considered to be the pro-peptide part.

In order to test KalbTGase cross-links activity, to find the subtrates that could react with this enzyme and compare its specificity with MTGase, Steffen et colleagues synthetized arrays with millions of spatially addressable peptides using a light-directed, digitally controlled process and developed methods for in situ analysis of enzyme activity and substrate specificity for both KalbTGase and MTGase. More in the details, after they confirmed that the mature KalbTGase possess basic microbial transglutaminase activity of at least 1.65 units/mg, they searched for specific recognition motifs by assaying KalbTGase with the NimbleGen peptide array technology. The turnover of the transamidation reaction between 1.4 million unique5-mer peptides and biotinylated amine donor N-(biotinyl)cadaverine used as a substitute for a Lysine substrate was quantified via fluorescence measurement of CyTM5–streptavidin binding. The experiments were performed on two arrays in parallel, and the sequences of the peptides with the highest turnovers were determined. The 9 best peptides were resynthesized and tested for KalbTGase activity in a stand-alone glutamate dehydrogenase (GLDH)-coupled assay. From these studies, it was possible to detect the best Glutamine 5-mer substrates: YRYRQ and RYRQR, with turnover rates of 3.52 ± 0.08 pmol of NADH/s and 3.60 ± 0.12 pmol of NADH/s, respectively.

A second round of maturation on the array was performed and the best substrate found was then resynthesized as biotinylated peptide to act as acyl donor for the discovery of optimized Lysine recognition motifs back on the 5-mer peptide array. Six of the best Lysine peptides from the array were resynthesized and tested in the GLDH-coupled assay, now using a peptide containing the optimized Glutamine recognition sequence YRYRQ as acyl donor. In this way, it was possible to detect also the best Lysine 5-mer substrates: RYESK, the sequence with the highest turnover (4.47 ± 0.16 pmol NADH/s) in the GLDH assay.

Further analyses have also demonstrated that KalbTGase shows poor or undetectable turnover with substrates recognized by conventional MTGase. Moreover, the top KalbTGase Glutamine substrates can be found in the midfield of the signal distribution on the array performed with MTGase, and, vice versa, the best-performing MTGase Glutamine substrates exhibit only signal close to background level on the KalbTGase array.

These results suggest that KalbTGase is a more selective enzyme than MTGase, which displays, instead, a broad substrate specificity for both acyl donor and alkyl amine groups.

Furthermore, the enzyme requires no additives, works well in standard buffers, such as Tris, MOPS, or PBS and is strongly inhibited by the addition of NH_4^+, the product of the transglutaminase reaction, making this enzyme very promising for the biotechnology industry, particularly for site-specific coupling applications.

2. Application of microbial Transglutaminase in the industrial fields

As mentioned in the previous paragraphs, until the discovery by Ando et al. in 1989 of the Ca^{2+} independent MTGase, an application of TGase enzyme in the industrial fields was not possible, due to the relatively small quantities obtained, the extensive separation and purification steps required, and the costs involved. However, after 1989, thanks to the rather easy purification of the enzyme extracted from *Streptomyces mobaraensis* via traditional fermentation, the production of MTGase has been established commercially.

Protein cross-linking catalyzed by MTGase has attracted the greatest interest, finding its application in food and industrial processes (Camposa et al., 2013). Due to its effects on the physical and chemical properties of proteins, MTGase has many biotechnological applications particularly in the food processing industry, in medicine, and in cosmetics (Martins et al., 2014). Although, it is associated predominantly to food industry, as a food additive (texturing agent), it is also applied in wool textiles and biopolymers (Carvajal et al. 2011). Moreover, the interest in these enzymes is also focused on several biological processes (blood clotting, wound healing, epidermal keratinization, curing membranes) and clinical applications such as neurodegenerative diseases and blood coagulation disorders, bone tissue healing processes and cell differentiation processes (Calmolezi Gaspar et al., 2015).

As regards the food processing industry, today, MTGase is mainly used in meat, fish, dairy, and baking industries. In the meat and fish industries, the main applications of MTGase are to alter mechanical properties of meat and as a bonding agent. Altering mechanical meat properties and meat bonding is used in production of sausages, improving texture and allowing utilization of lower quality meat. In dairy applications, MTGase modulates texture, structure, curd yield, and consistency of yogurts, ice cream, milk, and cheese. In baking, MTGase is used for improving flour properties such as elasticity and dough resilience, bread texture and volume, and pasta texture (Strop, 2014).

In research and biotechnology field MTGase is used for the construction of Protein-DNA conjugates, protein-polymer conjugates, full-length IgG conjugates, radioimmunoconjugates and antibody drug conjugates. This latter application has great therapeutic potentiality, above all on the treatment of cancer and has been one motivation to develop an enzymatic method for site-specific antibody drug conjugation using MTGase.

More in the details, in the case of protein-DNA conjugates, a successful strategy has been the employment of the approach that uses a synthetic nucleotide analogue, Z-QG-dUTP, that is incorporated into DNA via PCR reaction, resulting in multiple sites of attachment on DNA. Z-QG-labeled DNA is conjugated to multiple alkaline phosphatases and used in hybridization experiments, resulting in comparable detection sensitivity to digoxigenin labeled probes.

In the case of protein-polymers conjugates, instead, a successful strategy is represented by MTGase-based site-specific PEGylation of pharmaceutical proteins. In most cases, the number of reactive glutamines is typically low in comparison to the total number of surface-exposed glutamines and, in several cases, a single glutamine is identified resulting in site-specific conjugates. A good example of MTGase-based PEGylation is its application in human growth hormone where two glutamine residues are identified as major conjugation sites (Q40 and Q141) (Strop, 2014).

Even if the amount of lysine and glutamine residues available may be limited, especially on the surface of wool fibres, MTGase is industrially a very important enzyme in textile industry for wool fabrication, since it recovers wool damaged by chemicals and protease. In addition, it helps to incorporate amines, proteins, and antimicrobials to wool to bring desired properties in wool.

MTGase is also used in the treatment of silk to solve the problem of the poor elastic recovery. Both solo MTGase treatment and treatment with MTGase followed by hydrogen peroxide, protease and ultrasonic exhibit that MTGase can improve the crease resistance of silk fabric. Moreover, it enhances its tensile breaking strength or amends damage in the tensile breaking strength caused by pretreatments (Tesfaw and Assefa, 2014).

2.1 MTGase enzymatic effects exploiting by the food-processing industry

The major interest and the resulting widest fields of application of the MTGase is in the food-processing industry.

The functional properties of proteins are the physicochemical properties that affect their behavior in food systems, depending on the conditions of preparation, processing, storage and consumption, contributing to the quality and sensory characteristics of foods. Among these properties, MTGase acts in principle on solubility, gelation, emulsification and foam formation, water-holding capacity and viscosity.

More in the details, deamidation promoted by the action of MTGase can increase the solubility of proteins and thus their ability to stabilize emulsions and foams. Proteins reach in glutamine and asparagine can be deamidated into glutamic and aspartic acid respectively. The resulting deamidated proteins present a lower isoelectric point, which increases the protein solubility in the majority of slightly acid food systems. Cross-links activity instead enables the obtainment of highly elastic and irreversible gels in different substrates, even at relatively low protein concentrations (Calmolezi Gaspar et al., 2015). Studies on the effect of adding MTGase to meat systems showed as this enzyme drastically alters the structure of myosin heavy chain, with a significant reduction in the content of the α-helix structure, and an increase in the β-sheet, allowing for the formation of high molecular weight polymers, producing a significant increase in hardness, springiness and cohesiveness and resulting in strong gels with a compact and ordered structural conformation (Herrero et al., 2007).

In emulsions or foams, as mayonnaise, milk, creams and soups, ice creams, meringues, mousses, marshmallows and cakes, proteins are the major surface-active agents that aid the formation and stabilization of the dispersed phase. MTGase increases the emulsification capacity, reduces surface tension and enables augmented binding ability to the water. Moreover, the use of the enzyme leads to the formation of high molecular weight peptides due to the cross-links it catalyzes, these peptides are adsorbed on the surface of the oil droplets and promote electrostatic repulsion, preventing the approximation of these droplets, and thus, their flocculation, coalescence and phase separation, thereby increasing the stability of the emulsion. MTGase is also able to increase the foaming ability of proteins.

With regard to the water-holding capacity, both the deamidation reactions and the formation of cross-links catalyzed by MTGase directly influence this property in different protein substrates. In appropriate concentrations, MTGase yields stable gels with higher porosity that are able to immobilize water more efficiently, obtaining better textural properties such as bond strength, stiffness, cohesion, chewability and elasticity of protein gels.

After MTGase treatment in many food products, as yogurt and ice cream, viscosity increases as a function of the increase in gel strength established by crosslinking, and proportionally to the increase in enzyme content. This polymerization results in the formation of high molecular weight polymers that can reduce water mobility in the protein network, providing greater flow resistance and giving the product a suitable consistency (Calmolezi Gaspar et al., 2015; Santhi et al., 2017; Motoki and Kumazawa, 2000).

From this description, it results that the reactions promoted by this enzyme create radical changes in the proteins in food matrices, leading to improved texture and stability in terms of temperature, syneresis, emulsifying properties, gelation and increased water-binding capacity, without changing the pH, color, flavor or nutritional quality of food. They may even render it more nutritious, due to the possibility of adding essential amino acids (Calmolezi Gaspar et al., 2015). These characteristics, allied to the fact that the enzyme has been recognised by the scientific community as a safe substance for human ingestion, make MTGase so attractive for the food industry that nowadays this enzyme is used in meat, fish, dairy, and baking industries.

In particular, in meat industry, MTGase finds its main application in the production of restructured meat allowing the preparation of meat like beef (or pork) steaks, hamburgers or fish fillets from their smaller pieces. This procedure concerns the simultaneously use of MTGase and caseinate. Caseinate, when treated with MTGase, becomes viscous, and the viscous caseinate acts as a glue to hold different foodstuffs together. Moreover, even if MTGase treatment shows a synergistic effect, when combined with salt and phosphates, meat pieces can be also bound together by MTGase without salt (sodium chloride) and phosphates, resulting in more 'healthy' meat products. In addition, the use of crosslinked caseinate as a fat substitute enables low-fat meat production (Kuraishi et al., 1996; Weiss et al., 2010). In the fish products instead, seems that MTGase treatment improves and maintains the texture-quality of fish products, which strictly depends on freshness of raw materials (Motoki and Kumazawa , 2000; Tokay and Yerlikaya, 2017).

Many researchers have shown as milk casein is a very good substrate for the MTGase, demonstrating how after MTGase treatment casein forms heat-resistant gel. An example of MTGase use in dairy product is the MTGase employment in yogurt production. Actually, MTGase by improving the water-holding capacity of the gel, can solve problems of serum separation with a change in temperatures or physical impacts that affect yogurt. Moreover, MTGase reaction also makes it possible to produce dairy products, such as ice cream and cheese, with low-fat contents or a reduced content of non-fat-solid (Motoki and Seguro, 1998; Abd-Rabo et al., 2010).

Another important application of MTGase, above all for the Asian market, regards the soybean products. Tofu is prepared through coagulation of soybean proteins adding Ca^{2+}, Mg^{2+} and/or glucono-δ-lattone. However, it is very difficult to produce long-life tofu, because its smooth texture is easily destroyed by retort sterilization. The addition of MTGase solves the problem, enabling the maintenance of the smooth texture of retorted tofu for a long time (Motoki and Kumazawa, 2000).

However, MTGase finds its applications also on wheat products, some research found that MTGase treatment of noodles and pasta prevented deterioration in textures after cooking, and improved the strength of the product, even when low-grade flours were used. It was also suggested that the loaf volume of several breads may be increased or maintained by the addition of MTGase, when some ingredients were substituted or reduced during mixing dough (Motoki and Seguro , 1998; Seravalli et al., 2011).

MTGase is also used to produce edible coating for fresh cut products as fresh cut apples, potatoes and carrots, that were coated by a blended whey protein/pectin film, in presence of transglutaminase in order to extend their shelf-life. Several studies confirm that the whey protein/pectin/transglutaminase edible coating is effective to avoid fresh cut fruits and vegetables spoilage during ten days of storage, as demonstrated by reduction of weight loss, microbial growth prevention, antioxidant activity preservation and no change in fruit and vegetable hardness and chewiness (Rossi Marquez et al., 2017).

2.2 MTGase, functional food and allergy prevention

Recently, the interest on MTGase does not regards only its applications in the food industry just to solve manufacturing problems, but also its use to improve nutritional quality of food, and even render it more nutritious; thus, the interest nowadays is focused on the industrial application of MTGase to produce functional foods. Moreover, due to the great protein modifications that it is able to induce, the contemporary research is endeavoring to apply this enzyme in food allergy prevention, exploiting the possibility to elude the immunatary response masking the allergenic antigens by means of proper modifications catalyzed by mTGase.

As described in the previous paragraph, by treating meat or dairy products with MTGase it is possible to obtain foods with low-lipid and salt contents and to reduce the use of unhealthy gelling agents. However, by means of MTGase it is also possible to obtain fortified-food.

An example is the formulation of novel beef patties enriched with polyunsaturated ω-3 fatty acids and fiber. These petties show an optimal traditional texture as well as minimal effect on color and cooking loss and improved values of expressible water and were obtained by a pretreatment with 0.1 U/g of MTGase at 40°C for 17 minutes without the need of sodium caseinate addition (Martínez et al., 2011).

Another example is the MTGase employment for the probiotic microencapsulation to improve the survival of probiotics in simulated gastrointestinal conditions and yoghurt.

Chun Li and colleagues in 2016 demonstrated as the microencapsulation of probiotics by MTGase-treated soy proteins isolated can be a suitable alternative to polysaccharide gelation technologies. They prepared microencapsulation of *Lactobacillus rhamnosus* GG (LGG) by first crosslinking of soy protein isolate (SPI) using MTGase, followed by embedding the bacteria in cross-linked SPI, and then freeze-drying. The results showed the obtainment of a high microencapsulation yield (67,4%), no difference in water activity between free and microencapsulated LGG after freeze-drying and an improvement to 14.5 ± 0.5% of the survival of LGG in MTGase cross-linked SPI microcapsules during storage in yoghurt. Moreover, they demonstrated that, with this technique, the survival of microencapsulated LGG under simulated gastric juice (pH 2.5 and 3.6), intestinal juice (0.3% and 2% bile salt) and storage at 4°C were significantly higher than that of free cells.

MTGase is also used for the production of a material that promotes minerals absorption by the human body, exploiting its ability to deamidate casein. The casein is soluble in neutral and slightly acid conditions and can keep minerals solubilized in the intestine. Therefore, the resulting material promotes mineral absorption in the intestine and can be used in in mineral supplement formulations for adults, children and infants (Martins et al., 2014).

However, MTGase is also applied to induce strategically modifications in specific allergenic proteins.

From ‘90s, researchers started to develop method for reducing the allergenicity of some food proteins and/or peptides. One of the first discovery was related to the treatment of casein.

α_{s1}-Casein (23 kDa) treated with MTGase at 25°C for 20 h in water was less allergenic due to the production of cross-linked casein (approx. 90 kDa) (Zhu et al., 1995).

Recently Yuan F. and colleagues have demonstrated as MTGase-catalyzed glycosylation of β-Lactoglobulin (β-LG), recognized as the major milk allergen, has more potential, compared to glycation, for mitigating the allergenic potential of milk products.

Glycation is a non-enzymatic covalent interaction, which takes place following heat treatment, between the carbonyl groups of a reducing sugar and the amino groups of amino acids; the reaction products have the potential to modify the functional properties, digestibility, immunogenicity and allergenicity of food products.

Glycosylation, instead, is a conjugation reaction, mediated by an enzyme, performed at lower temperatures, that can site-specifically link the sugar. Yuri et al. in 2018 explored the changes in the conformational structure and potential allergenicity of glycated and MTGase mediated glycosylated β-Lactoglobulin with glucosamine. Both glycation and MTGase catalyzed glycosylation modified the linear and conformation structures of β-Lactoglobulin. However, the results showed that the glutamine residues of the immuno-activated peptides were modified after glycosylation. Subsequently, the allergenicity of the treated protein reduced substantially due to the alteration in the IgG/IgE binding epitopes. The most significant decrease in allergenicity was observed following the treatment of β-Lactoglobulin with MTGase and glucosamine at 37°C; suggesting the employment of this technique in hypoallergenic food processing (Yuan et al., 2018).

Preliminary studies demonstrate that MTGase could also be applied for shrimp detoxification, acting on tropomyosin, a myofibrillar protein recognized as the major allergen in shrimps (Yuan et al., 2017).

Also in this case, studies performed using MTGase-catalyzed glycosylation on tropomyosin shown that the reaction induced unfolding of the primary protein structure followed by loss of the secondary structure. These modifications can result in the reduction of IgG/IgE-binding capacity. Actually, western blotting and indirect ELISA with tropomyosin-specific polyclonal antibodies from rabbit and sera from patients allergic to shrimp demonstrated that antigenicity and potential allergenicity of tropomyosin decreased. It indicated that alterations in linear and conformational structures would cause epitopes destruction, which was responsible for potential allergenicity.

These results suggest that MTGase-catalyzed glycosylation has the potential to serve as a mild method for reducing the allergenicity of shrimp products (Yuan et al., 2017).

MTGase treatment is a valid procedure even for soybean and wheat flavors allergenicity reduction; in particular MTGase treatment reduce allergenicity of wheat flours by modifying epitopic structures or opposing steric hindrance around the epitopes (Watanabe et al., 1994).

Due to the relevance and the economic impact related to the allergenicity reduction by use of MTGase and the extent of this field, next paragraphs will be dedicated to the deeper analysis of one of the most debated examples of MTGase application: the celiac disease and the MTGase flours detoxification.

2.3 Microbial Transglutaminase and innovative treatment for Celiac Disease

In order to understand the results achieved by MTGase employment in the flour detoxification and their social and economic impacts, it is necessary to understand what the celiac disease is, who are the gluten free consumers and how this consume impacts on the economic aspect and, not less important, what are the risks and the benefits related to a gluten free diet. In the next paragraphs all these aspects will be briefly analyzed, in order to give to the reader all the instruments necessary to have a wider view of the problem and a better comprehension about the relevant impact related to the results achieved so far.

2.3.a The Celiac Disease and the gluten free diet

Celiac Disease is an autoimmune disease which leads, in case of gluten assumption, to many gut damages, to impairs nutrients absorption and to many symptoms as anemia, bloating, diarrhea, nausea and constipation (further explanations about celiac disease are available at: Sollid 2000; Di Sabatino and Corazza 2009).

Gluten proteins are the storage proteins in certain cereals as wheat, barley and rye. Proteins present in wheat gluten can be divided in four fractions on the basis of their solubility: the water-soluble albumins, the salt-soluble globulins, the prolamins that were soluble in 70% aqueous ethanol, and the glutelins that remained in the flour residue. Of these fractions, the glutelins (glutenins) and prolamins (gliadins) are the most widely studied proteins due to their contribution to the rheological characteristics of dough made from wheat flour (Ribeiro et al., 2013).

However, gliadins are the topic of lots of scientific studies, not only for their essential role for giving bread the ability to rise properly during banking but also, for their important involvement into the trigger of the celiac disease.

Gliadins, in fact, represents the toxic fractions in gluten and are a mixture of alcohol-soluble proteins, which are rich in glutamine and proline residues that even the healthy human intestine cannot fully digest (Hausch et al., 2011). As a result, intact gliadin peptides are left in the lumen, and some cross the intestinal barrier. These fragments are recognized by hTGase2, which deamidates them, modifying glutamine residues into glutamic acid residues. This transformation generates peptides ideally suited to interact with the HLA DQ2 or DQ8 molecules, whose expression is observed in celiac patients and, consequently, is strongly associated with the celiac disease (Sollid, 2000). Once bound to DQ2 or DQ8, the complexes peptide-HLA DQ2/DQ8 are expressed on the cellular surface of the APC and gliadin peptides are presented to the $CD4^+$ T-cells, triggering the inflammatory reaction (Guandalini and Assiri, 2014) in people affected by coeliac disease. Actually, once activated, the $CD4^+$ T cells produce high levels of pro-inflammatory cytokines, thus inducing a T-helper-cell–type-1 pattern dominated by interferon gamma (IFN-γ). The T-helper-cell type 1 response leads to the development of coeliac lesions: intraepithelial and lamina propria infiltration of inflammatory cells, crypt hyperplasia, and villous atrophy (Di Sabatino and Corazza, 2009)

Currently, adherence to a gluten-free diet is considered as the first line and indeed only therapy for celiac disease, which has been proven to relieve the symptoms in most cases and effectively prevent potential complications (Rashtak and Murray, 2012).

From this little explanation of what is gluten, why in some people it represents a risk and where is possible to find this molecule, it is easy to understand that in nature already exist lots of products that are gluten free such as: meat, fish, eggs, rice, milk, potatoes, legumes, fruit and vegetables. Also, most beverages are gluten-free, including fizzy drinks, juices and alcoholic beverages. Therefore, in a gluten-free diet is necessary to avoid: wheat, spelt, barley, rye, malt, Kamut®, flours, starch, bran, semolina and ethnic derivates as couscous; first courses, breakfast products, sweets and salted prepared with prohibited cereals; in brief everything that is prepared with flour prohibited and/or ingredients unsuitable.

From this summarized list published by AIC (Celiac Italian Association), it is possible to understand how is not so easy the elimination of gluten in the diet, also because gluten contaminations are present in the most varied products.

In fact, in addition to the products which for their nature contain gluten two others important problems for people that need to follow a gluten-free diet are the cross contamination (frequent in the home but also in the industrial production plants) and the use of gluten as additive. Actually, processed foods commonly contain gluten as an additive (as emulsifiers, thickeners, gelling agents, fillers, and coatings). Unexpected sources of gluten are, among others, processed meat, vegetarian meat substitutes, reconstituted seafood, stuffing, butter, seasonings, marinades, dressings, confectionary, candies, and ice cream (Biesiekierski, 2017).

Because of this, moved to the aim of satisfy the needs of the Celiac patients and gluten intolerant people, lots of food industries started to produce gluten-free products, and so to eliminate or reduce the gluten component of all those products that normally are prepared with wheat, in particular, with gluten.

The rise of the birth of gluten-free products needed to be combine to a proper legislation, so even if nowadays gluten-free diet doesn't mean no bread, no pasta, and no pizza anymore, all these products need specific labeling according the related legislation. In 2008 the *Codex Alimentarius Commission*, a commission set up in 1963 by FAO and OMS, established that a food can be labelled as gluten-free only if its percentage of gluten do not overcome 20 mg/Kg.

If many years ago the range of products without gluten was very limited, and related to foods with an uncertain taste and an anonymous packaging, nowadays, aided by the wide spread of the gluten-free lifestyle, the offer of gluten-free products is really large and heterogeneous, with a range of products that goes from the gluten-free pasta or bread to all the others bakery products and also gluten-free cakes, biscuits, desserts, ice-creams and snacks. The great variety of products is also destined to grow always more, if we consider that the rise in demand for allergen-free processed meats, condiments, sauces and even dairy alternatives is anticipated to remain a key driving factor for the industry, as underlined by Gluten-Free Products Market Size & Share, Industry Report, 2014-2025.

Even if the availability, the prices and the range of gluten-free products is not a big problem as the last years for who has a gluten-free diet, the real challenge for the industries producers of this kind of products is to preserve the organoleptic properties of this food, despite their gluten deprivation, and finding suitable alternatives for gluten. Gluten is the main structure-forming protein in flour and is responsible for the elastic characteristics of dough and contributes to the desired appearance and crumb structure of many products especially the baked ones (Gallagher et al., 2004). Glutenins and gliadins, in fact, play a key role in baking quality characteristics, being responsible for water absorption capacity, cohesivity, viscosity, and elasticity of dough (Wieser, 2007). Hence, gluten removal results in major problems especially for bakers in terms of quality. So, one of the main challenges during development of gluten-free product is to ensure that the product has desired texture as well as mouth feel as the gluten-containing product (Jnawali et al., 2016).

Dry, rough and crumbly texture of the gluten-free pasta and bread still represent technical difficulties, some studies reported as gluten-free bread has a high staling tendency as compared to the gluten-containing one (Alborn et al., 2005) or that bread dough without gluten can only retain gas if another gel like substance replaces gluten (Rotsch et al., 1954).

In conclusion even if great strides have been made in the field of the gluten-free industrial foods, many researches and investments are ongoing in order to have a constant innovation, to improve production techniques and quality of these products and to find a proper substitute of gluten.

2.3.b The gluten free diet, its customers and its economic impact

The gluten-free diet was born to respond to the needs of the patients affected by celiac disease.

This kind of disease has registered a great spread during these last years, also due to the development of new diagnostics methods that increased the detectability of the pathology in people. Nowadays 1% of the European population is affected by this pathology (data AOECS) and in the USA the percentage is almost the same (CBS News reported). Actually, is possible to say that celiac disease occurs in about 1% of the population worldwide, although most people with the condition are undiagnosed (Lebwohl et al., 2015).

However, the last published data about the gluten-free market and trade show that the global gluten-free retail market has grown from $1.7bn in 2012 to $3.5bn in 2016 and is forecast to grow to $4.7bn in2020, according to Euromonitor, the consumer data group (Financial Time font).

This high request cannot be explained by the assumption that the reason of all this business volume is due to a client portfolio composed only of celiac patients. Actually, the gluten free-products are bought not only by people affected by coeliac disease but also by the so called “gluten sensitive”, it means those people who are intolerant to gluten, without having celiac disease or a wheat allergy (NCGS). These individuals may have similar gastrointestinal symptoms as the celiac patients, but no damage to the intestinal tract nor will they develop complications associated with Celiac Disease. However, even if the number of the people affected by NCGS is growing because of their more efficient identification, and that this number is estimated to be greater than the number of Celiac patients, it is still not possible to define the accurate percentage of this population.

Another important little slice of gluten-free products consumers is composed by the individuals affected by Dermatitis Herpetiformis, Gluten Ataxia or Wheat Allergy.

Dermatitis Herpetiformis (DH) or Duhring-Brocq disease is a chronic bullous disease characterized by intense itching and burning sensation in the erythematous papules and urticarial plaques, grouped vesicles with centrifuge growth, and tense blisters. It is an IgA-mediated cutaneous disease, in fact at the top of the dermal papilla of both affected and healthy skin it is possible to find in a granular pattern immunoglobulin A deposits. The same protein IgA1 with J chain is found in the small intestinal mucosa in patients with adult celiac disease, suggesting a strong association with DH. Moreover, specific antibodies such as antiendomysium, antireticulina, antigliadin and the epidermal and tissue transglutaminase subtypes, are common to both conditions, Celiac Disease and DH. Because of this, the chosen treatment is dapsone and a gluten-free diet (Berti Rocha Mendes et al., 2013).

A gluten-free diet is also necessary in people affected by Gluten Ataxia, an immune-mediated disease triggered by the ingestion of gluten, in genetically susceptible individuals (Hadjivassiliou et al., 2008). This pathology affects cerebellum and therefore causes problems with muscle control and voluntary muscle movement.

Differently from the celiac disease, wheat allergy is a food allergy. It is due to an undesirable response of the immune system against gluten or some other protein found in wheat as a disease-causing agent.

Generally, food allergic reactions to wheat can give way to an array of clinical manifestations that can range from immediate to delayed, and their strictness can vary from mild to life-threatening. Typical immediate symptoms include erythema, pruritus, eczema, gastrointestinal reactions, oropharyngeal symptoms, urticaria, angioedema, AD, rhinitis, asthma, and anaphylaxis (Pasha et al., 2013; Heffler et al., 2011).

Even if there is a big percentage of people that for different pathology must remove gluten from their diet gluten, as mentioned before, these categories cannot be responsible of a so exorbitant increasing of the gluten-free market.

Actually, those who really drive the sales of gluten-free is the part of consumers, in constant increase, who choose gluten-free foods for a personal food style, regardless of intolerance to this protein. An American study has underlined as about 1.8 million Americans are affected by celiac disease, but also that on the flip side, about 1.6 million people in the U.S. are on a gluten-free diet even though they haven't been diagnosed with celiac disease (Rubio-Tapia et al., 2012). This trend has not spare the Europe, the Financial Time journal with an article of the 30 April

2017 on the basis of the European forecast and Euromonitor data, explains that "gluten-free foods have been consumed for years by people suffering from coeliac disease [...] however, demand has now widened beyond medical need as food intolerances have become more widely accepted and more people opt for "free-from" and "clean-label" products — a category that encompasses organic and GM-free foods — as a lifestyle choice".

Nowadays the use of gluten-free products has become a real trade; Dr. Alessio Fasano, director of the Center for Celiac Research and Treatment at Massachusetts General Hospital, who has led world-renowned research on gluten, in an interview for the CNN, said that "the gluten-free diet is the most popular diet in Hollywood". This suggest how much rife and "contagious" this trade is. Moreover, to fuel this trade, there are many celebrities who avoid gluten for non-medical reason and release public declaration about this topic exhorting to follow their example, triggering a "viral effect" on the spread of this life-style above all among the youngest generations.

In conclusion the gluten-free diet is something that involves more people than we would expect and is a topic of huge impact for several aspects whether economic or healthy.

2.3.c Risks and benefits of a gluten free diet

As widely discussed in the previous paragraphs, in almost the 1% of the worldwide population affected by celiac disease and in the even higher percentage of people who presents a non-celiac gluten sensitivity, it is necessary the removal of gluten from the diet in order to avoid the trigger of the autoimmune response that recognizes erroneously this component as toxic and dangerous for the organism.

Several clinical trials, in fact, have shown that individuals with mild enteropathy celiac disease and positive serum EmA suffer from a gluten-dependent condition. In these studies, small-bowel mucosal structure and inflammation, serology, and clinical symptoms were investigated both by ongoing gluten intake and gluten withdrawal. In conclusion, their findings confirm that when gluten consumption was continued, progression of the mucosal damage was evident, and the abnormal serology and clinical symptoms persisted, in contrast, the beneficial effect of a gluten free diet was indisputable and in fact similar both in mild enteropathy and in overt celiac disease (Kurppa et al., 2009).

Nevertheless, there are still some limitation related to the use of a gluten-free diet for these patients and more and more studies are ongoing in order to find a non-dietary treatment or novel strategy to allow the reintroduction of gluten in the diet. Actually, although a gluten-free diet is effective when implemented correctly, many patients find it unsatisfactory; this kind of diet can in fact be expensive and socially isolating; moreover, the potential hidden gluten used as additive in unexpected sources such as meat, fish, milk products, drugs or sauces is a legitimate source of anxiety for patients and represent a real problem (Lebwohl et al., 2015; Moreno et al., 2014).

Another very important aspect that is remarkable is that the avoidance of gluten-containing cereals includes the exclusion of major food sources which are one of the major protein sources in the diet. The protein content in wheat is 10%–12% and eliminating wheat from diet completely for celiac patient would mean the exclusion of a very good protein source (Jnawali et al., 2016).

To make matters worse is also the unhealthy composition of many gluten-free products. Actually, the removal of gluten presents major problems for bakers, and currently, many gluten-free products available on the market are of low quality exhibiting poor mouthfeel and flavor (Gallagher et al., 2004).

In recent years the formulation of gluten-free cereal-based products has been the center of many researches, involving a diverse approach which has included the use of starches, dairy products, gums and hydrocolloids, other non-gluten proteins, prebiotics and combinations thereof, as alternatives to gluten, to improve the structure, mouthfeel, acceptability and shelf-life of gluten-free bakery products (Gallagher et al., 2004).

However, gluten-free cereal foods are frequently rich in carbohydrates and fats and they are made using refined gluten-free flour or starch not enriched or fortified (Thompson, 1999). As a result, many gluten-free cereal foods do not contain the same levels of B-vitamins, iron and fiber as their gluten-containing counterparts (Thompson, 2000). Further studies have also demonstrated as there is a significantly lower contribution of folate from bread consumed by coeliac patients than from that consumed by the general population, whilst they ate bread to a similar extent, moreover, these studies have also underline as, in adult coeliac patients on a strict gluten-free diet for years, it is possible to observe the raise of total plasma homocysteine level (Hallert et al., 2002). This last aspect not only is indicative of a poor vitamin status, but it may also imply an independent increased risk for cardiovascular disease in the same range as hypercholesterolemia and hypertension (Nygård et al., 1999).

If it is true that research and development for products gluten-free with a better quality is very fervent, is also true that in all these years the situation is almost unchanged. A recent study made by the researchers of the Spanish Institute "Investigación Sanitaria La Fe" brought to the attention of the scientific community during the congress of the European Society for Pediatric

Gastroenterology Hepatology and Nutrition (ESPGHAN) that took place on 11 May 2017, has pointed the accent on the quality of many gluten free products and on the dangerous risks that people with non-celiac disease ran making of the "gluten-free" a life style. The researchers have in fact analyzed 654 gluten-free products, which were compared with 655 gluten-containing products, and they found that: gluten-free breads have significantly higher content of lipids and saturated fatty acids, gluten-free pasta has significantly lower content of sugar and protein, and that gluten-free biscuits have significantly lower content of proteins and significantly higher content of lipids. A terrible shock for all those people who see in the gluten-free diet a novel life-style to reach a healthy body and a healthy mind and who trust in gluten-free as a good solution to lose weight. Now more than ever is necessary, as also said by ESPGHAN expert and lead researcher, Dr. Joaquim Calvo Lerma, "that foods marketed as substitutes are reformulated to ensure that they truly do have similar nutritional values, especially for children, as a well-balanced diet is essential to healthy growth and development" and, as underlined by the researchers during the ESPGHAN annual meeting, the nutritional labelling should be clearer, in both gluten-free and gluten containing products, so that the consumers could make a more aware purchase.

Another recent important study published on the British Medical Journal made at the Brigham and Women's Hospital di Boston has, instead, underline as the gluten restricted diet with a goal of reducing coronary heart disease risk should not be promoted and that there is no significantly association with risk of coronary heart disease and the consumption of foods containing gluten. To examine the association of long term intake of gluten with the development of incident coronary heart disease, these researchers performed a cohort study on 64714 women in the Nurses' Health Study (a prospective cohort of 121 700 female nurses from 11 US states who were enrolled in 1976) and on 45303 men in the Health Professionals Follow-up Study (a prospective cohort of 51529 male health professionals from all 50 states who were enrolled in 1986) without a history of coronary heart disease who completed a 131 item semiquantitative food frequency questionnaire in 1986 that was updated every four years through 2010. The cohort participants were divided into fifths of estimated gluten consumption, according to energy adjusted grams of gluten per day, and at the end of the study the scientists found no differences between the group that ate the most gluten and the group that ate the last. In conclusion there is no significant association between estimated gluten intake and the risk of subsequent overall coronary heart disease, non-fatal myocardial infarction, and fatal myocardial infarction (Lebwohlet et al., 2017). On the contrary, when they adjusted for refined grain intake, leaving the variance of gluten intake correlating with whole grain intake, they noted a significant inverse relation between estimated gluten intake and coronary heart disease, probably related to the fact that whole grain intake has been found to be inversely associated with coronary heart disease risk and cardiovascular mortality (Aune et al., 2016).

For this reason, the promotion of gluten-free diets for the purpose of coronary heart disease prevention among asymptomatic people without celiac disease should not be recommended, also because going gluten-free tends to reduce the number of whole grains introduced in the diet. Moreover, the association between the reduction of the risk of myocardial infarction and death from cardiovascular disease in the celiac patients, who furthermore are less likely to have classic cardiac risk factors such as smoking and dyslipidemia, and the beneficial effect of a gluten-free diet is itself still controversial (Lebwohlet al., 2017).

Another important topic is the relation between gluten and diabetes. Celiac disease is routinely perceived to be more common in children with Type 1 diabetes (Sud et al., 2010), but recent genetic and epidemiologic trends as well as screening data suggest that Celiac disease is highly prevalent in adults with Type 1 diabetes as well (De Melo et al., 2015). The main reason of this correlation probably is due to the fact that Type 1 diabetes and celiac disease are autoimmune diseases with shared genetic origins (Smyth et al., 2008). Lots of studies are oriented to analyze the positive effects of the gluten-free diet on the patients affected by both Celiac disease and Type 1 diabetes.

A study published in 2016 on The Journal of Pediatrics showed as youth with Type 1 diabetes (T1D) and celiac disease reported similar generic and diabetes-specific Quality of Life (QoL) to T1D only. Gluten-free nonadherent vs adherent youth reported lower diabetes-specific Quality of Life (QoL) (mean score 58 vs 75, P = .003) and lower general well-being (57 vs 76, P = .02), as did their parents (50 vs 72, P = .006), and hemoglobin A1c was higher (9.6% vs 8.0%, P = .02), in summary, youth with T1D and celiac disease who do not adhere to the gluten-free diet have lower QoL and worse glycemic control (Pham-Short et al., 2016).

Moreover, in a study of adolescent patients with T1D and celiac disease who were treated with gluten-free diet for one year compared to patients with T1D, it was demonstrated as the adolescents who followed a straight gluten-free diet there were the lower albumin-to-creatinine ratios and the decreased plasma advanced glycation end products, which were known to be associated with glycemic-related end-organ renal tissue damage (Malalasekera et al., 2009).

However, if on the one hand the role of a gluten-free diet in symptomatic patients is supported by medical care guidelines, on the other hand it is unclear whether asymptomatic patients identified by screening alone will experience meaningful short- or long-term clinical improvements after treatment with gluten-free diet (De Melo et al., 2015), and further studies are necessary to understand the real impact of the gluten-free diet in people affected by T1D only.

A different story is instead the use of a gluten-free diet to prevent the risk of Type2 diabetes in people not affected by celiac disease. From a deeper analysis of the previously mentioned cohort study made at the Brigham and Women's Hospital di Boston, the researchers have also demonstrated an inverse association between gluten intake and Type2 diabetes or excess weight gain, and they also made the conclusion that limiting gluten from diet is thus unlikely to facilitate Type2 diabetes prevention and may lead to reduced consumption of cereal fiber or whole grains that help reduce diabetes risk (Zong et al., 2017).

In conclusion gluten-free diet is considerable until now the only adequate therapy for celiac disease patients and people who present a gluten intolerance, this diet lets them to have a better quality of life, a reduction of common cardiovascular risk factors, including weight, alterations in lipid profiles and restoration of the normal nutrient assimilation by means of the regression of the mucosal damage.

However, patients should make a more consciousness selection of the gluten-free foods, trying to avoid those products that are reach on fat and carbohydrate as a consequence of the gluten substitution.

People without celiac disease or any non-celiac gluten sensitivity should avoid gluten-free products because these products can raise the risk of cardiovascular disease, increased weight, and pathology as Type2 diabetes.

Briefly, to improve health, reduce the risk of diseases, have a keen mind and a toned body the only strategy is not to follow expensive and straight diet but do physical activity and have a good and healthy relationship to food, avoiding the abuse of fatty and caloric foods and preferring those foods which are rich in fiber and vitamins. Gluten-free diet is not a trend: is a medical treatment, and as such must be considered and followed.

2.3.d Wheat flour detoxification: its results and its limits

The scientific attempts to produce hypoallergenic flour date back to 1990 when Watanabe and colleagues started to perform tests using controlled enzymatic treatment of wheat proteins for producing hypoallergenic flour. From their results, it was possible to say that among all the treatments performed, the collagenase- and the transglutaminase-treated products were preferable as material for food processing and that the enzyme-treated soft flour can thus retain a certain degree of dough properties useful for food processing and can be used as a functional food material for wheat-associated allergic patients.

Of course, since 1990 much progress has been made, in fact, nowadays several studies have demonstrated how the MTGase can be useful to detoxify flour and block the trigger of the celiac disease in some patients affected by this pathology.

Actually, an innovative gluten detoxification called Gluten Friendly ^{TM} has been developed. By this method, structural modifications are induced in gluten, abolishing its antigenic capacity and reducing in vitro immunogenicity of the most common epitopes involved in celiac disease, without compromising nutritional and technological properties (Costabile et al., 2017).

The Gluten Friendly ^{TM} detoxification is an enzyme strategy involving wheat flour, MTGase and an appropriate amine donor, represented in this case by lysine methyl ester (K-CH_3), that allow to preserve the integrity of the protein structures *via* the wheat flour transamidation. This technique, actually, exploits the MTGase ability to form cross-links to modify glutamine residues present in gluten. More in the details, this technique modifies glutamine residue by cross-link formations between these residues and the amine donors K-CH_3. In this way the resulting modified glutamines are not able to be recognized by the hTGase2 which, in turns, can't deamidate them; immunogenic peptides composed by glutamine residues not deamidated are not recognized by the HLA DQ2 or DQ8 molecules, thus the trigger of the inflammatory response is blocked (Gianfrani et al., 2007).

The authors extracted gliadin samples from various flour preparation, after enzymatically digestion and treatment with MTGase, they incubated them with iTCLs for the assessment of the IFN-γ production. iTCLs are gliadin-specific intestinal T-cell lines generated from biopsy specimens from 12 adult patients with celiac disease and challenged in vitro with different antigen preparations. Notably, gliadin from flour treated with MTGase and K-CH_3 was ineffective in inducing IFN-γ expression, as reflected in the values that were statistically indistinguishable from the negative control. To verify the effectiveness of the proposed treatment on other cytokines that can be induced by gliadin, in vitro levels of IL-2, IL-4, and IL-10 were also tested. The results shown that when iTCLs were challenged with gliadin from flour treated with MTGase and K-CH_3, an almost complete block of production for all analyzed cytokines was observed in all examined iTCLs. Moreover, MTGase was also tested on the α-gliadin 33mer peptide, an immunogenic gliadin peptide containing 3 distinct T-cell epitopes, demonstrating that residues transamidated by MTGase were the same ones deamidated by hTGase2 (Q10, Q17, and Q24). Actually, MTGase seems to be unable to catalyze the deamidation reaction; in fact, the gliadin peptide 56-68, which contains an immunodominant epitope, was found to be unmodified when treated with MTGase in H_2O (MTGase modified this peptide only by Q65 transamidation). (Gianfrani et al., 2007).

In 2016 Moscaritolo, Rossi et al. started a pilot scale production of transamidated wheat semolina. In this study they adopt a two-step transamidation protocol.

According to this new protocol semolina was suspended in two volumes of water containing 8 U/g MTGase and 20 mM lysine ethyl ester (K-C_2H_5), and incubate in a reactor plant, having a nominal capacity of 16 liters, for two times, the first and the second step respectively.

In the first step the incubation was performed for 2 h at 30°C and the suspension was recovered by centrifugation (1000xg, 10 min). After an extensive washing of the reactor, a second enzyme step was conducted for 3 h at 30°C with fresh enzyme and K-C_2H_5 at the same concentrations. The suspension was finally centrifuged (15,000xg, 10 min) and dough recovered.

The effectiveness of the enzymatic reaction was tested by means of consolidated biochemical and immunological methods on isolated prolamins, finding that the production of isopeptide bonds from the catalytic activity of MTGase dramatically decreased the gliadin yield to 29.3% ±1.9% and 5.9% ±0.3% after the second step, and that the glutenins yield was fairly affected after the first enzyme step (86.6% ±1.6%), and it decreasing to 11.6%±0.1% after the second step. Moreover, using DQ8 transgenic mice as a model of gluten sensitivity, the authors observed a dramatic reduction in IFN-γ production in spleen cells challenged in vitro with the residual insoluble gliadin from transamidated semolina.

The pilot-scale study ended with the technological properties assay of treated wheat semolina by manufacturing classical pasta (spaghetti), which demonstrated that the manufactured spaghetti showed only minor changes in its features before and after cooking.

From their results is possible to say that the two-step transamidation reaction modified the immunogenic epitopes of gliadins also on a pilot-scale level without influencing the main technological properties of semolina. (Moscariolo et al., 2016).

To corroborate MTGase employment in the flour detoxification in addition to the results obtained there is also the commercially availability of this enzyme, that is commonly used as a dough improver that adds stability and elasticity to the dough. Additionally, bread volume and crumb texture are positively influenced by the addition of MTGase, especially for flours with low gluten content and poor baking performance (Caputo et al., 2010). Moreover, the occurrence of the isopeptide linkage in protein has been demonstrated that does not impair the digestibility of the gliadin, highlighting the safety of the proposed treatment (Gianfrani et al., 2007).

However, as most new strategies, this approach still presents some limitations that need to be investigate more in the details in order to improve its application.

Lombardi and colleagues in a study on the modifications induced by MTGase transamidation on wheat flour analyzed the effects on intestinal specimens, derived from celiac patients, of the soluble protein fraction (spf) and K-gliadins fractions obtained after extensive transamidation of wheat flour. Even if in their previous studies performed using DQ8 mice as model of gluten sensitivity it was observed a dramatic reduction of IFN-γ production in gliadin-specific spleen cells challenged with spf and K-gliadins in vitro, in human specimens those results haven't been replicated. Actually, the spf inhibitory activity, despite the one of the K-gliadins, was not confirmed in all the sample tested (Lombardi et al., 2013). Moreover, the specimens tested are only four, thus the sampling used is not composed of a number large enough to be considered statistically robust.

In the same period, other studies demonstrated that gluten peptide treated with MTGase but without lysine applied to cultured intestinal biopsies from celiac disease patients, induced a 15-fold increase in INFγ release, and 2.5- and 2.1-fold increases in medium hTGase2 antibody levels and endomysial antibody positivity, respectively. Addition of lysine to the enzymatic modification of gluten normalized interferon γ, antibodies, transglutaminase activity and immunohistochemical expression of transglutaminase type 2 (Elli et al., 2012).

Most recently, a study performed in 2016 by Matthias T. and colleagues instead report as the MTGase used as industrial food additive may mimics hTGase2 and be immunogenic in celiac patients. In this research the serological activity of MTGase, hTGase2, gliadin complexed MTGase (MTGase neo-epitope) and gliadin complexed hTGase2 (hTGase2 neo-epitope) were studied in 95 pediatric celiac patients, 99 normal children, 79 normal adults and 45 children with nonspecific abdominal pain. Sera were tested by ELISAs, detecting IgA, IgG or both IgA and IgG. MTGase neo-epitope were prepared by dilution of MTGase stock solution to a final concentration of 0.1 U/ml in a suitable reaction buffer; gliadin peptides were added to react with the reaction mixture overnight at room temperature. From their results it is highlighted that comparing pediatric celiac disease patients with the 2 normal groups: MTGase-neo IgA and IgG antibody activities exceed significantly the comparable MTGase ones and both MTGase-neo and hTGase2-neo levels were significantly higher than the single antigens' titers. Moreover, it seems that the IgG isotype against MTGase-neo are much higher compared to the IgA isotypes and that the antiMTGase neo-epitope antibodies levels positively correlate to the degree of the intestinal injury in celiac disease. However, the authors underline that the immunogenicity of those antibodies is operative only in celiac patients and not in comparable, non-celiac, symptomatic children and not in the pediatric and adult control groups (Matthias et al., 2016). This study is very important because shows, for the first time, that MTGase stimulates the human immune system and induces specific antibodies. Actually, despite declarations of the safety of MTGase for industrial use, direct evidence for immunogenicity of the enzyme is lacking; thus, studies like this are necessary to understand the limitations of this new technique in order to try to overcome those limits and make these applications safer and healthier.

AIM OF THE PROJECT

3. MTGase a wide topic for a wide scientific research: the aim of this project

As described in the first two chapters TGase are a class of enzyme widespread in eukaryotic and prokaryotic organisms and has been found in various tissues of animal and plant origin. Due to its facility of expression and purification, via traditional fermentation, despite the fervent scientific research on novel sources of TGase the only enzyme widely used for industrial applications is the MTGase, i.e. the TGase extracted from *Streptomyces mobaraensis* discovered in 1989 by Ando and colleagues. Therefore in the last thirty years many things changed: the industrial applications of MTGase have increased, covering more and more fields, from the food-processing industry to the biotechnology, from the pharmaceutical field to the allergy preventions and nutraceutical treatment; the scientific research about mTGase is now more active than ever, starting from the research of novel sources, or from the improvement of MTGase production and/or expression ending to novel applications.

Evidence of this are easy to find by a simple bibliographic search that shows as the number of scientific reports and studies on this topic are constantly increasing: attempts to obtain MTGase from other microorganisms by conventional fermentation or by means of genetic modification using host microorganisms are in progress, as are in progress studies on MTGase safety and ability to modify more and more allergenic substrates or proteins of industrial interest.

All this arises from the attempt to meet specific needs. Actually, nowadays, for an industrial utilization, a more efficient system for MTG production would have to be implemented, and variations of MTG have been considered a prerequisite to give preferable properties to foods (Dube et al., 2007); moreover the wider application fields (food processing industry, medicine, cosmetics, biotechnology) are increasing the demand for an inexpensive, efficient, and safe source of recombinant enzymes (Martins et al., 2014).

The present work proposes an analysis of the wide microbial transglutaminase world, developing the first classification of the mTGase based on their sequence features and their specific predicted secondary structures, representing a big innovation from this point of view. In fact, despite the wide research on this topic and the large distribution in nature of these enzymes, so far, they lack of a proper classification, especially for what concerns the microbial enzymes, that among all TGases are the most studied because of their easier industrial application.

Moreover, aim of the project is the functional and structural characterization of novel mTGases that could become an alternative to that in use for the food industry, the allergy prevention or any other possible application arising from the novel properties discovered from these novel enzymes. In order to do that, both computational and experimental techniques have been used, involving sequence analysis, comparative studies, building of phylogenetic trees and homology models, molecular dynamic simulations, expression, purification and experimental assays of specific proteins.

MATERIALS AND METHODS

4. Sequence and structural analyses

The first part of this project is based on sequence analyses and comparative studies, starting from a collection of all the microbial protein sequences that may have a transglutaminase function. To do that several specific databases have been searched. Moreover, databases employment was also important in order to perform allergenic epitopes screening.

The collected sequences have been analyzed on the base of their sequence features and have been clustered by a reiterative procedure involving phylogenetic tree constructions. The protein sequences forming the resulting groups were analyzed by motif researches and among them some sequences by specific criteria have been selected to be modelled. Models obtained were assessed and the secondary structures obtained were analyzed in order to find specific features for each group. From the analysis of models obtained, some sequences have been selected for experimental activity assays.

4.1 Databases searches

Databases searches have been performed in order to identify amino acid sequences of potential transglutaminases in microorganisms. NCBI databases, EBI databases, UniProt and specific microbial databases of protein sequences, as MicroScope and Microbes online, have been searched by querying the sequence of the MTGase and the sequence of the hTGase2, as reference to human form to evaluate the possible interference of MTGase activity with human health. The usage of different databases, although with some redundancy, offers the widest coverage of knowledge on microbial protein and genome sequences.

UniProt: (https://www.uniprot.org/) Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data. The UniProt databases are the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR). EMBL-EBI and SIB together produce Swiss-Prot, manually annotated and reviewed, and TrEMBL (Translated EMBL Nucleotide Sequence Data Library), automatically annotated and not reviewed, while PIR produce the Protein Sequence Database (PIR-PSD) (The UniProt Consortium, 2017).

NCBI protein database: is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB (NCBI Resource Coordinators, 2016). In this database the search was performed by a screening against the whole database, when the query was the MTGase, just the Bacteria section when the query was the hTGase2. (https://blast.ncbi.nlm.nih.gov/Blast.cgi)

EBI protein database: also in this database, it is possible to find records from different sources as UniProt, Enzyme Portal, ChEMBL, MEROPS, IPD, PDB, Patent Protein Sequences, and others (Cook et al., 2018). However, for this analysis the searches have been performed selecting The UniProt Knowledgebase, which includes UniProtKB/Swiss-Prot and UniProtKB/TrEMBL databases, in the case of the MTGase as query. When the query was represented by the hTGase2, UniProtKB bacteria taxonomic subsets was instead selected. (https://www.ebi.ac.uk/Tools/sss/ncbiblast/)

Despite NCBI, EBI databases and UniProt represents the most used resources, to obtain the wider knowledge possible about the hypothetic microbial TGase sequences that exist, results obtained from more specific database as MicroScope and MicrobsOnline have been integrated with the ones already obtained.

MicroScope: (http://www.genoscope.cns.fr/agc/microscope/home/index.php) it is a powerful database where for each microbial organism several data as complete sequences, non-coding DNA, coding sequences (nucleic or proteic), annotated data on genomic objects, are stored in a specific database named PkGDB:Prokaryotic Genome DataBase (Vallenet et al., 2017). In this database the search, for both the queries, was performed selecting all the organisms whose proteome was present in the database.

MicrobesOnline: (http://www.microbesonline.org/) it is another specific database in which are present over 1000 complete genomes of bacteria, archaea and fungi and thousands of expression microarrays from diverse organisms (Dehal et al., 2010). The deriving proteome from each of the organism present is collected in the VIMSS database (7.278.452 sequences); the one exactly chosen to perform the researches required from this project.

In order to collect all the microbial protein sequences, annotated as having a TGase core domain or a hypothetical TGase core domain, necessary to perform the clustering procedure that will described in the next paragraphs, PFAM database was used.

PFAM:(https://pfam.xfam.org/) it is a database composed of a large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs), it includes annotations and additional family informations from a range of different sources. The PFAM database, called Pfamseq is based on UniProt Reference Proteomes (Finn et al., 2014).

In addition to these databases, the data bank PDB was used to collect the protein structures necessary to perform the homology modelling strategy, on the sequence selected, and all the already known TGase structures, to compare their structure to the models obtained. Moreover, COMPARE and IEDB.org databases were used to research the epitopes that could react with specific proteins of interest.

PDB: (https://www.rcsb.org/) Protein Data Bank established in 1971 is a public repository for three-dimensional structure data of biological macromolecules, obtained predominantly by X-ray crystallography and NMR (Berman et al., 2003). The information stored in the PDB archive is a file containing the atomic coordinates for each protein present (147073 proteins present until December 2018).

COMPARE: COMprehensive Protein Allergen REsource is a database of allergens, that presents an exhaustive listing of clinically relevant and peer-reviewed protein allergens with citation support and species identification. Moreover, COMPARE reviewers make available the descriptions of the allergens and of their amino acid sequences.
(http://comparedatabase.org/).

IEDB.org: The IEDB (https://www.iedb.org/) is a resource funded by a contract from the National Institute of Allergy and Infectious Diseases. It offers easy searching of experimental data characterizing antibody and T cell epitopes studied in humans, non-human primates, and other animal species. In this database epitopes involved in infectious disease, allergy, autoimmunity, and transplant are included (Vita et al., 2018).

4.2 Sequence alignments

All databases have been searched by BLAST tool, being available at all web sites hosting the search databases.

BLAST: The acronym stands for The Basic Local Alignment Search Tool. The program finds regions of local similarity between sequences and compares nucleotide (nBLAST) or protein sequences (pBLAST) to those present in a selected target database, calculating the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships about the structure and the function of the sequences used as query as well as help identify members of gene families (Altschul et al., 1990).

The alignments performed by pBLAST tool were done using parameters differing from default settings. When necessary, the research was restricted just to the microbial organisms selecting, in the search set, under the heading organism, the taxid 2 corresponding to the Bacteria tax. The max target sequences parameter was increased from 100 to 500 in the general setting, in order to collect all the sequences stored in the database which could be homologues to the query, and the word size was changed from 6 to 3 or 2 depending on the length of the sequences chosen as query sequence. The searches with hTGase2 were performed using either the whole protein or each domain from which it is composed, to find also protein sequences that could be similar to single domains of hTGase.

The sequences obtained from the alignment were analyzed in terms of sequence length, presence or absence of the catalytic residues, number of identities and positives (i.e., pairs of amino acids with similar properties), gap presence, E-value, the kind and the function of the source organism, and its habitat and availability.

Multiple alignments were performed by Clustal Omega tool and MUSCLE tool using the default parameters.

Clustal Omega: (https://www.ebi.ac.uk/Tools/msa/clustalo/) it is a multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between sequences (Sievers et al., 2011).

MUSCLE: (https://www.ebi.ac.uk/Tools/msa/muscle/) MUltiple Sequence Comparison by Log- Expectation. MUSCLE is a multiple sequence alignment program claimed to achieve both better average accuracy and better speed than T-Coffee multiple alignment tool (Edgar 2004).

The results obtained from the two tools have been compared in order to find hypothetical key residues and similar sequences from the point of view of their hypothetical function.

During the modelling procedure all the secondary structure alignments were performing manually, starting from the T-Coffee multiple alignment results, because starting from that one, better results have been obtained.

T-Coffee: (http://tcoffee.crg.cat/apps/tcoffee/do:expresso) it is a multiple sequence alignment package able to combine sequence information with protein structural information (Expresso). With T-Coffee a data set of all pair-wise alignments between the sequences is pre-processed, providing a library of alignment information that can be used to guide the progressive alignment. Intermediates alignments are then based not only on the sequences to be aligned next but also on how all the sequences align with each other (Notredame et al, 2000).

4.3 Sequence clustering by phylogenetic tree constructions

In order to make an accurate selection of the protein, which better fit for the purpose and to have a wider knowledge of the different features present among the various mTGases, a clustering approach by means of phylogenetic trees construction was employed.

4.3.a Workflow of the clustering procedure

At first, all the protein sequences annotated as hypothetical transglutaminase in Pfam database (≈8000) were divided in different groups according to the phylum belongs in. The groups at the beginning of this selection works were Actinobacteria, Proteobacteria, Bacteroides, Firmicutes, Cyanobacteria, Planctomycetales, Spirochete and Acidobacteria.

For each group different multiple alignments were performed, and different phylogeny trees created. Therefore, it was possible to build many different trees for each group, choosing gradually only the sequences present in the same cluster and which have the longest and the shortest branch. In this way, the most representative sequences were selected and were employed for the building of new trees, all this process has been replicated until the obtaining of a significant reduction of the redundancy within the clusters in the tree (*Fig.10*). Obviously, the presence and the features of the catalytic triad

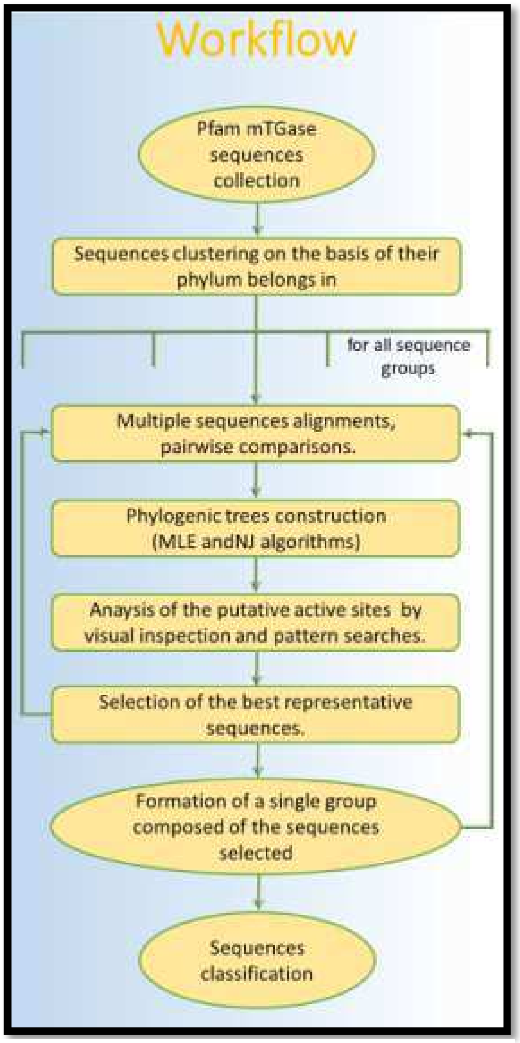

Workflow
Pfam mTGase sequences collection
Sequences clustering on the basis of their phylum belongs in
for all sequence groups
Multiple sequences alignments, pairwise comparisons.
Phylogenic trees construction (MLE andNJ algorithms)
Anaysis of the putative active sites by visual inspection and pattern searches.
Selection of the best representative sequences.
Formation of a single group composed of the sequences selected
Sequences classification

***Fig.10*: Workflow of the clustering procedure.**

and the features of the amino acids around them were observed singularly in order to prevent a possible loss of information during the selection.

At the end of this kind of clustering approach, the initial number of about 8000 sequences was reduced at about 300 sequences, and on the base of their features, it has been possible to make a preliminary classification that brings to the identification of the main MTGase groups.

4.3.b Methods applied for phylogenetic tree constructions

Multiple alignment was performed using MUSCLE and Clustal Omega tools, in order to compare their results, and the TGase cores of all the protein sequences have been analyzed by an integrated procedure including multiple sequence alignments, pairwise comparisons, database searches and pattern searches, by BLAST and Clustal Omega tools and by visual inspection. The analysis of evolutionary relationships among the examined proteins has been performed by MEGA6.0 tool and PhyML, using aLRT SH-like for the branch support.

MEGA: Molecular Evolutionary Genetics Analysis is a software distributed with a nominal fee. It contains facilities for building sequence alignments, inferring phylogenetic histories, and conducting molecular evolutionary analysis (Tamura et al., 2013). In the present work, the construction of phylogenic trees by MEGA6.0 was performed using Maximum Likelihood and Neighbor Joining algorithms, using the default parameters, except for the phylogeny test, where the option Bootstrap method was selected with a number of bootstrap replications set to 100.

PhyML: it is a phylogeny software based on the maximum-likelihood principle (Guindon et al., 2010). It is simple to use and represents a fair compromise between accuracy and speed. It is fast, accurate, stable and available for free as web server from http://www.atgc-montpellier.fr/phyml/.

PhyML parameters used for the present work are the following:

- in the section input data: the files generated by the multiple alignments and converted by AliView in .phylip format were uploaded and the option “amino-acids” was selected for the data type parameter
- in the section substitution model: the automatic model selection by SMS (Lefort et al., 2017) was selected and the option AIC (Akaike Information Criterion) (Akaike, 1973) was chosen.
- in the section tree searching: BIONJ (Gascuel, 1997) was left as default for the option starting tree, instead for the type of tree improvement SPR was chosen and the number of random starting tree fixed to 10. SPR, acronym of subtree pruning and regrafting,

even if slower, generally finds better tree topologies compared to other oldest option available, i.e. NNI (Nearest Neighbor Interchanges) (Guindon et al., 2010).

- in the section branch support: aLRT SH-like was chosen as fast likelihood-based method

aLRT is a new, fast, approximate likelihood-ratio test for branches and is presented as a competitive alternative to nonparametric bootstrap and Bayesian estimation of branch support (Anisimova and Gascuel, 2006). The confidence of the aLRT statistics is estimated by the SH-like algorithm (Guindon et al., 2010). aLRT SH-like is a fast, nonparametric version of the aLRT (Shimodaira–Hasegawa [SH]-aLRT), which was developed and implemented in the PHYML phylogenetic inference software (Guindon et al., 2010). SH-aLRT is derived from the SH multiple tree comparison procedure (Shimodaira and Hasegawa 1999) and is fast due to the RELL technique based on the resampling of estimated log likelihoods (Kishino and Hasegawa 1989) (Anisimova et al., 2011).

The trees obtained by PhyML were visualized by FigTree a graphical viewer of phylogenetic trees (http://tree.bio.ed.ac.uk/software/figtree/).

4.4 Motif researches

At the end of the clustering procedure, it was possible to obtain a first classification of all the best representative mTGase sequences in several groups. All these sequences have been analyzed in order to find specific motifs that could be shared among all the proteins belonging to the same group. Moreover, the presence of motifs shared among different groups was also checked. To do that, protein sequences of the same group and protein sequences belonging to different groups were analyzed by MEME 5.0.1 tool.

MEME: it is the acronym of Multiple EM for Motif Elicitation, where EM stands for Expectation Maximization alghoritm. It is a Motif-based sequence analysis tool that provides motif discovery algorithms using probabilistic model. MEME discovers novel, ungapped motifs (recurring, fixed-length patterns) in the sequences submitted, and is able to split variable-length patterns into two or more separate motifs (Bailey and Elkan, 1994). For this project, the motif discovery mode selected was the “classic mode”, where it is necessary to provide only one set of sequences for MEME to discover motifs enriched in this set, the site distribution chosen was “zero or one occurrence per sequence” or, in case of no result from this one, “any number of repetitions”. The number of motifs to research was set equal to 9.

4.5 The source organism as selection criterion for proteins to model

Before starting the homology modelling procedure, the clustered sequences were analyzed not only on the base of their sequence features but also of their microorganism of origin.

Actually, the construction of 3D models was performed only for those sequences that show the best features also in terms of the source organism. For each organism, in fact, it has been controlled the presence of a Qualified Presumption of Safety (QPS) certification by the European Food Safety Authority (EFSA), if some of them are producers of enzyme already used in the food industry (and so already exposed to the article 8(5) (c) of regulation EU No234/2011), and the legislation about their use as producers of food enzyme where present.

To be granted QPS status, a microorganism must meet the following criteria: its taxonomic identity must be well defined, the available body of knowledge must be sufficient to establish its safety, the lack of pathogenic properties must be established and substantiated and its intended use must be clearly described.

Microorganisms that are not well defined or for which it is not possible to conclude if they pose safety concern to humans, animals or the environment are not considered suitable for QPS status.

For the organisms not present in the EFSA list of Biological Hazards statement on QPS but not judge as unable to obtain it, their safety and the possibility to be procurable were checked.

Using this approach of analysis, many different sequences have been selected, and on the basis of their sequence features, i.e. active site preservation, sequence similarity, presence of absence of key amino acid residues, some of them have been modelled.

4.6 Models building

In order to build the models, the selected sequences have been aligned by BLAST tool. At first the alignment was performed versus the sequence of the MTGase (for this protein, the PDB crystallographic structure 3IU0 has been used in the following steps; this structure present however an inhibitor pre-peptide, so the first 50 amino acids have not been considered for the modelling procedure) and then versus the hTGase2 sequence (PDB structure: 1KV3). This step gives also the information about percentage of identity between query and template and the coverage of the alignments, useful for evaluating the opportunity of applying the modelling phase (template-based procedure). For all the protein sequences that result very similar to MTGase and belonging to the genus Streptomyces another analysis before the modelling has been performed. By ProtParam ExPASy tool (https://www.expasy.org/tools/) (Gasteiger et al., 2005), in fact, a research based on the sequences, about the prediction of the instability index, the aliphatic index, the isoelectric point and more in general about stability has been performed in order to evaluate which sequences present the best features, and to select them for the modelling.

For protein sequences that have very low identity with these two templates, a research of novel templates has been performed using the protein fold recognition server Phyre2 (Kelley et al., 2015) and the I-TASSER (Iterative Threading ASSEmbly Refinement) server, which uses a hierarchical approach to predict protein structure and function (Yang et al., 2015).

After an accurate analysis of the structural features and the overall quality of the templates suggested by the two servers, one or more novel templates have been chosen for the modelling procedure. Before modelling, each sequence selected has been aligned with its template by means of T-Coffee server, because considering its ability to combine sequence information with protein structural information, it allows to obtain better alignments and so better models. When the percentage of identity between query sequence and template was very low (lower or equal to 30%) an alignment of the secondary structure was also performed. To apply this kind of approach it was necessary to make a prediction of the secondary structure of the query sequence, this was obtained by comparison of the results of secondary structure prediction obtained using GOR4 (based on a statistical method; Garnier et al., 1996) and JPred (neural network-based predictor; Drozdetskiy et al., 2015) server.

After setting the alignment, protein models have been created by means of the program Modeller 9.18, a predictor used for homology or comparative modeling of protein three-dimensional structures based primarily on their alignment to one or more proteins of known structure (Webb and Sali, 2016).

4.7 Models validation

The models obtained have been assessed in terms of backbone and side-chain stereo-chemical characteristics, conformational energy, and suitability with known properties. In particular model quality has been evaluated by means of the web servers ProSA-web(Wiederstein and Sippl, 2007) (https://prosa.services.came.sbg.ac.at/prosa.php), QMEAN(Benkert et al., 2008) (https://swissmodel.expasy.org/qmean/) and PROCHECK (Laskowski et al., 1993) (https://servicesn.mbi.ucla.edu/PROCHECK/).

From ProSa-web the main results considered were the Z-score of the model, that indicates the overall model quality in comparison of the Z-scores of all the protein structures obtained by NMR or X-ray, and the plot of residue scores that shows the local model quality by plotting energies. Instead, from all the PROCHECK outputs, the main analysis evaluated was the Ramachandran Plot. Models are considered of good quality if they have a negative Z-score and similar to that of the template, an energy plot in function of amino acid sequence position mainly negative, and a very low percentage of Phi and Psi angles in the generously allowed and disallowed regions of the Ramachandran Plot. Furthermore, they should have a QMEAN value as more as possible close to 1. QMEAN is the acronym of Quality Model Energy ANalysis. It is a composite scoring function which is able to derive both global (i.e. for the entire structure) and local (i.e. per residue) absolute quality estimates on the basis of one single model. QMEAN global scores are originally in a range from 0 to 1, with one being good.

To evaluate the stability of some model a molecular dynamics approach has been also employed, by means of GROMACS 5.0 software. This is part of the abroad training period I spent at ETH Zürich, under the supervision of Prof. A. Caflish (see below, paragraph 5.0).

4.8 Secondary structure analysis

After the models construction, all the secondary structures obtained were analyzed and compared. These analyses regard:

a- proteins modelled belonging to the same group, in order to verify the presence of similar topology;

b- proteins modelled belonging to different groups, in order to analyze which are the differences or the similarity among their structures, above all in those regions that could play a central role in the catalytic activity

c- proteins modelled and proteins whose structures were already known, belonging both to same and to different groups, in order to obtain a reference point and validate what results from the points a and b

All the topologies have been analyzed by Pro-Origami tool and the TGases with a known structure were download from the PDB archive.

Pro-Origami: (http://munk.csse.unimelb.edu.au/pro-origami/) is a system for automatically generating protein structure cartoons. Very user friendly, it provides simplified topology maps of the submitted protein structures aiding in the comparison between structurally related proteins (Stivala et al., 2011).

In the present work, all the structures have been uploaded and submitted choosing the default parameters; i.e. choosing to decompose the structure into domains, the DSSP as secondary structure program, including 3_{10}-helices and π-helices.

4.9 Analysis of the catalytic site pocket

A specific analysis of the active site pocket has been performed for each protein modelled.

More in the details, the hypothetical catalytic amino acids of the modelled sequences have been superimposed, by pyMOL molecular viewer, on the catalytic residues of their relative templates, in order to calculate the Root Mean Square Deviation of the atomic positions (RMSD), i.e. the measure of the average distance between the atoms of superimposed residues.

Moreover, the volume pocket of the templates has been calculated and compared to that one of the models obtained from them. These measures have been done by means of POCASA 1.1 (http://altair.sci.hokudai.ac.jp/g6/service/pocasa/).

POCASA is a program implementing an algorithm called Roll, which can predict binding sites by detecting pockets and cavities of proteins with a rolling sphere. POCASA is not only able to predict the volume of the pockets detected but it presents also a new concept of volume: the volume depth (Vd). This parameter directly and quantitatively describes the position and volume together of a pocket (Yu et al., 2010).

The parameters selected in POCASA for this work differs from one proteins group to another due to the different structural features that occur among the models.

However, in general for small pockets the probe radius parameter (the radius of the probe spheres) was decreased from 2Å to 1Å; the PDF parameter, used to recover the useful pocket points deleted by SPF (single point flag: used to remove noise points of the search result), was modified from a minimum of 10 to a maximum of 18 according to the noise spread; the SPF parameter was fixed to the value of 16, as recommended, for all the analyses.

5.0 Molecular dynamics (MD) simulations

Several molecular dynamics simulations have been performed for both model assessment and structural investigations. At first, the MD was used to assess and to analyze the stability of *Kutzneria Albida* hypothetical mTGase model, in order to verify if this uncharacterized protein showed a behavior similar to the characterized MTGase and of course if, being a model of an uncharacterized protein, this model unfolded during the simulations, showing an implausibility of the predicted structure. However, when the structure of KalbTGase was discovered and so became available, MD simulations were performed on both the structures, KalbTGase (PDBcode:5M6Q) and MTGase (PDBcode:3IU0), in order to analyze their structural features: as the most flexible regions, stability of the active site pocket, conformational change when exposed to different conditions. All the MD simulations have been performed at the Biochemistry Department of the University of Zurich, in the computational and structural biology laboratory under the supervision of Prof. A. Caflisch and his research group (compulsory abroad training period). MD simulations ran on Piz Daint supercomputer (https://www.cscs.ch/computers/piz-daint/), the sixth fastest supercomputer in the world in June 2018 and the third until January 2017. The package used for the simulations and for the most of their analyses was GROMACS 5.0.

GROMACS is one of the most widely used open-source and free software codes in chemistry, used primarily for dynamical simulations of biomolecules. It provides a rich set of calculation types, preparation and analysis tools (Abraham et al, 2015). Developed in 1995, over the last two decades it has evolved from small-scale efficiency to advanced heterogeneous acceleration and multi-level parallelism targeting some of the largest supercomputers in the world (Páll et al., 2015). GROMACS software supports simulations with velocity Verlet, leap-frog Verlet, Brownian and stochastic dynamics, as well as calculations that do energy minimization, normal-mode analysis and simulated annealing. In a typical MD simulation, the user chooses an initial molecular configuration, prepare the simulation describing the atomic interactions and model physics, runs the simulation, and at the end makes observations from the trajectory. Several techniques are available for regulating temperature and/or pressure. Simulations may employ several kinds of geometric restraints, use explicit or implicit solvent, and can be atomistic or coarse-grained. Moreover, multiple simulations can run as part of the same executable, which permits generalized ensemble methods such as replica-exchange. (Abraham et al, 2015).

5.1 Description of the MD simulations performed

In the present work, as mentioned before, two different MD simulation strategies have been employed. The first concerned the model validation of the protein sequence UniProt Code: W5WHY8 from *Kutzneria Albida* mTGase, before its 3D structure availability, instead the second concerned the comparison between MTGase and KalbTGase, both with known structures (experimentally solved).

In the first case seven different MD simulations have been performed:

- the first simulation was performed on the MTGase structure and lasted130ns,
- the second on *Kutzneria Albida* mTGase model and lasted 278ns,

- the third simulation was performed starting from the conformation assumed by the model at the 50th ns of the first simulation and lasted 230ns,
- the fourth starting from the conformation assumed by the model at the 75th ns of the first simulation and lasted 230ns,
- the fifth starting from the conformation assumed by the model at the 100th ns of the first simulation and lasted 230ns,
- the sixth starting from the conformation assumed by the model at the 125th ns of the first simulation and lasted 230ns and
- the seventh starting from the conformation assumed by the model at the 150th ns of the first simulation and lasted 230ns.

All the simulations have been performed at the temperature of 300K in NPT condition.

In the second case, MD simulations have been executed on both KalbTGase crystal structure 5M6Q, become available since the last part of the year 2017, when the Roche Diagnostics GmbH published the discovery of a novel form of mTGase (Steffen et al. 2017), and on the crystal structure of MTGase from *S. Mobaraensis* (PDB code: 3IU0) as a comparison.
MD simulations performed for each structure were:

- 5 independent simulations in NPT condition at a temperature of 300 K, lasting 350ns,

- 1 simulation lasting 300ns at the temperature of 335K

- 1 simulation lasting 300ns too but at a temperature of 355K.

All the .pdb files (both structures and the model) were prepared for the submission using CHARMM-GUI Simulation Input Generator and, after a molecular visual inspection, by manual editing. Before the dynamics began, the topology file has been generated, the box defined and filled with water molecules and the ions added by GROMACS programs. Moreover, to ensure that the so assembled, solvated, electroneutral system has no steric clashes or inappropriate geometry, the structure has been relaxed through a process called energy minimization (EM). The EM was judged as successful when the potential energy value and the maximum force value, printed by GROMACS program at the end of the process, were negative (on the order of 10^5-10^6 kJ/mol^{-1}) and no greater than 1000 kJ mol^{-1} nm^{-1} respectively. An equilibration conducted in two phases, in NVT and in NPT, followed the EM and preceded the dynamics. Particularly, the NVT (constant Number of particles, Volume, and Temperature) equilibration phase ended when the temperature of the system reached the plateau at the specific given value, and the NPT (constant Number of particles, Pressure, and Temperature) equilibration phase ended when the average value of the pressure was around 1 bar and the density reached the expected value. All the simulations ran in NPT. Positional restrains have been included only during NVT and NPT equilibrations. During the NPT equilibration, the barostat chosen for the pressure coupling was Berendsen (Berendsen et al. 1984), during the NPT run, instead, it was replaced with the Parrinello-Rahman (Parrinello and Rahman, 1981) (no pressure coupling is applied during NVT equilibration). More in general, the parameters selected for the simulations were: a leap-frog algorithm (Hockney et al., 1974) for integrating Newton's equations of motion (md) as integrator parameter in Run control section; the LINear Constraint Solver (lincs) (Hess et al., 1997) as constraint-algorithm in the Bond parameters section; Verlet as cutoff-scheme in the Neighbor searching section; PME (Darden et al., 1993) as coulombtype in the Electrostatics section; temperature coupling using velocity rescaling, i.e. v-rescale as tcoupl parameter in the Temperature coupling section. Only in case of parallel multiple simulation runs, velocity generation in NPT production run has been set equal to yes starting from a random seed; otherwise no velocity has been generated after the NVT equilibration phase.

5.2 MD analyses

Many popular simulation file formats can be read via VMD, a molecular visualization program for displaying, animating, and analyzing large biomolecular systems (Humphrey W. et al., 1996). At the end of the simulations, the trajectories have been analyzed by means of Principal Component Analysis (PCA) on the first two eigenvector, Root Mean Square Deviation (RMSD), Root Mean Square Fluctuation (RMSF), by GROMACS programs. DSSP analysis, variation of the dihedral angles analysis, and the SAPPHIRE Plots constructions (Blöchliger N. et al. 2013) have been performed by means of Campari (Vitalis et al., 2009) and CAMPARI analysis tools in R (CampaRi) (http://campari.sourceforge.net/V3/documentation.html).

5.2.a RMSD and RMSF analysis

GROMACS has a built-in utility for RMSD calculations called rms. The command gmx rms compares two structures by computing the root mean square deviation (RMSD). More specifically, with this command, each structure from the trajectory is compared to a reference structure. Thus, for a correct execution of the command it was necessary to provide two different files as input:

- the .tpr file, i.e. the structural file from which the reference structure is taken
- the .xtc file (the file of corrected trajectories, which take into account any periodicity in the system) i.e. the trajectory file from which each structure is taken.

The -tu flag, added to the command, outputted the results in terms of ns, even though the trajectory was written in ps. The RMSD was calculated for all the Cα atoms of the backbone. The lower and more stable the RMSD values, the greater the molecular stability during the simulations and its folding preservation. The RMSD results have been plotted and compared using R.

Even if GROMACS has a built-in utility also for computing the root mean square fluctuation (RMSF), a specific script has been used for this calculation. The RMSF is a measure of the deviation between the position of particle *i* and some reference position, which typically is the time-averaged position of the same particle *i*. The peculiarity of the used script is that the RMSF of the selected atoms is recalculated several times; in fact, the calculation is performed in a time window that increments of 2000ps until the end of the run. Also in this case, the RMSF was calculated only for the Cα atoms of the backbone but the obtained results have been analyzed by XMGrace software. The higher the picks in a specific region shown by the resulting plot, the higher the flexibility of that region, and vice versa.

5.2.b PCA

PCA is a linear transform that extracts the most important elements in the data using a covariance matrix constructed from atomic coordinates, such as the Cartesian coordinates that suggest atomic displacements in each structural conformation of a trajectory.

The eigenvalue decomposition of the covariance matrix leads to a complete set of orthogonal collective modes (eigenvectors), each with a corresponding eigenvalue (variance) that characterizes a portion of the motion. Larger eigenvalues describe motions on larger spatial scales. When the original (centered) data is projected onto an eigenvector, the result is called a principal component (David C. C. and Jacobs D. J., 2014).

Using GROMACS, it was possible to calculates and diagonalizes the (mass-weighted) covariance matrix by the command gmx covar. However, it was necessary to provide the .gro file (the structural file) and the .xtc file (the trajectory file) related to the backbone of the protein. Thereafter, using the GROMACS command gmx anaeig, it was possible to analyze the eigenvectors. Particularly, by the -extr option it was possible to calculate the two extreme projections along the trajectory on the average structure and interpolate all the frames between them (using the -nframes N option; where N stands for the number of the frames comprising the trajectory); the options -first and -last have been used to set the first and the last eigenvector for the analysis, respectively. The command gave as output .gro files, read by VMD, that were useful to detect the principal movements of the analyzed molecules, thus their conformational changes during the simulations.

The option -2d, instead, was used to calculate a 2d projection of the trajectory on the first two eigenvectors, that was written in a .xvg file. XMGrace was used to analyze them.

5.2.c SAPPHYRE PLOT

The SAPPHIRE (States And Pathways Projected with HIgh REsolution) plot provides a comprehensive picture of the thermodynamics and kinetics of complex, molecular systems exhibiting dynamics covering a range of time and length scales. It provides an efficient means of identifying the statistically reliable states visited by a dynamic system giving also an idea of the reoccurrences of these states (Blöchliger N. et al. 2013). Briefly, all snapshots are arranged by geometric criteria that use as only input a definition of pairwise distance between conformations. The resultant sequence (called progress index) is annotated with structural information, times of occurrence (kinetic trace), and a cut function able to highlight basins and barriers (Langini C. et al., 2017). In the present work the SAPPHIRE plot constructions have been performed by means of Campari (Vitalis et al., 2009) and CAMPARI analysis tools in R (CampaRi) (http://campari.sourceforge.net/V3/documentation.html). The plots show the progress index (a specific ordering of the snapshots obtained from the MD simulations), of about 35000 or 30000 snapshots from 350 or 300 ns of MD data, according to the different simulations analyzed, annotated with kinetic information (cut functions, i.e. red and blue curves), dynamical trace (black dots), DSSP assignment or variation of the dihedral angles analysis, depending on the legends present in the plots. In particular, the cut functions allow the identification of metastable states visited by the structure in exam, the dot pattern gives the sampling times and displays the recurrently, i.e. which are the most populated metastable states that are visited multiple times (in contrast to many of the smaller ones), both DSSP and dihedral angle annotations reveal how these states differ from each other.

Moreover, it is necessary to underline that all these annotations are shown using a subsampling factor in order to maintain readability at fixed figure resolution.

5.2.d Volume Analysis (fpocket ed Md pocket)

Analysis of the variation of the volume pocket during all the MD simulations have been performed. To do that two different web servers have been used: Fpocket at first and then MDpocket (Schmidtke et al., 2011) (http://bioserv.rpbs.univ-paris-diderot.fr/services/fpocket/).

Fpocket was used to perform the pocket detection and its output was a useful reference to prepare the input pocket file that is necessary provide in the next steps.

MDpocket was used to track the given pocket in molecular dynamics.

More in the details, Fpocket has a very easy to use interface, requiring just to provide the .pdb file of the protein structure under analysis, in that case the structures from which the MD simulations started. Among its outputs, there is the pdb file of the pockets detected; here it is possible to choose, if present, the pocket of the catalytic site.

Detected the pocket, the next steps require two MDpocket job runs: the first in order to obtain the .pdb file of the pocket grid points and the second to follow the evolution of this pocket during all the dynamics.

For both these steps, it is necessary to prepare a multiple pdb file composed of all the MD simulation snapshots. MDpocket is limited to up to 500 snapshots, because of that, several strategies have been pursued. To make the run faster, the input file of the first run was a multiple .pdb composed by only ten snapshots taken from all the trajectory at regular intervals. This file has been manually edited to adapt its format to one legible from the tool. At the end of this first run the grid pocket file provided as output was used for selecting grid points which defined the zone of interest using PyMOL; for this selection the pocket detected by Fpocket was very useful as a reference to define the grid points to taken and the ones to discard. After the modification of this file, keeping just the grid points of interest, the second run was ready to start.

However, in this case the multiple .pdb file was composed of 500 snapshots.

Therefore, they have been prepared:

- 5 multiple .pdb files; dividing the trajectory in five one after the other parts, was created, for each part, a multiple .pdb file composed of 500 snapshots collected every 20 frames
- 1 multiple .pdb file; it was composed of 500 snapshots taken from the whole trajectory, at a regular interval of about 60-70 (according to the dimension of the simulation).

These multiple .pdb files have been created using both trajectory with and without ions.

Thus, multiple second MDpocket jobs ran, having as input, for the pocket to trace, always the same pdb file composed of the grid points of interest, and as input file for the structures to analyze a different multiple pdb file for each run. Among the output files provided by the second run of MDpocket there is also a .txt file concerning the pocket descriptors. This file contains all pocket descriptors calculated by MDpocket for each of the input snapshots (1 per line), and from its analysis it was possible to see the evolution of the pocket volume.

6.0 Experimental tests

After the mTGase classifications, two proteins have been selected for deeper experimental studies: the first was KalbTGase and the second an automatically annotated protein-glutamine gamma-glutamyltransferase from *SaNDy* (organism not disclosed for patent opportunity).

As regards KalbMTGase, after the discover by Roche, a collaboration with Dr. W. Steffen started, it was possible to obtain the purified protein and directly test it on substrates of interest for this project. The first assayed substrate of interest was the gliadin peptide 56-68, largely studied for its involvement in the coeliac disease onset. The spectrometry mass CeSMA-ProBio lab at the CNR of Avellino, performed the reactions and analyzed the results by mass spectrometry.

The hypothetical mTGase from *SaNDy* was instead cloned, expressed and purified during my thesis work spent in the Laboratory for Molecular Sensing of Dr. S. D'Auria at the CNR of Avellino under the supervision of Dr. A. Pennacchio.

6.1 KalbTGase: mass spectrometry assays

The reaction of transamidation using KalbTGase was conducted in a solution composed of TrisHCL 50 mM, dithiothreitol 1 mM, spermine 50mM, for 1h at 37°C, pH8.00. The following ratio enzyme substrate was used: 4.3 µM enzyme/1.3 mM substrate (almost 1/300 in molarity) and a concentration equal to 1µg/µL for the two peptides: ROCHE control peptide (sequence: *Z-*GGGYRYRQGGGGG*-OH*; molecular weight 1317.5924) and the gliadin peptide 56-68. The reaction of transamidation using MTGase was conducted in a solution composed of Ambic 25mM, spermine 50mM, for 1h at 37°C, pH=8.30. The following ratio enzyme substrate was used: 0.0015Uenzyme/10 µg substrate and a concentration of the two substrate peptides equal to 1µg/µL. The mass spectrometry analyses have been performed using a mass spectrometry with a high resolution and sensibility Q-Exactive basic MS system with an Orbitrap analyzer and a nanoESI source, connected to the capillary-chromatography system UltiMateTM 3000 RSLCnano (THERMO-SCIENTIFIC).

6.2.a Clone and Expression of the protein

Due to the great similarity of this protein with the MTGase, the condition indicated as the best genetic organization for the MTGase vector design was used (Liu S. et al.,2011). Thus, it was constructed a vector pSPRO-STG in which the co-expression was initiated in the order of pro-peptide and mTGase (active form). Both pro-peptide and mTGase (active form) were fused with the pelB signal peptide. The gene was cloned in pET-22b(+) by Nde1/BamH1 by the

GenScript company and delivered to the laboratory for Molecular Sensing at the CNR of Avellino in a lyophilized form.

The lyophilized plasmid was centrifuged at 6000 rpm for 1 minute; in order to resuspend it 20µl of H_2O were added, and the mixture was vortexed for 1minut.

Thereafter, for the transformation:
- 12µg of plasmid was added to 200µl of BL21DE3 and TOP10 competent cells, respectively.
- the mixtures were placed on ice for 30 minutes
- heat shocked at exactly 42°C for exactly 40 seconds.
- placed again on ice for 2 minutes
- in each tube 800µl of room temperature LB medium (10g NaCl, 5g Yeast extract, 10g Tryptone, H_2O until 1L final volume) were added and the tubes were placed at 37°C for 1h.
- After one hour, cells have been centrifuged again, 900µl of supernatant medium removed and resuspended in the remaining 100µl. From each one, 30µl and 70µl have been spread onto selection plates (with ampicillin) and incubate overnight at 37°C.

To start the growth:
the day after, BL21DE3 and TOP10 cells that have expressed the plasmid and so were able to grow, were picked and placed in several different labelled tubes (filled with 10ml of LB medium and ampicillin 100µg/ml) and incubate overnight at 37°C and 160 rpm. After this time, cells concentration has been evaluated by optical density (OD) measurements, performed setting the spectrophotometer at a wavelength of 600 nm. When the OD reached the value of 1.2/1.6, 1ml of cell culture is added to 1L of LB medium + ampicillin 100µg/ml and incubated at 37°C.

To induce the expression:
Several tests of induction and grown at different condition have been performed, in order to decide which was the best competent cell line between TOP10 and BL21DE3.

In order to find the good percentage of growth at which it was better to induce the expression of the vector, a growth curve for the BL21DE3 cell line was performed and after the analysis of the best OD value at which it was possible to induce expression, several tests at different time of induction and concentration of the IPTG inductor were performed. More in the details 3h, 5h, overnight and 30 hours inductions with 0.1mM – 0.3mM – 0.5mM – 0.7mM – 1mM and 1.5mM IPTG at 37°C and 25°C have been tested. The analysis performed suggested that the best expression was obtained when the cell culture reach an OD value equal to 0.7 inducing with 1mM IPTG at 25°C for 30 hours.

TOP10 cell line was dropped due to insufficient results.

Expression tests:

In order to verify the expression of the protein a cell sample (1ml) for each test performed to induce the expression was taken. The samples have been centrifuged at 7000 rpm for 10min and after removing the supernatant, resuspended in 50µl of H_2O. To test the expression each sample was prepared for the SDS-PAGE. 2µl of sample were added to 8µl of H_2O and 5µl of sample buffer to denature the samples, heated for 10 minutes at 100°C and loaded into the wells of the gel at 12% acrylamide prepared for the run. Runs were performed at 100V for the first 30minuts and at 120V until the downmost sign of the protein marker almost reached the foot line of the glass plate. The obtained gels have been rinsed with ddH_2O, microwaved on high power for 1 minute, left in ddH_2O for 10minutes on a rocky table, after removing ddH_2O, Coomassie Stain was added, thus gels have been microwaved on high power for another 1 minute (until the Commassie Stain boiled) and left in incubation with this dye for at list 1 hour on the rocking table. When the gels were well stained the Comassie was removed and the gels incubate a second time in ddH_2O on a rocking table until the level of destaining was sufficient.

6.2.b Purification

In order to purify the protein an osmotic shock was performed, and the obtained fraction was purified by anion-exchange chromatography.

Shock osmotic protocol:

To obtain a reasonable protein amount, 2L of cell cultures were prepared, induced at 0.6OD with 1mM IPTG for 30h at 25°C.

- The culture has been centrifuged at 4000 rpm for 30 minutes at 4°C, the supernatant medium discarded and the pellet weighted.
- Pellet has been resuspended in a solution composed by 30mM Tris-HCl, 1.0mM EDTA, 20% saccharose, pH 8.0 (200ml per liter of the origin culture).
- The mixture has been incubated for 30 minutes at room temperature on a rocking table.
- After this time the cell suspension has been centrifuged for 20minutes at 5000rpm, 4°C, and the obtained supernatant stored in a tube at 4°C
- The obtained pellet has been resuspended in cold water containing10mM $MgCl_2$ (5ml per gram of the original culture pellet) for the osmotic shock and incubated at 4°C for 30 min on a roller mixer.

- After this time the cell suspension has been centrifuged for 20minutes at 5000rpm, 4°C, and the obtained supernatant, thus the periplasmic fraction, has been stored in a tube at 4°C (here should be present the target protein).
- Remaining pellet has been resuspended in an equal volume of H_2O

At the end of the osmotic shock procedure, to analyze if the protein of interest was really present in the periplasmatic fraction, SDS-PAGE assays have been performed, using as samples: a not induced sample (fraction of cell culture collected before the induction), an induced sample (fraction of cell culture collected after the induction but before the first centrifugation step), an input sample (a fraction collected after the pellet resuspension in the TRIS/saccharose/EDTA buffer), the first supernatant obtained, the periplasmatic fraction (second supernatant obtained) and a sample of the resuspended final pellet.

After verifying by SDS-PAGE results that the protein target was actually present in the periplasmatic fraction, this fraction underwent dialysis overnight at 4°C in a water solution containing 20mM $NaH_2PO_4 \cdot H_2O$, pH7.1.

Anion-exchange chromatography:

The anion exchange chromatography was performed using ÄKTA pure protein purification system, where a HiPrep DEAE TF 16/10 column (volume=20ml) was loaded, a weak anion exchanger prepacked with DEAE Sepharose Fast Flow, ready-to-use for fast, preparative separations of proteins and other biomolecules using ion exchange chromatography.

Before starting the chromatography, binding buffer consisting in a solution of 20mM $NaH_2PO_4 \cdot H_2O$, pH 7.0, elution buffer consisting in a solution of 20mM $NaH_2PO_4 \cdot H_2O$, 1.5M NaCl, pH 7.0 and the sample have been made flat and filtrated. Moreover, the column was washed with 20ml of ddH_2O with a flow rate of 0.8ml/min, the pumps and the column were equilibrated with the binding buffer (flow rat 5ml/min per 5 column volume).

The sample was loaded directly into column with a flow rate of 2ml/min (a fraction collector was used). For the column wash a flow rate of 5ml/min per 5 column volume was selected and the fractions were collected. The elution was performed with the elution buffer applied with a gradient step and a flow rate of 1ml/min, fractions were collected as well. Column was washed again with a flow rate of binding buffer equal to 5ml/min.

Fractions which showed a pick in the OD read were analyzed by SDS-PAGE in order to select the ones were the target protein is present. After identifying the correct fractions, they underwent dialysis overnight at 4°C in a water solution containing 20mM $NaH_2PO_4 \cdot H_2O$, pH7.1. After this period, protein solution was stored.

RESULTS AND DISCUSSION

7. Sequence analyses

In order to characterize novel forms of mTGase, the first step to make was the analysis of those protein sequences that were more similar to proteins whose TGase activity was already known. After this phase, that has allowed to have a first overview of how the mTGase world could be wide, a thorough classification of all the protein sequences annotated as hypothetical mTGase was performed and the main features of these sequences, belonging to different groups obtained from the classification, have been characterized by an intensive motifs research.

7.1 Databases searches and sequence alignments results

After a thorough analysis of the state-of-the-art and on the basis of several literature studies on the structure and the organization of both *Streptomyces Mobaraensis* MTGase and hTGase2 proteins, sequences analyses were done, in order to find in specific databases some proteins homologues to MTGase, with potential sequence features suitable for the project aims. As already mentioned, this kind of research was performed on different specialized databases in order to collect the largest number of protein sequences homologues to our query.

However, in order to study the presence of possible interactions between new MTGases and hTGase2 and so, to avoid selecting microbial proteins which could be a threat to human health, search for microbial proteins similar to hTGase2, by specific databases, was performed too. And so, it was possible to understand the existence of many different kinds of mTGases, thus, how wide and heterogeneous the mTGase world is.

7.1a *Streptomyces Mobaraensis* TGase (MTGase) as query

As a result of the search performed to find proteins that could be homologues to MTGase, it was found that 68% of the subject sequences are extracted from bacteria of the Actinobacteria phylum, 60% of whom belong to the Streptomyces genus, as expected. The remaining part is instead extracted from bacteria belonging mainly to the Proteobacteria phylum. The sequence of MTGase presents three different amino acids (C 140, D 331, H 350) making up the catalytic triad in the active site. For each sequence found, an analysis of the conservation of the catalytic residues was done, in order to select only protein sequences that maintain this catalytic region for at least two residues. From this research it was found that all the sequences of proteins belonging to bacteria of the Streptomyces genus preserve all the three catalytic residues and show high query coverage, from 80% to 100%, and so a high number of identical or positive residues, and very low E value, often equal to 0.0. These were expected results especially because the most of these sequences are recorded as protein-glutamine gamma-glutamyltransferase or as transglutaminase. The other protein sources belonging to the Actinobacteria phylum preserve the whole catalytic triad only for the protein recorded as hypothetical and extracted by *Actinobacteria bacterium OV450* and *Kutzneria albida* (discovered as novel mTGase producer only one years ago). *Actinobacteria bacterium OV450* is a specific strain of Streptomyces isolated from the rhizosphere; *Kutzneria albida* is a bacterium of the Actinomycetales order that represents an amazing source of biologically active compounds, which are produced as secondary metabolites (Demainand and Adrio, 2008). The broad spectrum of structural features and wide array of activities of these metabolites attract attention as a limitless source of novel chemical scaffolds as well as new drugs for human and veterinary medicine, and agriculture (Rebets et al., 2014). In fact, nowadays the strain DSM 43870T, is used for the antibiotic production. The protein sequences from *Kutzneria albida* show query coverage of almost 70% and E value equal to $2e^{-12}$, but identity of 28%. Another genus of Actinobacteria, which shows similarity with MTGase is *Nocardia*; different protein sequences from *Nocardia* genus show E value between $1e^{-08}$ and $9e^{-11}$, query coverage of almost 65% but identity of 25-26%. Moreover, these sequences maintain just two of the three catalytic residues (C140 and D331). However, it is important to consider that many *Nocardia* species have been shown to be agents of human diseases, such as *N. asteroides*, *N. farcinica* and *Nocardia nova* (Schaal and Lee, 1992; Wallace et al., 1991), although it has also been shown that some species produce secondary metabolites of potential industrial value (Isik et al., 1999; Kinoshita et al., 2001), e.g. *Nocardia uniformis*, which can produce nocardicin, a β-lactam antibiotic. An additional important protein sequence found is the protein, reported as transglutaminase, extracted from *Thermoactinomyces viridis*, a member of the genus *Thermoactinomyces*. This sequence maintains all the catalytic residues, has E value and identity equal to 0.0 and 100%, respectively. This organism was already employed since 1955 for the antibiotics production and was recently investigated for its transglutaminase activity within a project EU funded named “High performance industrial protein matrices through bioprocessing”. As regards the protein sequences derived from bacteria belonging to the Proteobacteria phylum, these show higher E value, identity of almost 25-35% and query coverage of almost 24%; furthermore, these sequences do not maintain the catalytic triad except for protein sequences from *Pseudoalteromonas rubra*, which preserve D 331 and H 350. Also this genus is well known to be a producer of active compounds. In 2018, after a whole genome sequencing of the organism *SaNDy* (organism not disclosed for patent opportunity), a bacterium not belonging to the Streptomyces genus, a novel sequence extracted from this organism has been annotated as protein-glutamine gamma-glutamyltransferase. This sequence maintains all the catalytic residues, has E value and identity equal to 0.0 and 76%, respectively.

7.1b hTGase2 as query

As previous mentioned, to study the presence of possible interactions between new MTGases and hTGase2, search for microbial protein sequences similar to hTGase2 was performed. It has been found that 56% of microbial protein sequences, which result similar to hTGase2 sequence, belong to Proteobacteria phylum (above all β and δ proteobacteria), 20% to Firmicutes phylum, 12% to Actinobacteria phylum and the remaining 12% is equally divided among Cyanobacteria phylum, Nitospirae phylum and Thermodesulfobacteria phylum. Also for hTGase2 the preservation of the catalytic core among the different sequences, which align with it, was investigated. HTGase, in fact, present a different catalytic triad compared to the MTGase catalytic residues, consisting of: C277, H335 and D358. Therefore, it can be seen the presence of an inversion in the active site between the residues Histidine and Aspartate. All the protein sequences extracted by bacteria belonging to Proteobacteria phylum, maintain the catalytic residues, except for the genus *Ideonella*. Furthermore, proteins, which derived from bacteria belonging to the genus *Methylocaldum szegediense* and *Duganella*, maintain only two of the three catalytic residues that are C277 and H335 for *M. szegediense*, H335 and D358 for *Duganella*. Regarding proteins from Firmicutes phylum, all of them, except for protein extracted by bacteria belonging to the *Candidatus Stoquefichus*, which preserve only the catalityc cysteine and histidine, maintain all three residues. All the remaining proteins extracted from bacteria of the other phylum, analyzed until now, preserve the catalytic triad. In general, all the microbial protein sequences that result similar to hTGase2 sequence show query coverage of almost 30% and identity of 28%, except for the hypothetical protein WP_058554739.1 extracted by *Thiohalocapsa sp. ML1* that present query coverage of 80%, identity of 39% and E value equal to 1e^{-48}. Another noteworthy result is the presence of proteins extracted by *Nocardia sp. BMG111209* in both the outputs of *S. mobaraensis* and hTGase2 similarity search. This suggested to remove these proteins from the set of proteins to be investigated because of the presence of similarity between these proteins and the hTGase2.

7.1c Hypothetical transglutaminase dissimilar to both hTGase2 and MTGase

Starting from the analysis of the sequences similar to hTGase2 and MTGase, it was possible to see that many of these sequences are annotated in Pfam database as sequence with a transglutaminase domain but that the number of sequences with this annotation was very higher (≈ 8000) than the number of sequences found until that moment. Thus, the next step was to download all these sequences and align them with the human and the *Streptomyces* one. From these alignments it was possible to notice that there is a very high number of bacterial protein sequences similar to hTGase2, a small number similar to *S. Mobaraensis*, but that there is also an important number of sequences annotated as hypothetical transglutaminase that does not show similarity with any of the two. From these sequences it was possible to identify a moderate number of protein sequences belonging to the phylum Firmicutes that show similarity with a small microbial transglutaminase, the one from *Bacillus Subtilis*, also called TGl, whose structure was discovered in 2015 and deposited in the PDB as entries 4P8I. As explained more in details in the 1.2.b paragraph, this TGl works through a unique partially redundant catalytic dyad formed by Cys116 and Glu187 or Glu115, even if seems that His200 also plays an important role in the function of the enzyme (Fernandes et al., 2015).

7.2 Sequence clustering

All the results obtained by sequence alignments reinforced the need to make a proper classification, lacking in the mTGase world, of all those sequences that, from the annotation, seem to have a hypothetical transglutaminase domain. Actually, only starting from a thorough clustering procedure it is possible to analyze and thus make a structural and, wherever possible, a functional characterization of novel forms of mTGase.

Therefore, in order to obtain a classification it was used a clustering approach that exploits an iterative procedure (the workflow employed is described under the section: “materials and methods” paragraph 4.3 “Sequence clustering by phylogenetic trees construction”), at the end of the which, it has been possible to build a tree, composed by the best representative sequences of all the ≈8000 present in Pfam database, which in turn allowed to identify five main groups and to make a preliminary classification of these enzymes (Giordano and Facchiano, 2018).

The obtained tree (*Fig.11*) divides the protein sequences as the groups listed below:

1. A first group is composed by mTGases extracted by different Flavobacteria and Sphingobacteria that show the characteristics to be very similar to the mTGase of *Chryseobacterium sp*. (see paragraph 1.2c), a novel form of mTGase, which is very different from all the other mTGases known but whose activity has been experimentally proved by S. Yamaguchi et al. (Yamaguchi et al, 2001).

2. mTGases extracted by bacteria from different phylum compose the second group. They are very similar to the mTGase of *Bacillus Subtilis* (TGl) (see paragraph 1.2b), (Fernandes et al. 2015), but some of them do not preserve all the catalytic residues.

3. A third group is composed by the most common mTGases, which are the mTGase which preserve the main catalytic triad of *Streptomyces mobaraensis* MTGase, in the order Cysteine-Aspartate-Histidine (see paragraph n°1.2a). All of them are Actinobacteria and most of them belong to the genus Streptomyces. They also share good sequence similarity.

4. A little group of proteins from Proteobacteria composes the fourth group. These mTGases differ from all the other mTGases but maybe preserve the catalytic residues in the order Cysteine-Histidine-Aspartate, even if serine could replace cysteine.

5. The last group, the fifth, is the biggest; it is composed by all the mTGases which present a similarity to the eukaryotic TGase and preserve the catalytic triad order typical of the eukaryotic TGase, i.e. Cysteine-Histidine-Aspartate. Aspartate in these proteins is often replaced by Glutamate. Sequences present in this group differ from each other also for the sequence length (double in the half of the cases). It was not possible to obtain a good clustering of these sequences, so this last group would need further deep analyses in order to evaluate the opportunity to obtain a more detailed classification.

This classification represents the basis for starting the model constructions of those sequences, which show the best features also in terms of the source organism.

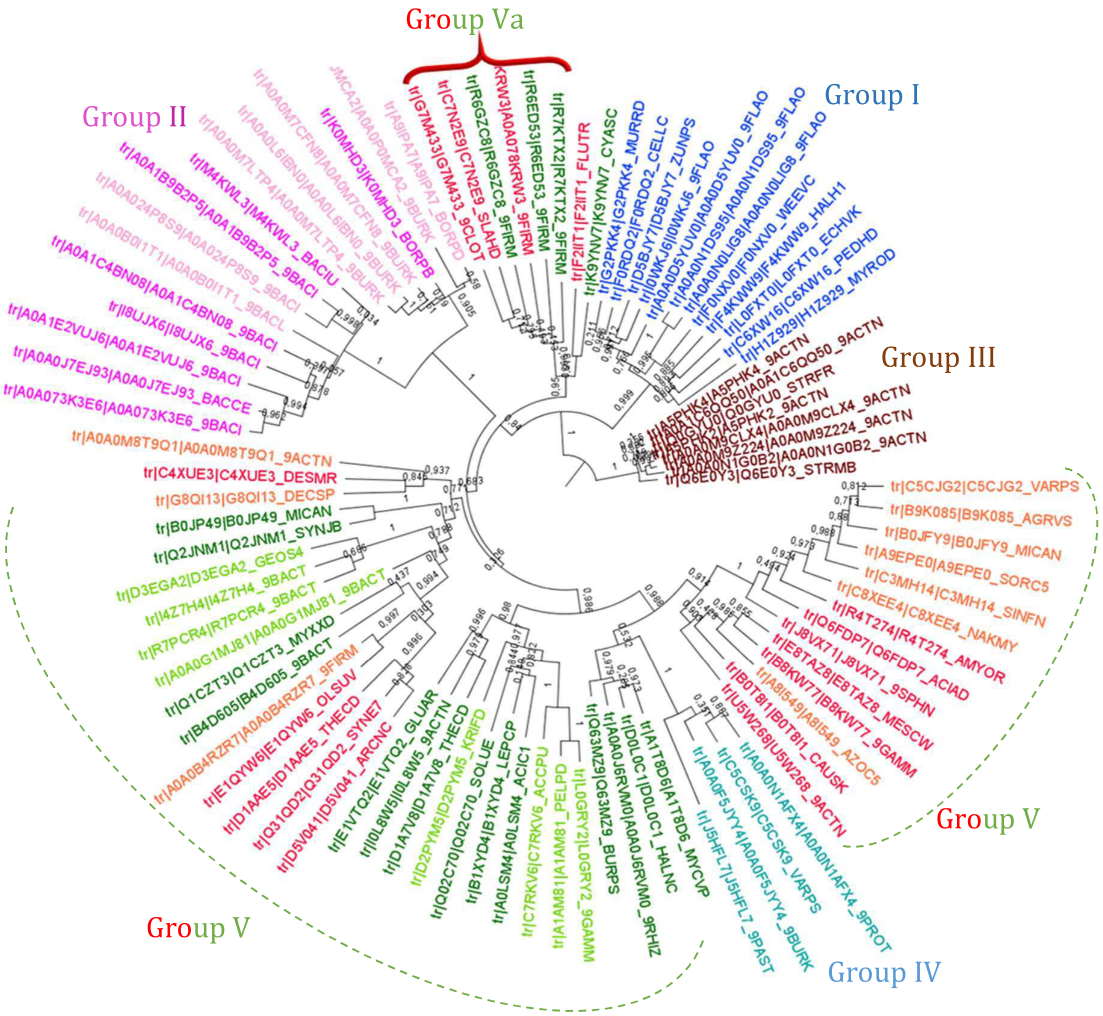


Fig.11: **Phylogenetic tree showing sequence clustering.** The obtained tree is composed of the main sequences annotated as having a transglutaminase domain, it divides the protein sequences in five main groups: labelled in brown, sequences belonging to the third group (MTGase similar); labelled in blue, sequences belonging to the first group (*Chryseobacterium sp*. mTGase similar); labelled in pink and fuchsia, sequences belonging to the second group (TGl similar), where pink labels indicate not complete preservation of the catalytic residues; labelled in cyan, sequences belonging to the fourth group (Protobacteria sequences with no similarity with TGase already known); labelled in green, red and orange sequences belonging to the fifth group (hTGase2 similar): sequences labelled in green has a double length than the other in orange or red, sequences in light green and orange do not preserve all the catalytic residues. Highlighted by a red bracket a small group of sequences that even if are similar to the hTGase2 are located in a separate branch of the tree.

7.3 Motif research

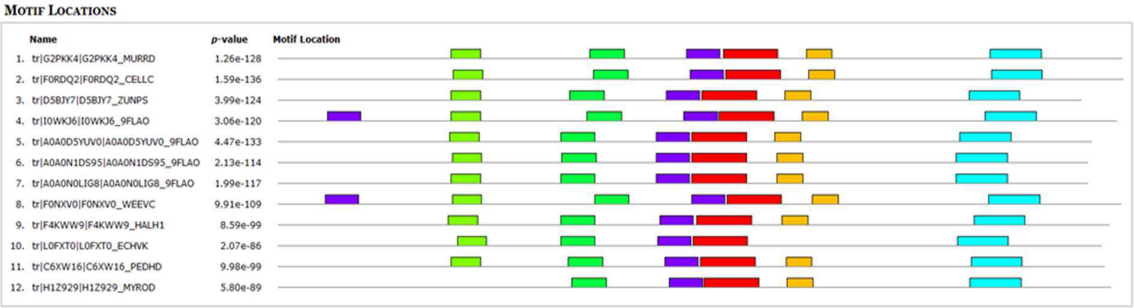
After the obtainment of a first classification the following step was the analysis of the sequences features that allows some sequences to be pooled in the same group and the others to form separated groups. In order to do that a thorough search for motifs has been performed involving many different comparisons by MEME tool (see paragraph 4.4 "Motif research"). The search, as explained in more details in the following paragraphs, has underlined as each group shares specific motifs that make it different from the others (Giordano and Facchiano, 2018). No motifs are shared among different groups, except for the group highlighted by the red bracket in figure 11 (group Va), this group shares, as expected, some motifs with the group V. This is a very important case to study because from its analysis it is possible to hypothesize which motifs are the most ancient as well as the most preserved (see paragraph 7.3.f). In the next paragraphs a description of the motifs found for each group will be dealt.

7.3.a Motif research in the 1st group

As explained in the paragraph related to the preliminary classification of the mTGase, the first group is composed by mTGases extracted by different Flavobacteria and Sphingobacteria that show the characteristics to be very similar to the mTGase of *Chryseobacterium sp.*
This protein is a novel form of mTGase, whose deamidation activity has been experimentally proved by S. Yamaguchi et al. More properly seems that this enzyme is not a real Transglutaminase but a protein-deamidating enzyme, because no transglutaminase activities were detected based on the lack of hydroxyamate formation and lack of cross-linked product formation from caseins (Yamaguchi et al, 2001).

However, also for this group a motif research has been performed, and the found motifs have been searched in the other sequences of the phylogenetic tree in order to verify the unicity of the group. From the obtained results it is possible to identify six different motifs showing the features explained in *fig.12(A/B)* but, unfortunately, because of the unicity of these proteins, it is not possible to say where are located the catalytic residues.

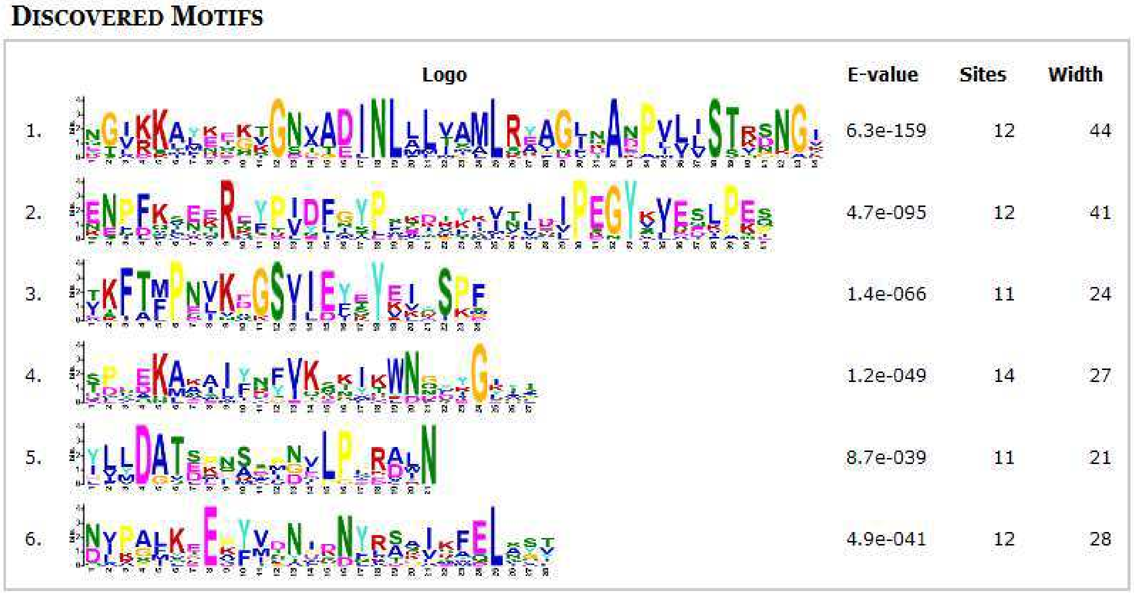
A



Motif Locations

	Name	p-value	Motif Location
1.	tr|G2PKK4|G2PKK4_MURRD	1.26e-128	
2.	tr|F0RDQ2|F0RDQ2_CELLC	1.59e-136	
3.	tr|D5BJY7|D5BJY7_ZUNPS	3.99e-124	
4.	tr|I0WKJ6|I0WKJ6_9FLAO	3.06e-120	
5.	tr|A0A0D5YUV0|A0A0D5YUV0_9FLAO	4.47e-133	
6.	tr|A0A0N1DS95|A0A0N1DS95_9FLAO	2.13e-114	
7.	tr|A0A0N0LIG8|A0A0N0LIG8_9FLAO	1.99e-117	
8.	tr|F0NXV0|F0NXV0_WEEVC	9.91e-109	
9.	tr|F4KWW9|F4KWW9_HALH1	8.59e-99	
10.	tr|L0FXT0|L0FXT0_ECHVK	2.07e-86	
11.	tr|C6XW16|C6XW16_PEDHD	9.98e-99	
12.	tr|H1Z929|H1Z929_MYROD	5.80e-89	

B



DISCOVERED MOTIFS

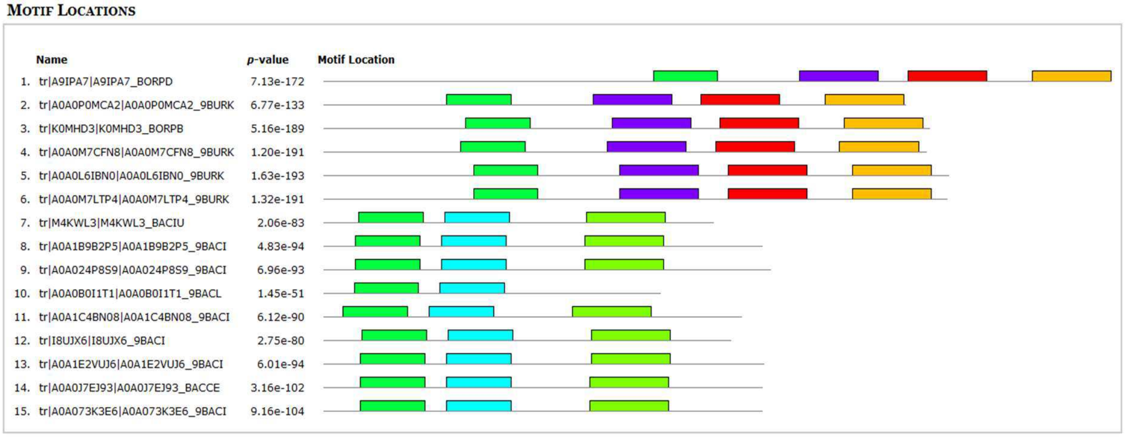
	Logo	E-value	Sites	Width
1.		6.3e-159	12	44
2.		4.7e-095	12	41
3.		1.4e-066	11	24
4.		1.2e-049	14	27
5.		8.7e-039	11	21
6.		4.9e-041	12	28

Fig.12: **Motifs found in the 1st group of sequences. A**: motifs locations along the sequences, different box colors correspond to different motifs **B**: the consensus sequences of the six best motifs are reported with related statistical parameters: motif no.1 is referred to the motif highlighted by the red boxes, motif no.2 to the motif highlighted by the light blue boxes, no.3 by light green boxes, no.4 by the violet boxes, no.5 by the yellow boxes, no.6 by the dark green boxes.

Moreover, the results showed that there is the absence of motifs belonging to protein sequences present in this group in any other sequences of the tree and so in any other of the five groups. Therefore, the sequence analysis and the motif searches confirm this group as independent from the other microbial transglutaminase under investigation.

7.3.b Motif research in the 2nd group

As mentioned in the previous pages the second group is composed by the mTGase similar to the TGl, which functions through a unique partially redundant catalytic dyad (C116 and E187/E115). From the analysis of the sequence features is possible to find six motifs (*Fig.13*)



MOTIF LOCATIONS

	Name	*p*-value	Motif Location
1.	tr|A9IPA7|A9IPA7_BORPD	7.13e-172	
2.	tr|A0A0P0MCA2|A0A0P0MCA2_9BURK	6.77e-133	
3.	tr|K0MHD3|K0MHD3_BORPB	5.16e-189	
4.	tr|A0A0M7CFN8|A0A0M7CFN8_9BURK	1.20e-191	
5.	tr|A0A0L6IBN0|A0A0L6IBN0_9BURK	1.63e-193	
6.	tr|A0A0M7LTP4|A0A0M7LTP4_9BURK	1.32e-191	
7.	tr|M4KWL3|M4KWL3_BACIU	2.06e-83	
8.	tr|A0A1B9B2P5|A0A1B9B2P5_9BACI	4.83e-94	
9.	tr|A0A024P8S9|A0A024P8S9_9BACI	6.96e-93	
10.	tr|A0A0B0I1T1|A0A0B0I1T1_9BACL	1.45e-51	
11.	tr|A0A1C4BN08|A0A1C4BN08_9BACI	6.12e-90	
12.	tr|I8UJX6|I8UJX6_9BACI	2.75e-80	
13.	tr|A0A1E2VUJ6|A0A1E2VUJ6_9BACI	6.01e-94	
14.	tr|A0A0J7EJ93|A0A0J7EJ93_BACCE	3.16e-102	
15.	tr|A0A073K3E6|A0A073K3E6_9BACI	9.16e-104	

Fig.13: **Motifs found in the 2nd group of sequences.** Motifs locations along the sequences: different box colors correspond to different motifs

The results show the presence of a motif *(Fig.14)* shared by all the sequences that explain their clustering in the same group (first green boxes in *fig.13*). This motif has a length of 41 amino acids and an E-value equal to 1.1e^{-121}, but it does not involve any catalytic residues.

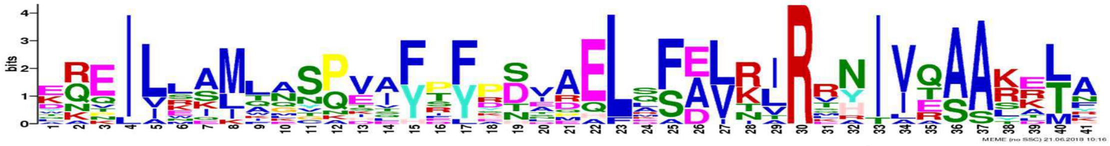


Fig.14: **Sequence motif shared by all the sequences belonging to the 2nd group.** It is referred to the motif highlighted by the green boxes in fig.13, here its composition is shown.

The other motifs, showed as results, lead to a division of the sequences analyzed in two main groups: in the first, it is possible to see three motifs (violet, red and yellow), in the motif highlighted in violet (*Fig.15A*) the catalytic cysteine is present. This motif is really well preserved. It has a length of 50 amino acids, and an E-value equal to 2.3e^{-131}. The following motif, the red one (*Fig.15B*), concerns the other catalytic residue, however, many of these sequences do not preserve it. Also in this case it is well preserved and the length is of 50 amino acids, however this motif shows a better E-value, equal to 2.3e^{-183}. The last motif, the yellow (*Fig15.C*), has an E-value equal to 6.8e^{-114} and the length is of 50 amino acids too.

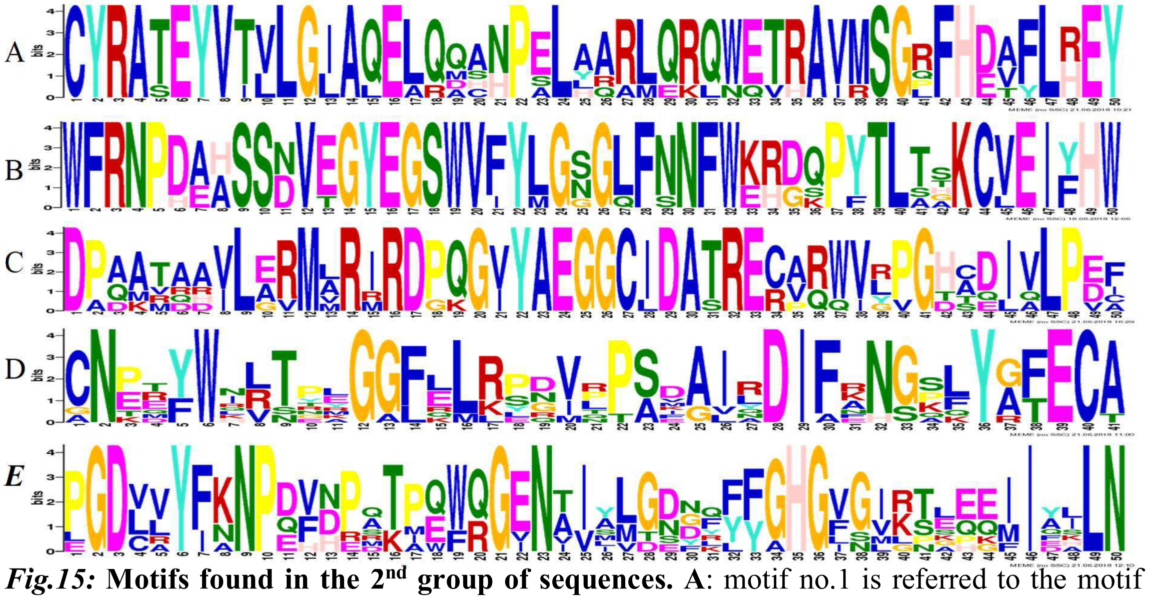


Fig.15: **Motifs found in the 2nd group of sequences. A**: motif no.1 is referred to the motif highlighted by the violet boxes **B:** motif no.2 to the motif highlighted by the red boxes, **C**: no.3 by yellow boxes, **D**: no.4 by the light cyan boxes, **E**: no.5 by the light green boxes

Considering the sequences starting from number seven to fifteen represented in *fig.13*, the presence of two different motifs shared by all of them is easy to detect. In the motif highlighted in cyan (*fig.15D*) there are two important residues for the catalysis i.e. Glutamine followed by

Cysteine matching Glu115 and Cys116 of the TGl. The motif has length of 41 amino acids and good E-value, equal to 2.8e^{-129}.

In the second motif (*Fig.15E*), highlighted in light green, it is present not only the second catalytic residue Glutamine corresponding to Glu187 but also the Histidine corresponding to His 200 of the TGl, this amino acid as mentioned before seems play an important role in the function of the enzyme. This motif is composed by 41 amino acids and has E-value of 3.0e^{-137}.

If we compare these results to the clustering obtained in the tree, it is possible to see that we find the same division in the branches of the tree: so that the first group detected by the motif analysis correspond to the sequences gather by the green circle and the second to the sequences gather by the light blue one (*Fig.16*).

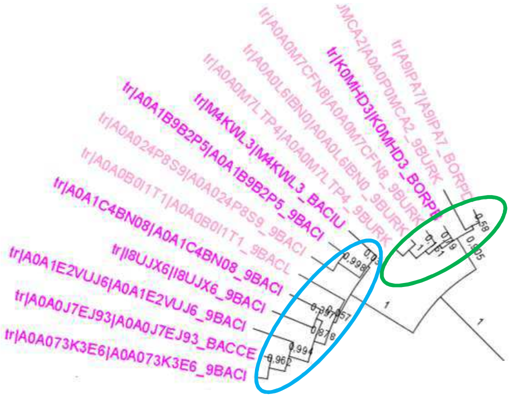
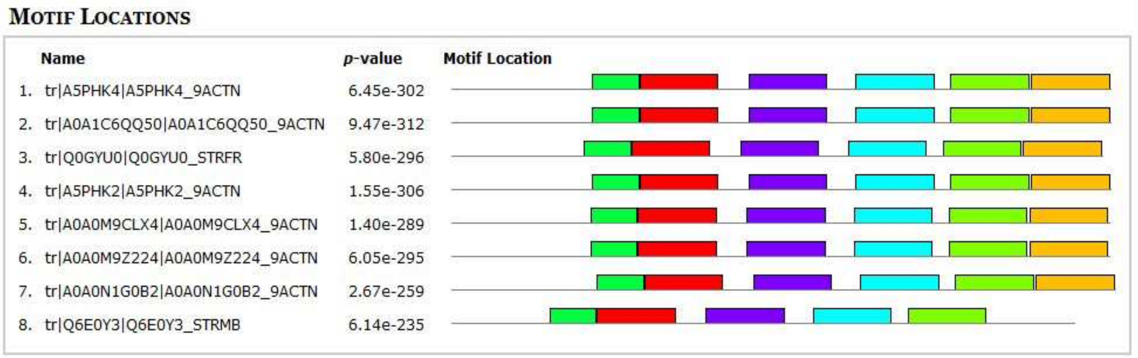

tr|A9IPA7|A9IPA7_BORPD
JMCA2|A0A0P0MCA2_9BURK
tr|K0MHD3|K0MHD3_BORPB
tr|A0A0M7CFN8|A0A0M7CFN8_9BURK
tr|A0A0L6IBN0|A0A0L6IBN0_9BURK
tr|A0A0M7LTP4|A0A0M7LTP4_9BURK
tr|M4KWL3|M4KWL3_BACIU
tr|A0A1B9B2P5|A0A1B9B2P5_9BACI
tr|A0A024P8S9|A0A024P8S9_9BACI
tr|A0A0B0I1T1|A0A0B0I1T1_9BACL
tr|A0A1C4BN08|A0A1C4BN08_9BACI
tr|I8UJX6|I8UJX6_9BACI
tr|A0A1E2VUJ6|A0A1E2VUJ6_9BACI
tr|A0A0J7EJ93|A0A0J7EJ93_BACCE
tr|A0A073K3E6|A0A073K3E6_9BACI
0.58
0.79
0.905
1
0.998
0.397
0.957
0.878
0.994
0.962
1
1

Fig.16: **The 2nd group branch.** A detail from the tree in fig.11 which shows the splitting of the 2nd group in two subgroups.

Since most of the sequences belonging to this group are from bacteria belonging to the Firmicutes phylum, another important aspect that has been investigated is the hypothetical presence among these sequences of motifs shared with the other Firmicutes sequences, which are also present in group V due to their similarity with the hTGase2. The analysis performed demonstrated that there is no motif shared by these sequences, therefore there is no similarity among them and it is possible to state that they really belong to two different groups of mTGases.

7.3.c Motif research in the 3rd group

The third group is composed by the sequences similar to the microbial transglutaminase extracted from *S. Mobaraensis.* The most of this group is composed by sequences derived from organism belonging to the Streptomyces genus. Among them, the percentage of identity is very high ad so the motif found are all very well preserved. MEME tool shows for the proteins belonging to the Streptomyces genus 6 motifs (*Fig.17*) highlighted by dark green, red, violet, cyan, light green and yellow boxes



MOTIF LOCATIONS

Name	p-value	Motif Location
1. tr|A5PHK4|A5PHK4_9ACTN	6.45e-302	
2. tr|A0A1C6QQ50|A0A1C6QQ50_9ACTN	9.47e-312	
3. tr|Q0GYU0|Q0GYU0_STRFR	5.80e-296	
4. tr|A5PHK2|A5PHK2_9ACTN	1.55e-306	
5. tr|A0A0M9CLX4|A0A0M9CLX4_9ACTN	1.40e-289	
6. tr|A0A0M9Z224|A0A0M9Z224_9ACTN	6.05e-295	
7. tr|A0A0N1G0B2|A0A0N1G0B2_9ACTN	2.67e-259	
8. tr|Q6E0Y3|Q6E0Y3_STRMB	6.14e-235	

***Fig.17*: Motifs found in the 3rd group of sequences.** Motifs locations along the sequences: different box colors correspond to different motifs

For the protein belonging to other genus MEME recognizes the motifs highlighted in red and cyan, however it does not include them in the results because the p-value associated to these motifs is a little bit higher than the significance level.

However, it is important to underline that the red boxes showed the motif that include the catalytic residue cysteine (*Fig.18A*). Its width is 50 amino acids and its E-value is equal to 1.4e^{-271}. The second motif found by MEME, instead, covers a portion preceding the third motif discovered (highlighted in light green) where the catalytic histidine and aspartate are located (*Fig.18B* and *C* respectively).

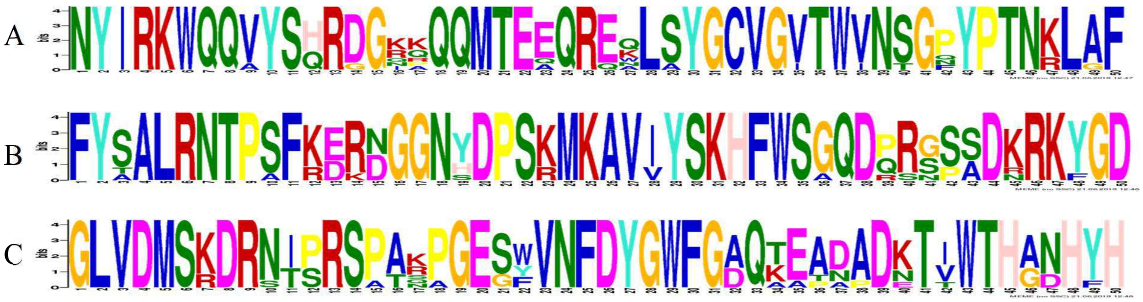


***Fig.18:* Motifs found in the 3^{3d} group of sequences. A**: motif no.1 is referred to the motif highlighted by the red boxes **B**: motif no.2 to the motif highlighted by the cyan boxes, **C**: no.3 by yellow boxes, **D**: no.4 by the light green boxes

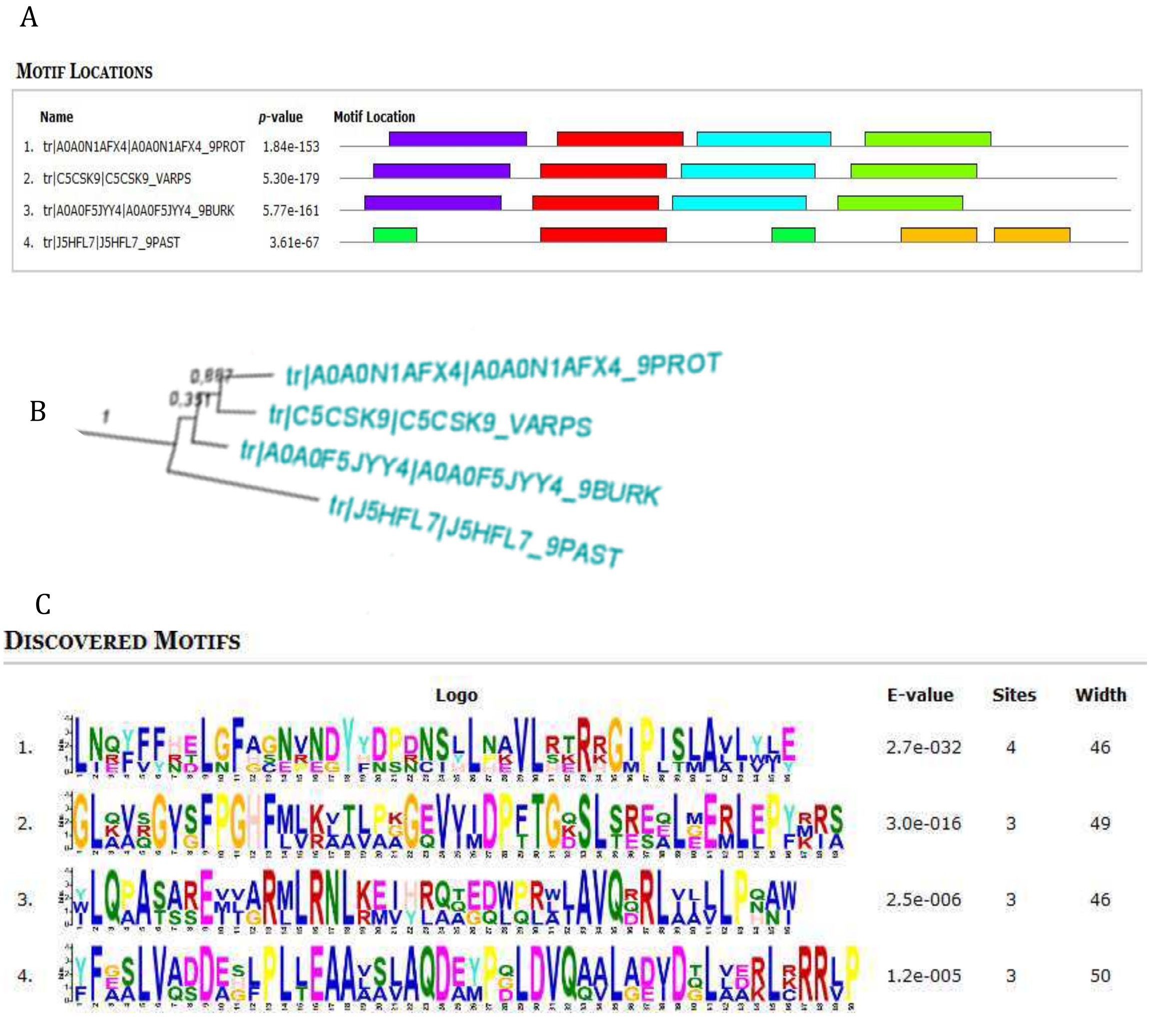
Both the motifs, the one highlighted in cyan and the other in light green, have a length of 50 amino acids, are really well preserved as the motif highlighted in red and have an E-value equal to 4.4e^{-230} and 9.3e^{-230}, respectively.

Another important aspect to underline is the absence of motifs belonging to mTGases similar to *Streptomyces Mobaraensis* MTGase in any other sequences of the tree. Therefore, also in this case, it is possible to conclude that the sequence analysis and the motif searches confirm this group as a particular class of microbial transglutaminase.

7.3.d Motif research in the 4th group

In this group, as mentioned in the previous paragraphs, there is a little group of proteins from Proteobacteria, which differ from all the other mTGases but that probably preserve the catalytic residues in the order Cysteine-Histidine-Aspartate, even if serine could replace cysteine.

In the tree, just four sequence are representative of this class, and they were searched for peculiar motifs. MEME tools give as output four motifs marked by violet, red, light blue and green in the graphical representation (*Fig.19A*). As it is possible to see from the results, only the red one is shared by all the sequences, more properly the sequence tagged by the UniProt Code J5HFL7 seems to be quite different from the others; observation that is also confirmed by its marginal position in the phylogenetic tree (*Fig.19B*)



Name	p-value
1. tr|A0A0N1AFX4|A0A0N1AFX4_9PROT	1.84e-153
2. tr|C5CSK9|C5CSK9_VARPS	5.30e-179
3. tr|A0A0F5JYY4|A0A0F5JYY4_9BURK	5.77e-161
4. tr|J5HFL7|J5HFL7_9PAST	3.61e-67

	E-value	Sites	Width
1.	2.7e-032	4	46
2.	3.0e-016	3	49
3.	2.5e-006	3	46
4.	1.2e-005	3	50

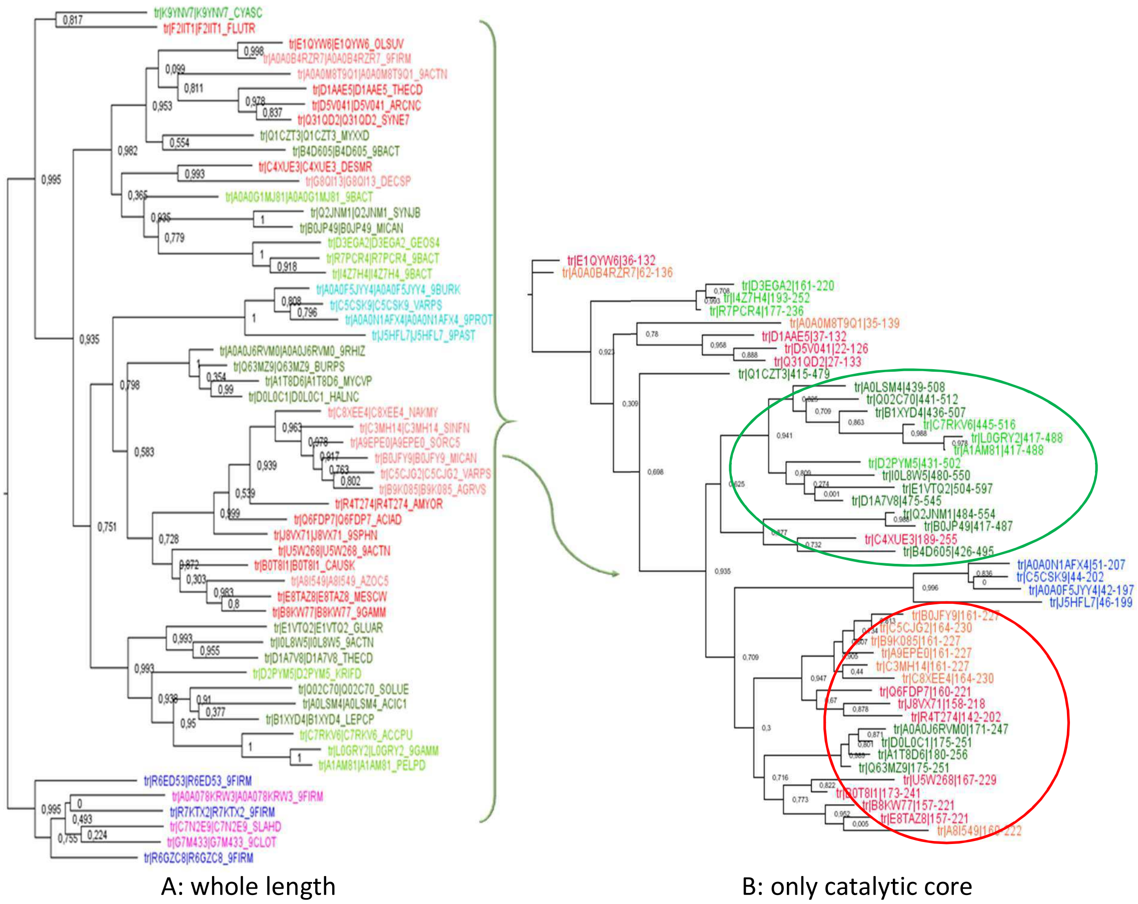
Fig.19: **Motifs found in the 4th group of sequences. A**: motifs locations along the sequences, different box colors correspond to different motifs **B**: A detail from the tree in *fig.11* which shows the separation of the sequence having UniProt code J5HFL7 from the rest of the group. **C**: the consensus sequences of the fourth best motifs are reported with related statistical parameters: motif no.1 is referred to the motif highlighted by the red boxes, motif no.2 to the motif highlighted by the light blue boxes, no.3 by light green boxes, no.4 by the violet boxes.

Analyzing the alignment and the preserved regions, it seems that the first catalytic residue could be the Serine present in the last part of the first motif in the sequence region: G-I-P-I-S-L-M-A-V-L (*Fig.19C*). The hypotheticals catalytic Histidine and Aspartate residues could be, instead, located in the motif highlighted by the light blue boxes in the sequence region F-P-G-H-F and V-V-I-D-P-F-T (*Fig.19C*).

Moreover, it is important to underline that the motifs found are typical of these sequences and are not shared by other sequences into the tree; therefore, it is possible to say that they form an independent group of mTGase.

7.3.e Motifs research in the 5th group

Since the 5th group is the biggest and the most heterogeneous group present in the tree, it was built also a phylogenetic tree related only to the catalytic core of all the sequences



A: whole length

B: only catalytic core

***Fig.20:* Phylogenetic trees showing sequence clustering in the 5th group.** The obtained tree is composed of the main sequences similar to the hTgase2 and annotated as having a transglutaminase domain: sequences labelled in green has a double length than the other in orange or red, sequences in light green and orange do not preserve all the catalytic residues; in violet the outgroup. **A**: Phylogenetic tree of the 5th group sequences used in their full length. **B**: Phylogenetic tree of the 5th group sequences built using only the catalytic cores.

Looking at the tree at the left side of *Fig.20* (A) it is possible to see sequences labelled in a different way. The orange/red ones are sequences with a small length, the dark and light green ones are instead sequences with a length at least double of the previous ones. However, the phylogenetic tree at the right side of *Fig.20* (B), composed just of the catalytic core sequences, shows the same sequence clustering; therefore, it is possible to say that all these sequences divide themselves independently from the sequence length.

From the previous mentioned tree, it is possible to see two different groups, labelled by a green and a red circle respectively (*Fig.20B*).

All the sequences inside the green circle share mainly two motifs: in the first it is located the Cysteine catalytic residue in the second Histidine and Aspartate/Glutamate (*Fig.21A* and *B* respectively).

The situation is almost the same also for the sequences present in the red circle, with two motifs one for the first catalytic residue Cysteine/Serine and the second for the catalytic Histidine (*Fig.21C* and *D,* respectively).

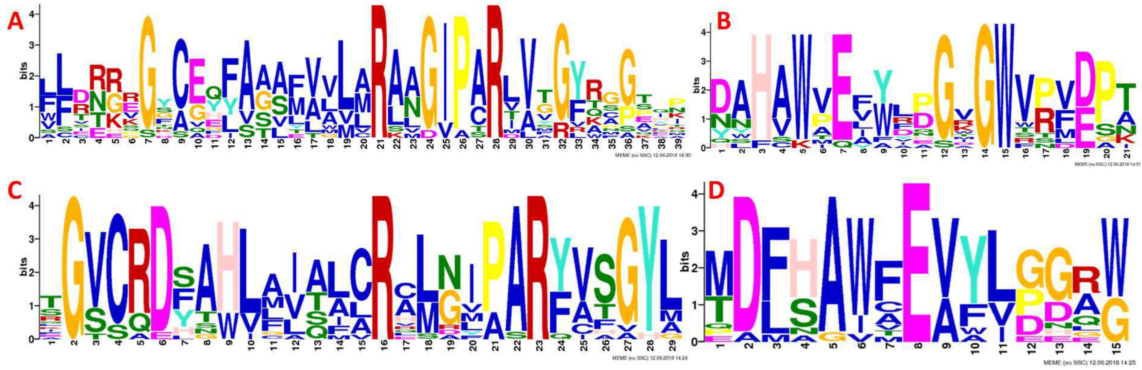
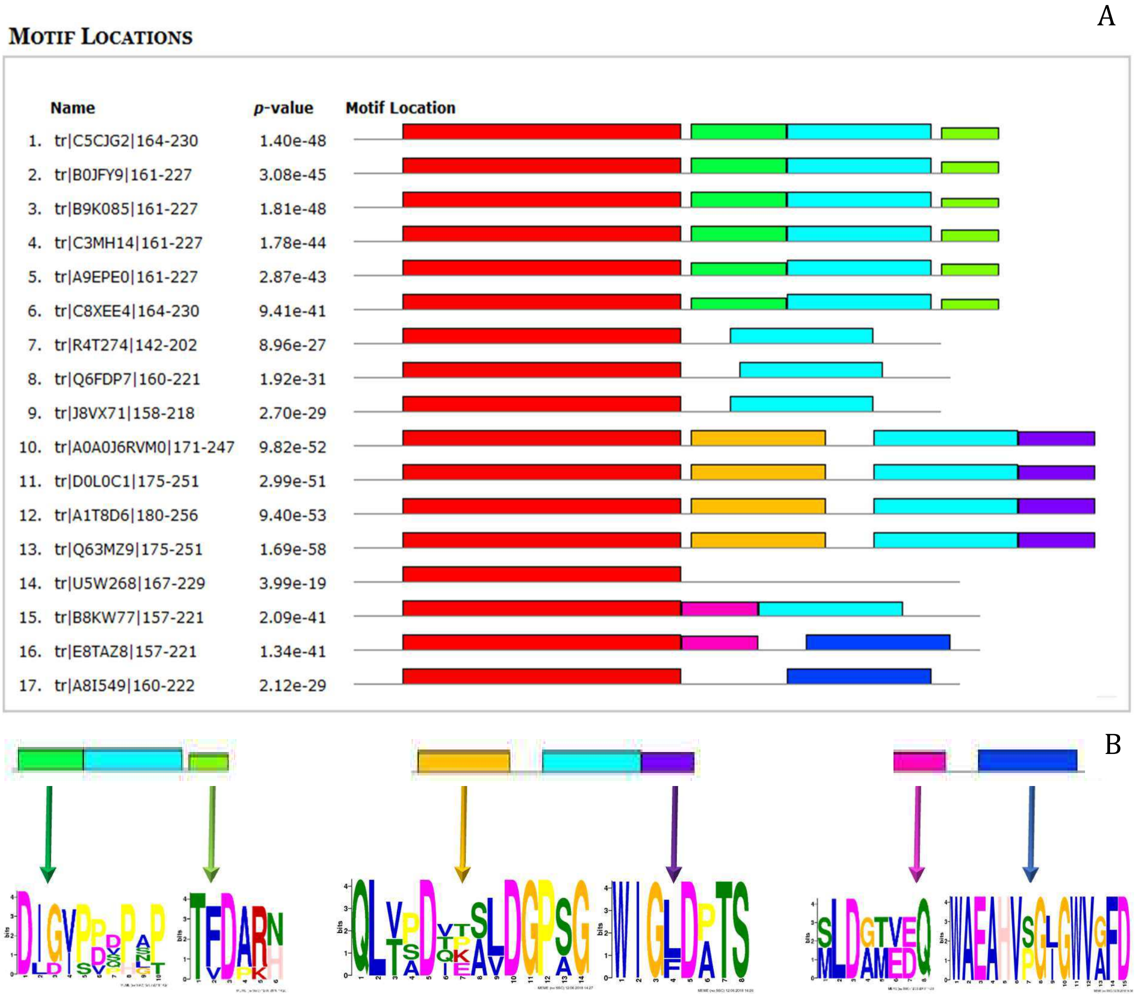


Fig.21: **Motifs found in the 5th group of sequences. A**: motif no.1 is in the sequences labeled by the green circle, here is present the catalytic Cys **B**: motif no.2 is in the sequences labeled by the green circle, here are present the other two catalytic residue **C**: referred to the sequences labeled by the red circle, this motif shows the preservation of the catalytic Cys **D**: also referred to the sequences labeled by the red circle, the motif shows the preservation of the catalytic His.

Looking at the motifs, if the first motif of the two groups are considered, i.e. motifs containing the Cysteine catalytic residue, it can be notice that the one from the green circled group is longer than the other (39 amino acids compared to the 29 of the red circle sequence motif) but they share some amino acids in particular positions. The preservation of these amino acids could suggest a particular relevance of these residues, which could be involved in the specificity of the enzyme, in the active site architecture preservation or in the regulation of the enzymatic activity.

An analogues situation regards the description and the comparison of the other two motifs: the catalytic Histidine and the amino acids closely related to it are involved in both. However, the motif related to the green group involves also the third catalytic residue and the residues around it, in the red group, instead, it is not possible the detection of an analogue clear motif common to all the sequences.

Analyzing the sequences in the red circle, it is possible to see also the existence of other motifs (green and yellow boxes *Fig.22A*) that validate the presence of two distinct branches and more specifically the clustering of the sequences labelled in orange and the sequences labelled in green.



Name	*p*-value	Motif Location
1. tr|C5CJG2|164-230	1.40e-48	🟥🟩🟦🟩
2. tr|B0JFY9|161-227	3.08e-45	🟥🟩🟦🟩
3. tr|B9K085|161-227	1.81e-48	🟥🟩🟦🟩
4. tr|C3MH14|161-227	1.78e-44	🟥🟩🟦🟩
5. tr|A9EPE0|161-227	2.87e-43	🟥🟩🟦🟩
6. tr|C8XEE4|164-230	9.41e-41	🟥🟩🟦🟩
7. tr|R4T274|142-202	8.96e-27	🟥🟦
8. tr|Q6FDP7|160-221	1.92e-31	🟥🟦
9. tr|J8VX71|158-218	2.70e-29	🟥🟦
10. tr|A0A0J6RVM0|171-247	9.82e-52	🟥🟧🟦🟪
11. tr|D0L0C1|175-251	2.99e-51	🟥🟧🟦🟪
12. tr|A1T8D6|180-256	9.40e-53	🟥🟧🟦🟪
13. tr|Q63MZ9|175-251	1.69e-58	🟥🟧🟦🟪
14. tr|U5W268|167-229	3.99e-19	🟥
15. tr|B8KW77|157-221	2.09e-41	🟥🟪🟦
16. tr|E8TAZ8|157-221	1.34e-41	🟥🟪🟦
17. tr|A8I549|160-222	2.12e-29	🟥🟦

***Fig.22* Motifs found in the 5th group of sequences labeled by the red circle. A**: motifs locations along the sequences, different box colors correspond to different motifs **B:** motif composition of each box found by MEME.

Among all the motifs detected, noteworthy are the two motifs marked by the light green and the violet boxes; here, it is possible to see two distinct motifs that include the catalytic Aspartate

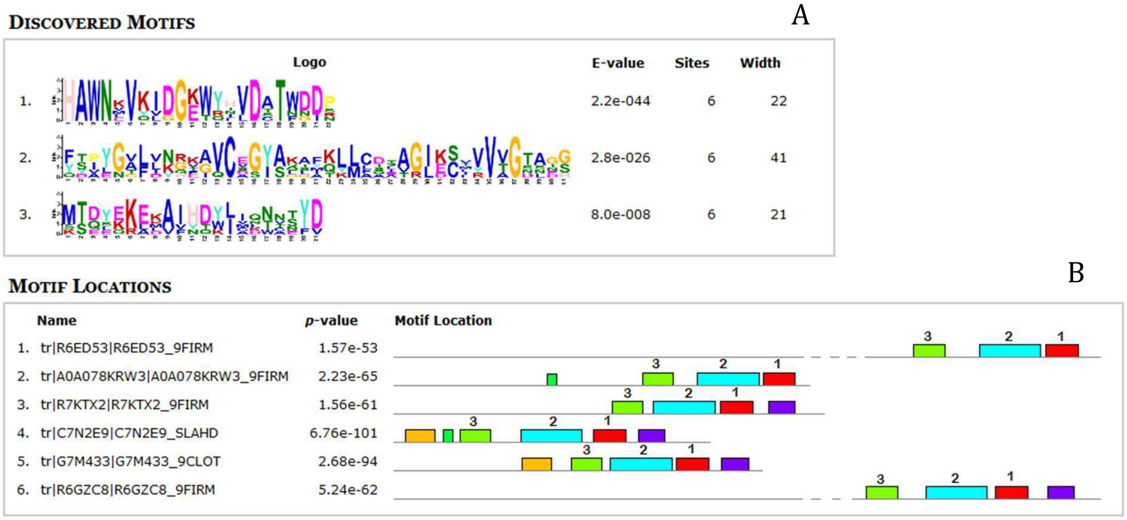
(D): T-[FV]-D-[AP]-[RK]-[NR] for the sequences labelled in orange and W-I-G-[LF]-D-[PA]-T-S for the ones labelled in green.

Peculiar motifs are present also between the motif related to the catalytic Cysteine and the other related to the catalytic Histidine as remarked by the dark green and yellow boxes in the *fig.22B*.

Even if all the motifs descripted are peculiar of this group, the motifs related to the catalytic Cysteine and Histidine (*Fig.21*), as described before, share some residues that are also shared in another group, the one labeled as group Va and highlighted by a red bracket in the main phylogenetic tree showed in *fig.11*. This group, even if present sequence features that associate it with the group V (the sequences belonging to it actually are similar to the hTGase2), is located in a different branch of the tree and because of that it deserves a separate description.

7.3.f Motifs research in the group Va

Sequences from organisms belonging to the Firmicutes philum cluster in this group. However, these sequences are clearly different from the sequences from Firmicutes present in group II. Actually, they present three not contiguous amino acids potentially indicating the catalytic triad, in the order Cysteine-Histidine-Aspartate, i.e. the order observed in eukaryote TGases, and in group V. The results of the search for motifs in the Firmicutes sequences of group Va (*Fig.23*) are summarized here.



DISCOVERED MOTIFS

	Logo	E-value	Sites	Width
1.		2.2e-044	6	22
2.		2.8e-026	6	41
3.		8.0e-008	6	21

MOTIF LOCATIONS

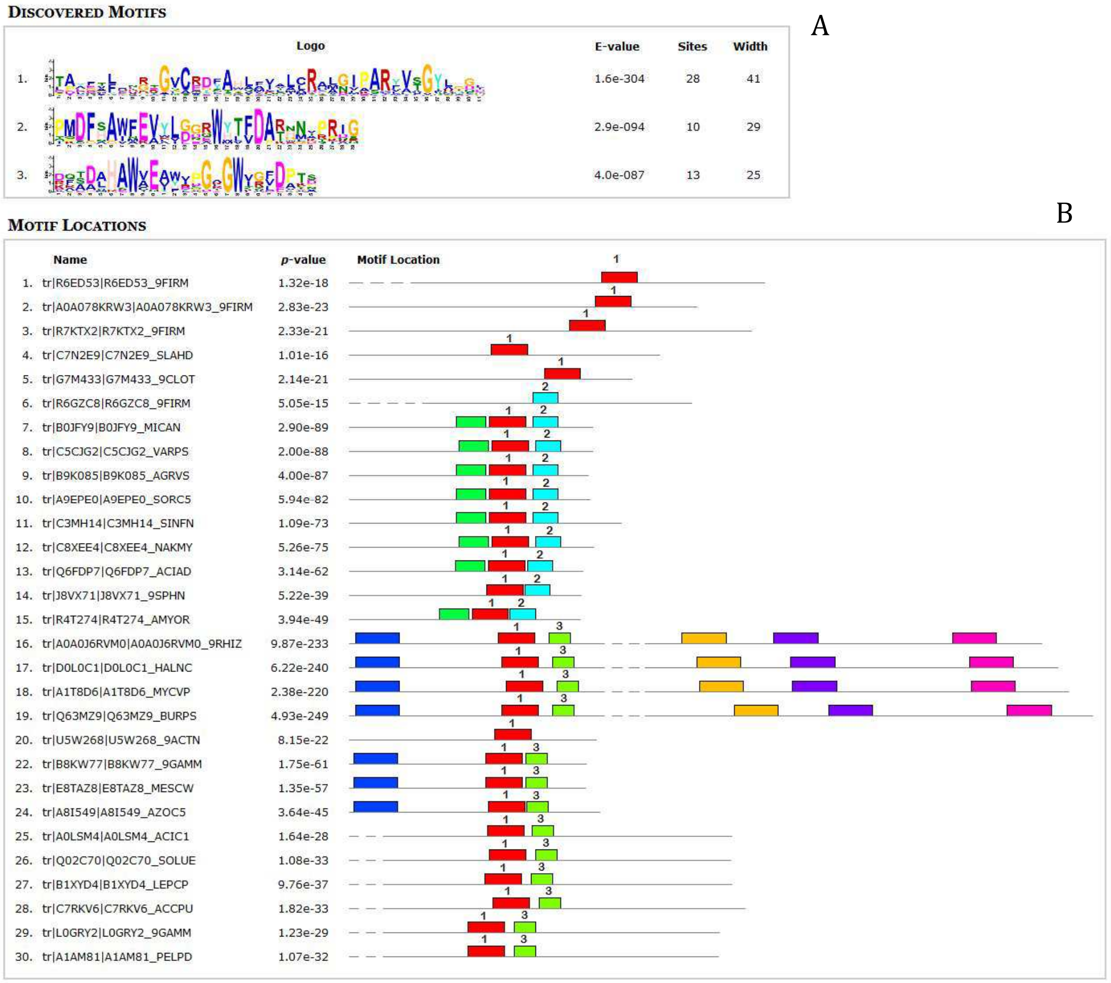
	Name	*p*-value	Motif Location
1.	tr|R6ED53|R6ED53_9FIRM	1.57e-53	
2.	tr|A0A078KRW3|A0A078KRW3_9FIRM	2.23e-65	
3.	tr|R7KTX2|R7KTX2_9FIRM	1.56e-61	
4.	tr|C7N2E9|C7N2E9_SLAHD	6.76e-101	
5.	tr|G7M433|G7M433_9CLOT	2.68e-94	
6.	tr|R6GZC8|R6GZC8_9FIRM	5.24e-62	

***Fig.23:* Sequence motifs detected in TGases from Firmicutes of group Va**. **A**: the consensus sequences of the three best motifs are reported with related statistical parameters. **B**: Motifs locations within sequences are shown. The colored boxes indicate the position within the amino acid sequences of the three best motifs (indicated by number, 1,2, 3) together with other motifs identified by MEME, having lower E-value and not conserved in all sequences.

Motif 1 shows an E-value equal to 2.2e^{-044} and is composed of 22 amino acids; this motif includes two residues of the potential catalytic triad (i.e. histidine, H and aspartic acid, D) and the amino acids closely related to them forming the consensus H-A-W-N-X-V-X(2)-[DG]-G-[KE]-[WT]-X(2)-[VL]-D-X-T. Motif 2 is the widest, is composed of 41 amino acids, has E-value equal to 2.8e^{-026}, and includes the other catalytic residue (Cysteine, C), with the consensus [AGI]-[VQ]-C-X-[GS]-[YI]-[AS]. Motif 3 has E-value equal to 8.0e^{-008} and is composed of 21 amino acids; it is located upstream from the first catalytic residue.

Two additional sequences belong to this branch of the tree, one from *Fluviicola Taffensis* (UNIPROT code: F2IIT1) belonging to the Flavobacteria phylum and one from *Cyanobacterium stanieri* (UNIPROT code: K9YNV7) belonging to the Cyanobacteria phylum. The analysis of motifs with these two additional sequences indicates that they share motifs 1 and 2, containing the three catalytic amino acids, but not the motif 3 (data not shown). Moreover, the catalytic aspartic acid in motif 1 is not conserved in these two sequences.

Analysis of motifs in group II and V, as discussed in the previous paragraphs, shows some heterogeneity, probably due to the high number of sequences that cluster in these groups and their probable sub-group organization. In these groups it is also observed variability on the potential catalytic amino acids. It is interesting to note that both II and V group include sequences of Firmicutes, that are also in group Va. By analyzing sequences of Firmicutes from group Va together with those from group II, the search does not find significant motifs including the potential catalytic residues (not shown). This result suggests that different TGases could exist within the same phylum. On the other hand, the analysis of Firmicutes sequences from group Va together with sequences from group V (*Fig.24*) identified a strongly significant motif that has a consensus similar to the one shown by the motif 2 in *fig.23A* (observed for group Va alone) and includes a conserved cysteine residue (motif 1 in *fig.24A*). This motif is conserved in all sequences (red motif in *fig.24B*). This suggests that the conserved cysteine residue may be the catalytic Cysteine, and that motif is conserved across the two groups. The sequences from group Va (in the top of *fig.24B*) do not present other motifs when analyzed together with sequences of group V, thus suggesting that motif 1 is the most conserved during evolution. Moreover, the analysis finds that all sequences from group V share motifs 2 or 3 (*Fig.24A* and *B*) at the C-terminal side of motif 1. Both motifs 2 and 3 present conserved Histidine and Aspartate residues and appear similar to Motif 1 of the analysis on Firmicutes only (*Fig.23A*). This finding suggests that group Va and group V, although clustered not contiguously, share significant similarities in the region containing the Cysteine residue potentially part of the catalytic mechanism.



DISCOVERED MOTIFS

	Logo	E-value	Sites	Width
1.		1.6e-304	28	41
2.		2.9e-094	10	29
3.		4.0e-087	13	25

MOTIF LOCATIONS

	Name	*p*-value	Motif Location
1.	tr|R6ED53|R6ED53_9FIRM	1.32e-18	
2.	tr|A0A078KRW3|A0A078KRW3_9FIRM	2.83e-23	
3.	tr|R7KTX2|R7KTX2_9FIRM	2.33e-21	
4.	tr|C7N2E9|C7N2E9_SLAHD	1.01e-16	
5.	tr|G7M433|G7M433_9CLOT	2.14e-21	
6.	tr|R6GZC8|R6GZC8_9FIRM	5.05e-15	
7.	tr|B0JFY9|B0JFY9_MICAN	2.90e-89	
8.	tr|C5CJG2|C5CJG2_VARPS	2.00e-88	
9.	tr|B9K085|B9K085_AGRVS	4.00e-87	
10.	tr|A9EPE0|A9EPE0_SORC5	5.94e-82	
11.	tr|C3MH14|C3MH14_SINFN	1.09e-73	
12.	tr|C8XEE4|C8XEE4_NAKMY	5.26e-75	
13.	tr|Q6FDP7|Q6FDP7_ACIAD	3.14e-62	
14.	tr|J8VX71|J8VX71_9SPHN	5.22e-39	
15.	tr|R4T274|R4T274_AMYOR	3.94e-49	
16.	tr|A0A0J6RVM0|A0A0J6RVM0_9RHIZ	9.87e-233	
17.	tr|D0L0C1|D0L0C1_HALNC	6.22e-240	
18.	tr|A1T8D6|A1T8D6_MYCVP	2.38e-220	
19.	tr|Q63MZ9|Q63MZ9_BURPS	4.93e-249	
20.	tr|U5W268|U5W268_9ACTN	8.15e-22	
22.	tr|B8KW77|B8KW77_9GAMM	1.75e-61	
23.	tr|E8TAZ8|E8TAZ8_MESCW	1.35e-57	
24.	tr|A8I549|A8I549_AZOC5	3.64e-45	
25.	tr|A0LSM4|A0LSM4_ACIC1	1.64e-28	
26.	tr|Q02C70|Q02C70_SOLUE	1.08e-33	
27.	tr|B1XYD4|B1XYD4_LEPCP	9.76e-37	
28.	tr|C7RKV6|C7RKV6_ACCPU	1.82e-33	
29.	tr|L0GRY2|L0GRY2_9GAMM	1.23e-29	
30.	tr|A1AM81|A1AM81_PELPD	1.07e-32	

***Fig.24:* Motifs detected by the analysis of selected Firmicutes sequences from group Va and group V. A**: the consensus sequences of the three best motifs are reported with related statistical parameters. **B**: Motifs locations within sequences are shown. The colored segments indicate the position within the amino acid sequences of the three best motifs (indicated by number, 1,2, 3) together with other motifs identified by MEME, not conserved in all sequences.

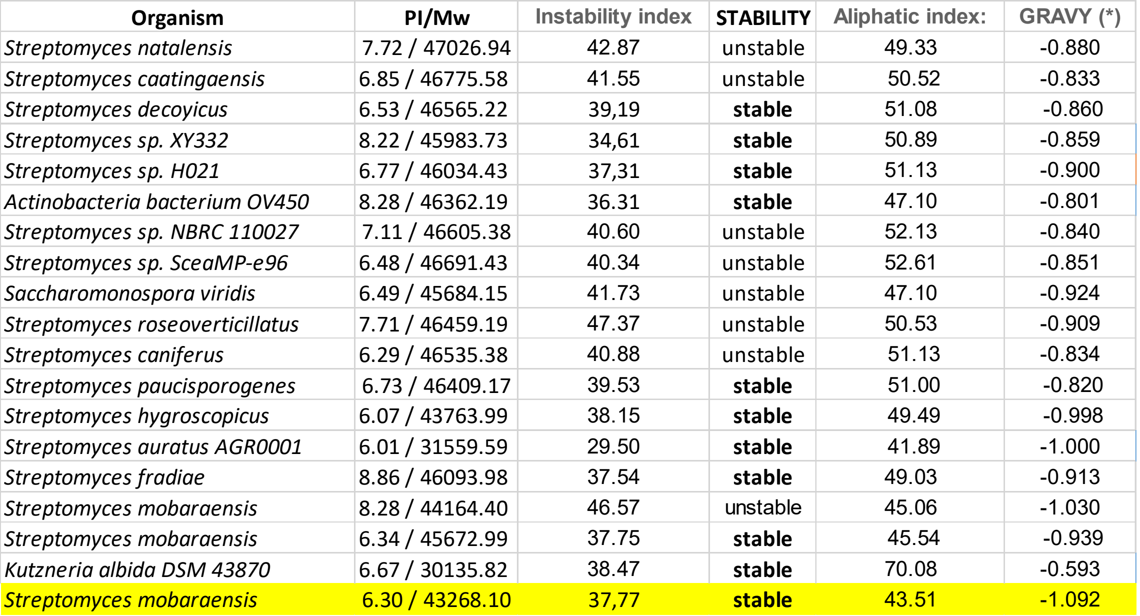
In conclusion from this massive sequence analysis it is possible to say that the presence of specific motifs in each group corroborate the classification of the mTGases in at least five main distinct groups. Moreover, the case of the group Va shows as different mTGases could exist within the same phylum, indicating that the species identity is not an index of how this enzyme has evolved.

8. Structural analyses

At the end of this classification, it was possible to have the basis for starting the model constructions of those sequences which show the best features also in terms of the source organism. For each organism, in fact, it has been verified the presence of a QPS certification by EFSA, if some of them are producers of enzyme already used in the food industry, or, for the organisms not judge as unable to obtain the QPS, simply if they are safe and procurable. Thus, for each group, proteins having the best features in terms of sequence features, predicted stability and source organism have been selected for the modelling procedure. However, for proteins belonging to group I and IV, because no suitable template was found, it wasn't possible to make a structural prediction. For the protein modelled, instead, an analysis of their secondary structure was done and a specific evaluation of their active site pocket was performed.

8.1 Protein 3D models and their validation

From all the evaluations mentioned before, it has been possible to select many hypothetical mTGase sequences, 19 of them are very similar to MTGase, from them it has been selected just 9 because by means of ProtParam Expasy tool the remaining 10 were judge as instable. From these 9 sequences, those from *Streptomyces sp., Actinobacteria bacterium OV450*, *S. Auratus*, and *S. fradiae* have stability index calculated on the basis of the sequences better than MTGase. A higher stability index is instead found for the protein sequences belonging to *S. decoyicus*, *S. paucisporogenes*, *S. hygroscopicus* and to *Kutzneria albida* (Table.1).

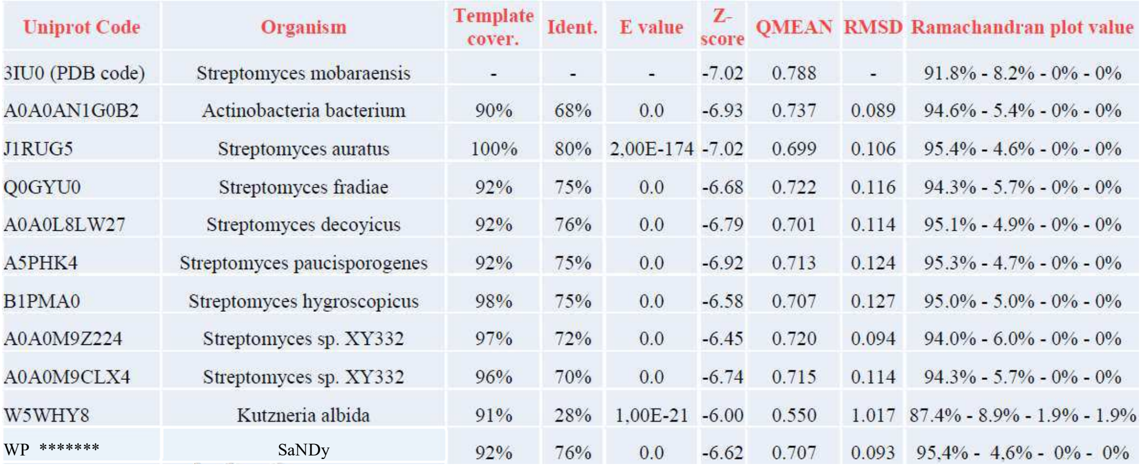


Organism	PI/Mw	Instability index	STABILITY	Aliphatic index:	GRAVY (*)
Streptomyces natalensis	7.72 / 47026.94	42.87	unstable	49.33	-0.880
Streptomyces caatingaensis	6.85 / 46775.58	41.55	unstable	50.52	-0.833
Streptomyces decoyicus	6.53 / 46565.22	39,19	**stable**	51.08	-0.860
Streptomyces sp. XY332	8.22 / 45983.73	34,61	**stable**	50.89	-0.859
Streptomyces sp. H021	6.77 / 46034.43	37,31	**stable**	51.13	-0.900
Actinobacteria bacterium OV450	8.28 / 46362.19	36.31	**stable**	47.10	-0.801
Streptomyces sp. NBRC 110027	7.11 / 46605.38	40.60	unstable	52.13	-0.840
Streptomyces sp. SceaMP-e96	6.48 / 46691.43	40.34	unstable	52.61	-0.851
Saccharomonospora viridis	6.49 / 45684.15	41.73	unstable	47.10	-0.924
Streptomyces roseoverticillatus	7.71 / 46459.19	47.37	unstable	50.53	-0.909
Streptomyces caniferus	6.29 / 46535.38	40.88	unstable	51.13	-0.834
Streptomyces paucisporogenes	6.73 / 46409.17	39.53	**stable**	51.00	-0.820
Streptomyces hygroscopicus	6.07 / 43763.99	38.15	**stable**	49.49	-0.998
Streptomyces auratus AGR0001	6.01 / 31559.59	29.50	**stable**	41.89	-1.000
Streptomyces fradiae	8.86 / 46093.98	37.54	**stable**	49.03	-0.913
Streptomyces mobaraensis	8.28 / 44164.40	46.57	unstable	45.06	-1.030
Streptomyces mobaraensis	6.34 / 45672.99	37.75	**stable**	45.54	-0.939
Kutzneria albida DSM 43870	6.67 / 30135.82	38.47	**stable**	70.08	-0.593
Streptomyces mobaraensis	6.30 / 43268.10	37,77	**stable**	43.51	-1.092

Table.1 shows the **property of the prediction of the stability index calculated for each sequence.** Stability index values, prediction of stability, predicted aliphatic index and grand average of hydropathicity (GRAVY) are shown. In yellow the reference parameters of MTGase.

For all these protein sequences, it was possible to create some models using the homology modelling approach. In fact, as results of the alignments performed by BLAST tool, the sequence of MTGase from *S. mobarensis* (PDB structure 3IU0) has high similarity with these sequences, in detail: 90% of query coverage, 68% of identity and E-value equal to 0.0 with protein sequence from *Actinobacteria bacterium OV450* (UniProt Code: A0A0AN1G0B2); 100% of query coverage, 80% of identity, and E-value of 2e^{-174} with protein sequence from *S. Auratus* (UniProt Code: J1RUG5); query coverage of 92%, identity of 75%, and E-value equal to 0.0 with the protein sequence from *S. Fradiae* (UniProt Code: Q0GYU0); 98% of query coverage, 75% of identity, and E-value equal to 0.0 with the protein sequence from *S. hygroscopicus* (UniProt Code: B1PMA0); 92% of query coverage, 75% of identity, E-value equal to 0.0 with the protein sequence from *S. paucisporogenes*, (UniProt Code:A5PHK4), and very similar results with the protein sequence from S. *decoyicus* (UniProt Code:A0A0L8LW27), differing only for the percentage of identity of 76%. Similar values have been obtained also for the protein sequences extracted by *Streptomyces sp*. In this last case two models have been constructed one for the protein sequence extracted by *Streptomyces sp. H021* (UniProt Code: A0A0M9Z224) and the other for the protein sequence extracted by *Streptomyces sp. XY332*, (UniProt Code: A0A0M9CLX4) because of the significant difference in the predicted isoelectric point of these sequences. The most recent alignment of MTGase with the hypothetical protein-glutamine gamma-glutamyltransferase from *SaNDy* (organism not disclosed for patent opportunity) available since 2018 shows high similarity too, in detail: 92% of query coverage, 76% of identity and E-value equal to 0.0 (the model obtained was selected for experimental characterizations that are still ongoing; see paragraph 10.2).

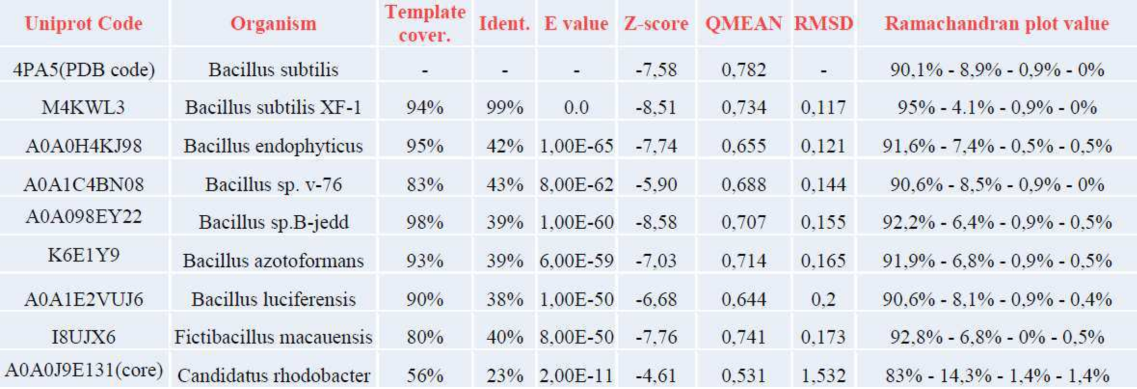
For each selected sequence, using MTGase as template, several models have been generated and all the models have been assessed, by the analysis of the structure built by Modeller 9.18 and the evaluation of their Ramachandran plot, their local and global energy, their QMEAN value. By means of this analysis, it was possible to select for each sequence the best model. The following Table 2 shows the property of the best models.



Uniprot Code	Organism	Template cover.	Ident.	E value	Z-score	QMEAN	RMSD	Ramachandran plot value
3IU0 (PDB code)	Streptomyces mobaraensis	-	-	-	-7.02	0.788	-	91.8% - 8.2% - 0% - 0%
A0A0AN1G0B2	Actinobacteria bacterium	90%	68%	0.0	-6.93	0.737	0.089	94.6% - 5.4% - 0% - 0%
J1RUG5	Streptomyces auratus	100%	80%	2,00E-174	-7.02	0.699	0.106	95.4% - 4.6% - 0% - 0%
Q0GYU0	Streptomyces fradiae	92%	75%	0.0	-6.68	0.722	0.116	94.3% - 5.7% - 0% - 0%
A0A0L8LW27	Streptomyces decoyicus	92%	76%	0.0	-6.79	0.701	0.114	95.1% - 4.9% - 0% - 0%
A5PHK4	Streptomyces paucisporogenes	92%	75%	0.0	-6.92	0.713	0.124	95.3% - 4.7% - 0% - 0%
B1PMA0	Streptomyces hygroscopicus	98%	75%	0.0	-6.58	0.707	0.127	95.0% - 5.0% - 0% - 0%
A0A0M9Z224	Streptomyces sp. XY332	97%	72%	0.0	-6.45	0.720	0.094	94.0% - 6.0% - 0% - 0%
A0A0M9CLX4	Streptomyces sp. XY332	96%	70%	0.0	-6.74	0.715	0.114	94.3% - 5.7% - 0% - 0%
W5WHY8	Kutzneria albida	91%	28%	1,00E-21	-6.00	0.550	1.017	87.4% - 8.9% - 1.9% - 1.9%
WP *******	SaNDy	92%	76%	0.0	-6.62	0.707	0.093	95,4% - 4,6% - 0% - 0%

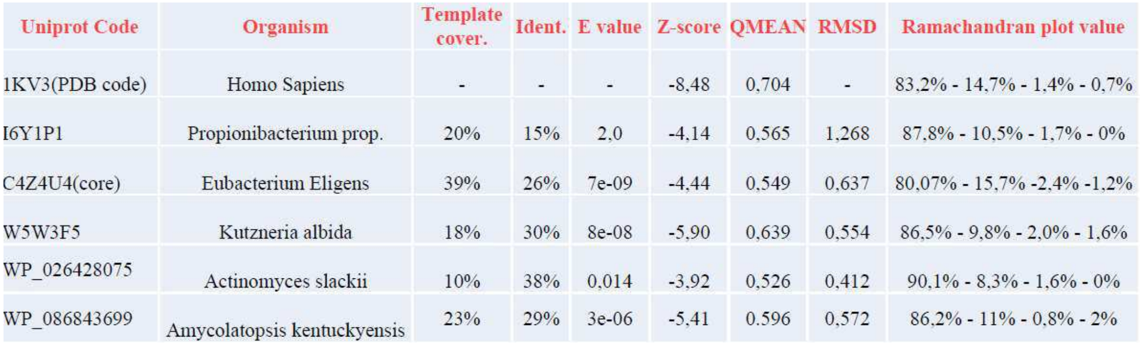
Table.2 shows the **property of the best models of proteins belonging to group III.** Z-score, QMEAN and Ramachandran plot values are used for evaluating the quality of the models. RMSD and Z-score from the template structure are also shown for comparison.

Similarly, also the protein sequences belonging to the second group have been modelled. In this case the template used for the homology modelling procedure was the TGl (PDB code: 4PA5). Quality of the models obtained was affected by the different percentages of identity, thus, the higher is the percentage of identity between sequence and template, the better is expected to be the model obtained. However, in the case of the protein sequence extracted from *Candidatus rhodobacter*, also due to low sequence coverage, it was possible to model only the active site. All the protein sequences modelled, and the features of the best models obtained are summarized in Table 3. In the case of the protein sequences belonging to the fifth group, the modelling procedures have been more complex due to the very low percentage of identity and the very low sequence coverage. In this case the protein chosen as template was the hTGase2 (PDB code:1KV3) but this sequence was able to cover just the catalytic domain of the sequence to model. Therefore, several different templates have been used to model the whole protein sequences, following the procedure discussed in the paragraph 4.2. A summary concerning the proteins that have been modelled and the features of the obtained models is shown in Table 4.



Uniprot Code	Organism	Template cover.	Ident.	E value	Z-score	QMEAN	RMSD	Ramachandran plot value
4PA5(PDB code)	Bacillus subtilis	-	-	-	-7,58	0,782	-	90,1% - 8,9% - 0,9% - 0%
M4KWL3	Bacillus subtilis XF-1	94%	99%	0.0	-8,51	0,734	0,117	95% - 4.1% - 0,9% - 0%
A0A0H4KJ98	Bacillus endophyticus	95%	42%	1,00E-65	-7,74	0,655	0,121	91,6% - 7,4% - 0,5% - 0,5%
A0A1C4BN08	Bacillus sp. v-76	83%	43%	8,00E-62	-5,90	0,688	0,144	90,6% - 8,5% - 0,9% - 0%
A0A098EY22	Bacillus sp.B-jedd	98%	39%	1,00E-60	-8,58	0,707	0,155	92,2% - 6,4% - 0,9% - 0,5%
K6E1Y9	Bacillus azotoformans	93%	39%	6,00E-59	-7,03	0,714	0,165	91,9% - 6,8% - 0,9% - 0,5%
A0A1E2VUJ6	Bacillus luciferensis	90%	38%	1,00E-50	-6,68	0,644	0,2	90,6% - 8,1% - 0,9% - 0,4%
I8UJX6	Fictibacillus macauensis	80%	40%	8,00E-50	-7,76	0,741	0,173	92,8% - 6,8% - 0% - 0,5%
A0A0J9E131(core)	Candidatus rhodobacter	56%	23%	2,00E-11	-4,61	0,531	1,532	83% - 14,3% - 1,4% - 1,4%

Table.3 shows the **property of the best models of proteins belonging to group II.** Z-score, QMEAN and Ramachandran plot values are used for evaluating the quality of the models. RMSD and Z-score from the template structure are also shown for comparison.



Uniprot Code	Organism	Template cover.	Ident.	E value	Z-score	QMEAN	RMSD	Ramachandran plot value
1KV3(PDB code)	Homo Sapiens	-	-	-	-8,48	0,704	-	83,2% - 14,7% - 1,4% - 0,7%
I6Y1P1	Propionibacterium prop.	20%	15%	2,0	-4,14	0,565	1,268	87,8% - 10,5% - 1,7% - 0%
C4Z4U4(core)	Eubacterium Eligens	39%	26%	7e-09	-4,44	0,549	0,637	80,07% - 15,7% -2,4% -1,2%
W5W3F5	Kutzneria albida	18%	30%	8e-08	-5,90	0,639	0,554	86,5% - 9,8% - 2,0% - 1,6%
WP_026428075	Actinomyces slackii	10%	38%	0,014	-3,92	0,526	0,412	90,1% - 8,3% - 1,6% - 0%
WP_086843699	Amycolatopsis kentuckyensis	23%	29%	3e-06	-5,41	0.596	0,572	86,2% - 11% - 0,8% - 2%

Table.4 shows the **property of the best models of proteins belonging to group V**. Z-score, QMEAN and Ramachandran plot values are used for evaluating the quality of the models. RMSD and Z-score from the template structure are also shown for comparison.

Among all the models built one was particularly interesting, i.e. the model of the uncharacterized protein from *Kutzneria Albida* (Uniprot code: W5WHY8). In fact, from the alignment between this sequence and MTGase*,* the query coverage is of 91%, the E-value of the alignment is equal to 1e^{-21}, but the percentage of identity is only equal to 28. With a low level of sequence identity, if homology between the target protein and the template protein can be assumed, it is possible to apply the homology modelling, but a careful procedure must be applied in order to obtain the optimal alignment of the two sequences. So, in this case it was necessary to predict the secondary structure for the sequence from *Kutznerya Albida*, compare it to the secondary structure of the 3IU0 structure, modify the sequence alignment according to the conformity of the secondary structures, and then perform the homology modelling by Modeller. After many trials, it was possible to obtain a model that presents 4 α-helices less compared to 3IU0 structure, but this was justified by the fact that this sequence is shorter than MTGase, and most important thing, this change does not interfere with the structural part involved in the active site, that is instead very well preserved. Also in this case, the models obtained have been assessed by means of ProsaWeb, ProCheck and QMEAN servers. The best model that has been chosen show a Z-score equal to -6, a QMEAN of 0.55, and from the Ramachandran plot it is possible to see that 87.4% of the torsion angles are in the low-energy regions, 8.9% in the allowed regions, 1.9% in the generously allowed regions and 1.9% in the disallowed regions. The value in the disallowed regions is however related to 5 amino acids, 4 located in the two different random coil regions (Glu96 – Ser99 - Ala 184 - His185) and 1 in the N-terminal part of the protein (Ala44). The RMSD of this model calculated against the template 3IU0 is equal to 1.017. This was an expected value because the built model shows 7 beta-strands and 7 alpha-helices (see *fig.25* – left panel), while the template 3IU0 is composed by 7 beta-strands and 11 alpha-helices. However, even if there is a significant difference related to the length of these sequences and so also in the number of the alpha-helices present in the structure, the features of the active site are not modified (see *fig.25* – right panel).

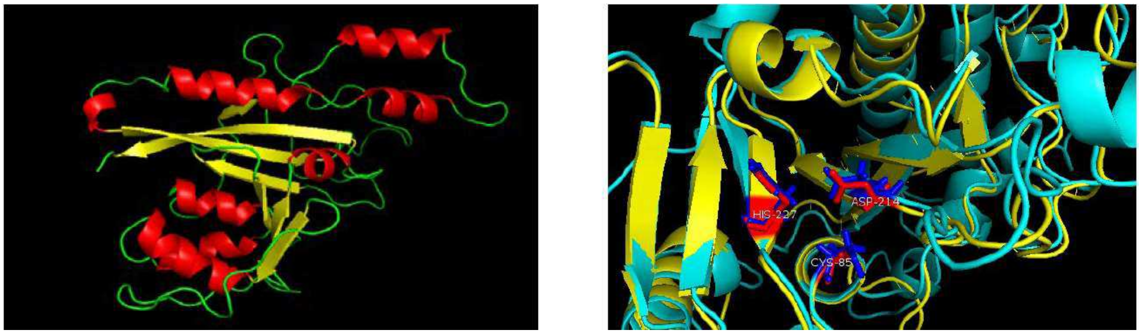


Fig.25: **model of the hypothetical mTGase from *K. albida*. Left panel**: shows the structure of the model create for the protein from *Kutznerya Albida* (highlighted in red the alpha helices and in yellow the beta strands). **Right panel**: shows the structural overimposition of the protein from *Kutzneria Albida* (in cyan) and the template 3IU0 (in yellow) from *S. mobaraensis*. The catalytic residues are highlighted as blue and red sticks for the protein from *Kutznerya* and of 3IU0, respectively.

After the building of this protein before starting the experimental characterization of this so particular protein, it was chosen to perform further model assessments by means of MD simulations in order to strongly validate the stability of this uncharacterized protein before starting a possible protein expression in bacteria; but this topic will be dealt in paragraph 9.1.

8.2 Secondary structure analysis

At the end of the modelling and validation procedure, the next step of the analysis performed was the investigation of the secondary structure topologies and their preservation along the phylogenetic tree. Therefore, models obtained have been structurally analyzed and compared with the structures already present in PDB. The comparison has followed the criteria of clustering adopted until now, i.e. each transglutaminase present in the PDB has been compared to the models of the corresponding group, based on the features of its amino acidic sequence.
Here are reported three examples of the secondary structure detected, one for each group of proteins modelled:

Example of protein topology – Group II

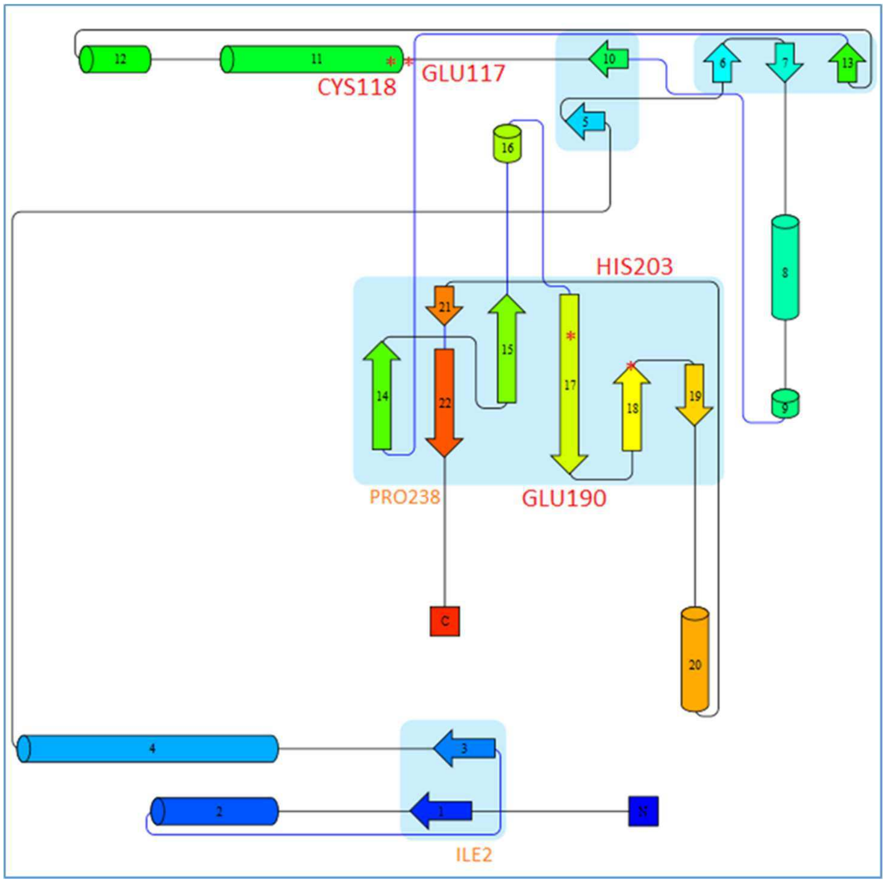


Fig.26: **Secondary structure topology of protein belonging to the group II**. Here is represented in cartoons the model of mTGase TGl-like from *Fictibacillus macauensis* (model no.7 in table3) similar to the *Bacillus Subtilis* TGl. The location of the catalytic amino acids (highlighted in red) is tagged by a star.

Example of protein topology – Group III

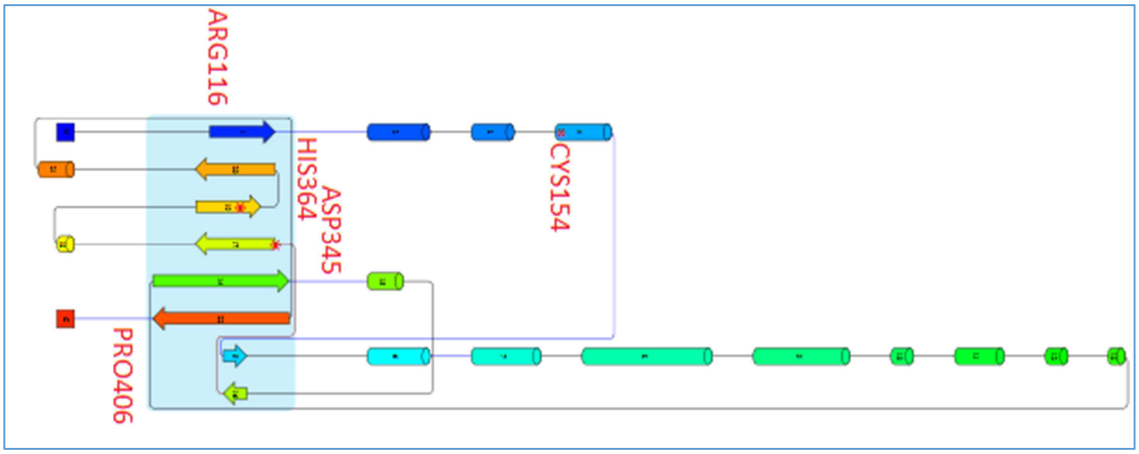


Fig.28: **Secondary structure topology of protein belonging to the group III**. Here is represented in cartoons the model of mTGase from *Actinobacteria Bacterium* (model no.1 in table 2) similar to *S. Mobaraensis* MTGase. The location of the catalytic amino acids (highlighted in red) is tagged by stars.

Example of protein topology – Group V

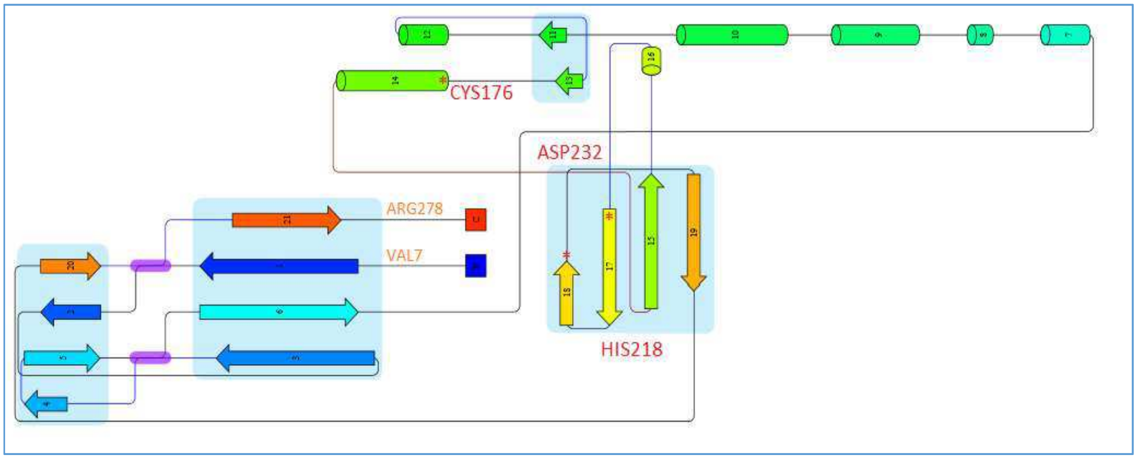


Fig.27: **Secondary structure topology of protein belonging to the group V**. Here is represented in cartoons the model of mTGase hTG2-like from *Kutzneria albida* (model no.3 in table4) similar to the mammalian hTGase2. The location of the catalytic amino acids (highlighted in red) is tagged by stars.

The comparison of the secondary structure of the catalytic core of the three groups (i.e., groups II, III, V) evidenced that the first catalytic residue (C/S) is always at the N terminal of an α-helix linked to a beta-sheet of at least three antiparallel β-strands, with the second and third strand hosting each another catalytic residue. The two catalytic regions, i.e. the α-helix and the β-sheet, are differently linked in the three groups. For group II (*Fig.26*), the linker is composed by few α-helices and several β-strands. For group III (*Fig.28*), the two regions are linked by a very long portion composed by eight α-helices. For group V (*Fig.27*), the catalytic helix may be linked directly to the catalytic β-sheet or by means of one additional β-strand. The topology of the three groups could suggest the hypothesis that the mTGase human-like (group V) and Streptomyces-like (group III), arise from a Firmicutes (group II) ancient form of TG-like. Moreover, looking at the little spread of the form of mTGase Streptomyces-like along the different phylogenetic phyla, it is possible to think that this is a less-specialized form of TGase, that has evolved, perhaps starting from the ancient form previous mentioned, in a different way than the one human-like (Giordano and Facchiano, 2018).

These observations open new views on the evolution of TGase superfamily and may help the comprehension of the structural properties (at least for the active site), also of the proteins belonging to the other groups, for which no structure or model is available, yet.

8.3 Analysis of the catalytic site pocket

The mTGase models have been investigated to observe in more detail the potential catalytic sites and compare them to the experimental structure of active forms (see Table 5).

The catalytic triad is conserved in all models for group III, while in some example of groups II and V it is modified in one or two amino acids. In all cases but one, the three amino acids are well overlapped in the models, thus indicating that the structure allows their correct geometry for putative catalytic function. In groups III and II, RMSD values for the overlap of the triad are always very low, in most cases lower than 0.1. Another analysis performed concerns the volume of the cavity that includes the triad, detected with two measures to take into account different depth. In comparison to the cavity of the related template, cavity volumes appear in some cases very similar, thus suggesting the possibility of similar activity (see *S. decoyicus* and *S. paucisporogenes* in group III). In other cases, the different volumes of the cavity suggest different selectivity for substrate, or inactivity (see the case of W5WHY8 from *K.albida*, for which experimental tests have demonstrated a strong specificity; Steffen et al., 2017). For group V, there are relevant divergences from the template features. For *K. albida* mTGase hTgase2-like, the overlap to catalytic triad of the template appears not possible. Deviations of the different parameters from the template suggest either the absence of function similarity with the template enzyme, i.e. the hTGase2, or difficulties to create correct models.

Even if these results could be merely indicative, it is possible to suppose that to smaller active site pockets could correspond more selective (or inactive) enzyme, and vice versa. For those proteins, whose active site pocket volume is similar to the one of the templates, it is instead possible to hypothesize a similar activity (Giordano and Facchiano,2018).

Uniprot Code	Organism	Preserved catalytic site	Overlapping catalytic site	RMSD catalytic site	Volume	Volume depth
Group II						
4PA5(PDB code)	Bacillus subtilis	EC--E--H	-	-	193	506
M4KWL3	Bacillus subtilis XF-1	YES	YES	0,252	162	508
A0A0H4KJ98	Bacillus endophyticus	YES	YES	0,277	227	672
A0A1C4BN08	Bacillus sp. v-76	YES	YES	0,098	238	634
A0A098EY22	Bacillus sp.B-jedd	NO (E→L)	YES	0,038	126	338
K6E1Y9	Bacillus azotoformans	NO (EC→DC) (E→V)	YES	0,058	274	684
A0A1E2VUJ6	Bacillus luciferensis	YES	YES	0,067	155	414
I8UJX6	Fictibacillus macauensis	YES	YES	0,075	265	705
A0A0J9E131(core)	Candidatus rhodobacter	NO (E→W)	YES	0,262	126 (c)	282 (c)

Group III						
3IU0 (PDB code)	Streptomyces mobaraensis	C--D--H	-	-	213	557
A0A0AN1G0B2	Actinobacteria bacterium	YES	YES	0,075	167	412
J1RUG5	Streptomyces auratus	YES	YES	0,074	133	331
Q0GYU0	Streptomyces fradiae	YES	YES	0,084	159	413
A0A0L8LW27	Streptomyces decoyicus	YES	YES	0,076	211	549
A5PHK4	Streptomyces paucisporogenes	YES	YES	0,084	208	562
B1PMA0	Streptomyces hygroscopicus	YES	YES	0,105	176	489
A0A0M9Z224	Streptomyces sp. H021	YES	YES	0,064	60	136
A0A0M9CLX4	Streptomyces sp. XY332	YES	YES	0,067	310 (a)	924 (a)
W5WHY8	Kutzneria albida	YES	YES	0,084/ 0.308	33+46 (b)	77+116 (b)
Group V						
1KV3(PDBcode)	Homo Sapiens	C--H--D	-	-	43	105
I6Y1P1	Propionibacterium prop.	NO D→E	YES	0,908	22+44 (b)	113+116 (b)
C4Z4U4(core)	Eubacterium Eligens	YES	YES	0,364	36	86
W5W3F5	Kutzneria albida	YES	NO	2,316	93	253
WP_026428075	Actinomyces slackii	NO D→E	YES	0,339	75	202
WP_086843699	Amycolatopsis kentuckyensis	YES	YES	0,251	123/ 49+28 (b)	311/ 128+64 (b)

***Table.5* Structural features of the putative catalytic site of modelled TGases, compared to those of related templates.** Sequences from groups II, III, and V, investigated by molecular modelling techniques, with structural features of the catalytic site. For each group, the related template structure is shown on the first line. The (a), (b), (c) labels indicate that:
(a) The cavity expands beyond the catalytic triad, so the value maybe overestimated.
(b) In some case, two cavities have been identified, so two values are reported.
(c) Default parameters did not identify a cavity, so it was needed to modify the standard parameters, and results may be underestimated.

9. Molecular dynamics (MD) simulations

As several time mentioned, the principal aim of this project is the structural and the functional characterization of novel forms of mTGase. Because of that, after the massive clustering of all the sequences annotated as possible mTGase, some of them have been selected, modelled and their structure investigated, in order to make hypothesis also on their function an accurate analysis of the active site was performed also by computational analysis of the catalytic pocket. However, to find a novel form of mTGase that could be an alternative to that in use, experimental analyses and comparative studies with the MTGase are necessary. Thus, from all the mTGase analyzed until now two models have been selected for further investigations: the model of the uncharacterized protein from *K. albida* (UniProt code: W5WHY6) and the other from *SaNDy*. The latter due to its high percentage of identity with MTGase, together with its more recent discovery (the genome of this organism have been sequenced only in 2018) became directly object of experimentally study. The model of the hypothetical mTGase from *K. albida*, instead, was object of a deeper assessment by MD simulations in order to test its stability and its truthfulness. However, at the end of 2017 when the structure of the KalbTGase was discovered, the continuation of this activity became unnecessary, so MD simulations started to be performed directly on the crystallographic structure in order to compare its results with the one obtained from MTGase MD simulations. In this way structural features as stability in different conditions, flexible regions, conformational change and active site property could be analyzed and compared.

9.1 Kutzneria Albida mTGase: model validation

The first MD simulation at 300K in NPT performed on MTGase structure (PDB code 3IU0) lasted 130 ns and showed immediately that this protein was very stable. Actually, RMSD values were very low, and no conformational change was observed. In the meantime, a first simulation, lasting 278ns in NPT at 300K, on *Kutzneria Albida* mTGase model was performed. From the results the model seemed to be stable (*Fig.29*). In fact, even if there was a peak in the RMSD, this was related to a conformational change of a helix that turns upside down, moreover, the position of this helix was not close to the active site and did not interest anymore it, which instead was very stable. These results could suggest a rearrangement of the model due to a not perfect construction during the modelling phase or just an adaptation to the conditions of the simulation. To understand if this problem could be related to the model, other MD simulations have been done but the structures used for starting these novel simulations were taken extracting some specific frames from the first simulation (for more detail see paragraph 5.1), this in order to check if the developing into a stabilized system in the first simulation was just by chance or if that model could be judge as really stable.

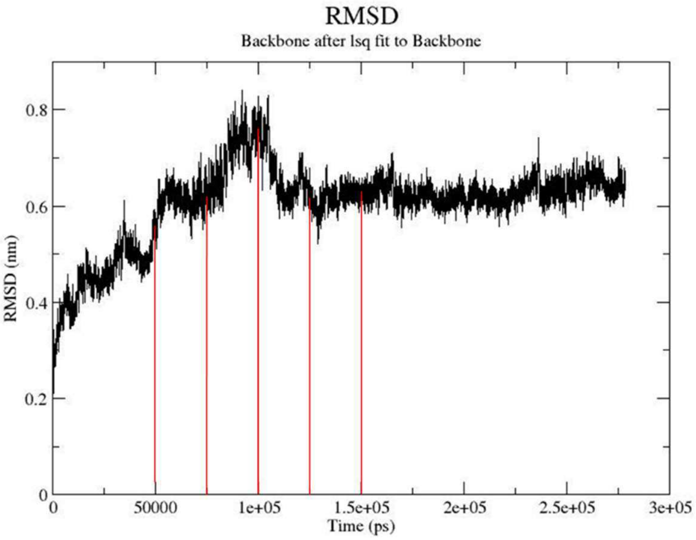


***Fig.29*: RMSD of MD simulation on mTGase from *K. albida*.** RMSD related to the backbone after 275ns of molecular dynamics simulation at 300K in NPT condition for the built model of the protein sequence extracted by *Kutznerya Albida*. The red lines indicate the time step in correspondence of which have been extracted the frames to restart the MD simulations.

After 230ns of MD simulation at 300K in NPT condition the model of the protein sequence extracted by *Kutznerya Albida* showed in all the simulation the same rearrangement of a specific peripherical α-helix, in the sequence region 86-121, showing the movement of this helix (pointed by the red arrow in *fig.30A*/*B*) and its fusion with the next one. Moreover, after this movement in all the simulation the system reached the stability (data not shown).

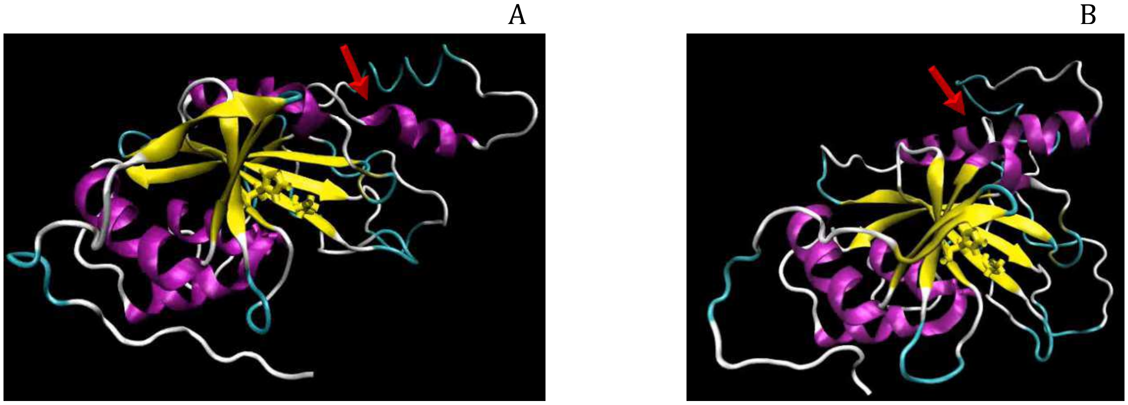


***Fig.30*: Structural evolution of mTGase model from *K. albida* after the MD simulation. A**: model structure at the beginning of the simulation. **B**: model structure at the end of the simulation. Data refers to the simulation starting from the frame caught at 75ns of the first simulation

These results suggested a rearrangement of the model that probably in that portion was not correctly folded, but additionally they suggested a stability of the hypothetical enzyme as well as a more suitable conformation.

However, after the publication of the discovery of a novel form of mTGase from *K. albida* by Roche Diagnostics GmbH (Steffen et al., 2017) it was possible to compare the real structure of this enzyme with the model built.

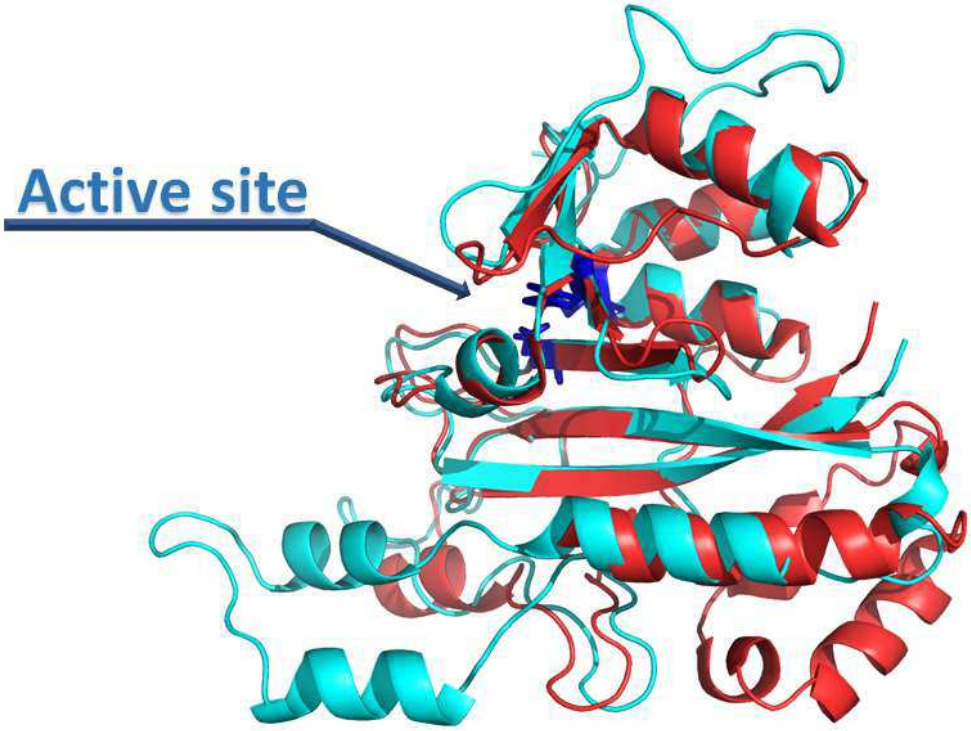


Fig.31**: Superimposition of Kalb mTGase model and its real crystal structure.** In cyan cartoon *Kutzneria Albida* mTGase model, in red cartoon the crystal structure of *Kutzneria Albida* mTGase, in dark blue sticks the amino acids of the active site.

From the superimposition of the two structures, it is possible to see as they are very similar except for the helices present in the regions 86-121. Unfortunately, even if from the alignment done the prediction of secondary structure was corrected, during the modelling procedure, the helices of that region folded in a different way; this explains the mismatch between the sequence and the model. However, the hypothesis that this uncharacterized protein could be a real mTGase, with different features than the MTGase, suggested by the good stability predicted, especially for the active site, has been confirmed by Roche's characterization of this enzyme (Steffen et al, 2017).

These results are a clear example of how powerful the MD could be: actually, not only MD simulations have represented a good method to corroborate or not the selection done, concerning which protein should be experimentally characterized, but they have also given specific information on the strengths and weaknesses of the model obtained.

9.2 KalbTGase and MTGase: MD simulations at 300K

KalbTGase is a very specific TGase for which many activity assays and tests to find substrate specificity and activity best conditions have been performed (Steffen et al., 2017). After its discovery, it was decided to perform MD simulations on both the crystal structure of KalbTGase (PDB code:5M6Q) and the crystal structure of the MTGase (PDBcodde:3IU0), in order to find the presence of similarity or differences between these two proteins also in order to explain their different specificity. Many MD simulations and analyses have been performed (see paragraphs 5.1), together with deeper analyses focused on the active site. Five simulations at 300K, each one lasted 300ns, have been performed, however from the **RMSD** plot (*Fig.32*), the five MD simulations seem to be very similar each other so it is possible to say that the evolution of the system during these different simulations seems to be the same.

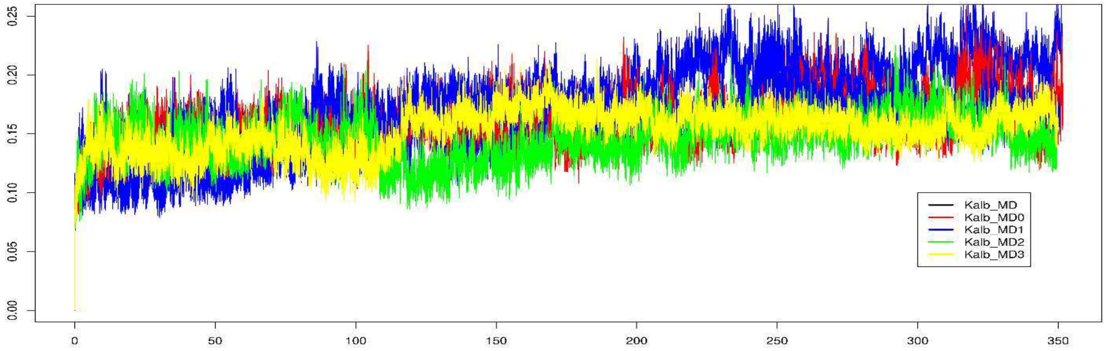


Fig.32: **RMSD plot of all the five different simulations of KalbTGase performed at 300K.** In the plot it is possible to see on the x-axis the duration of the simulation in ns, on the y-axis the RMSD values in nm.

Moreover, from all the RMSD plots it is possible to see that the system seems to be quite stable, in fact, no strong variation in the RMSD values is displayed and no rearrangement of the structure is visible.

An analogue result is found in the analysis of the **RMSF** and, because of that, in order to simplify the description, only one RMSF plot of the five made is shown. Moreover, for a more readily comprehensible of the plot just the RMSF plot related to the last 100ns of simulation is shown (*Fig.33*). This time frame is enough to show which are the most flexible regions and its analysis does not differ from the results obtained from the RMSF of the whole simulation.

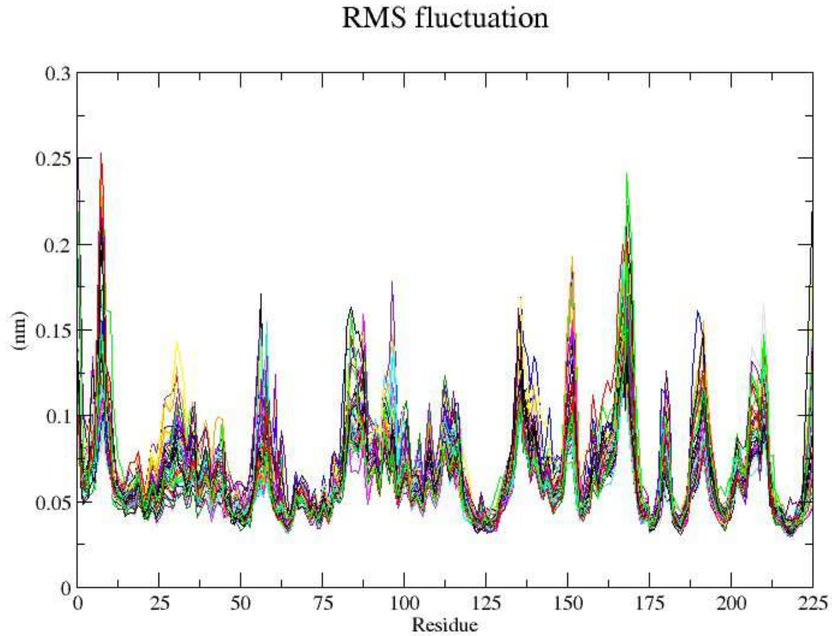


Fig.33: **RMSF plot of the last 100ns of MD simulation for KalbTGase at 300K**.
On the x-axis the residues, which form the protein, numbered consecutively, on the y-axis the RMSF values in nm. Each colored line corresponds to an analyzed time step of the simulation (see paragraph 5.2.a).

By the analysis of this plot *(Fig.33*) it is possible to see the most flexible parts of the protein and, in particular, which one of these parts has a lower, higher or medium level of flexibility (*Fig.34*). As displayed by the figure the most flexible regions are composed predominantly of the small peripherical loops. The core of the protein, instead, seems to be very stable, especially in the active site region, which show a particular rigidity.

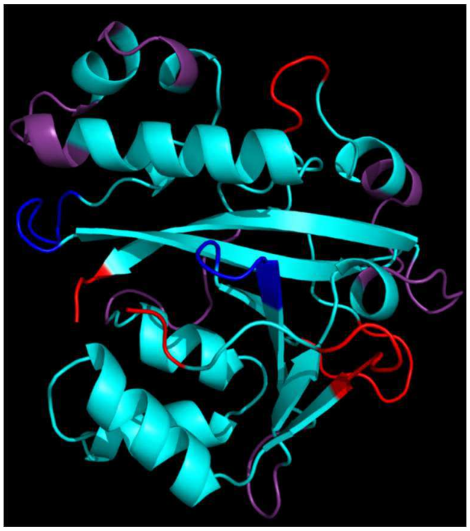


Fig.34: **Flexible regions of KalbTGase.** This figure shows the structure of KalbTGase: in red the regions with a higher flexibility, in violet the regions with a medium flexibility and in blue the regions with a lower flexibility. In all the other regions, highlighted in cyan, there is no significant fluctuation.

By the construction of a **Sapphire plot** (*Fig.36*), using the dihedral angles of the active site and the regions closest to it (*Fig.35*), is possible to see which dihedral angles, present in this region, change during the simulation.

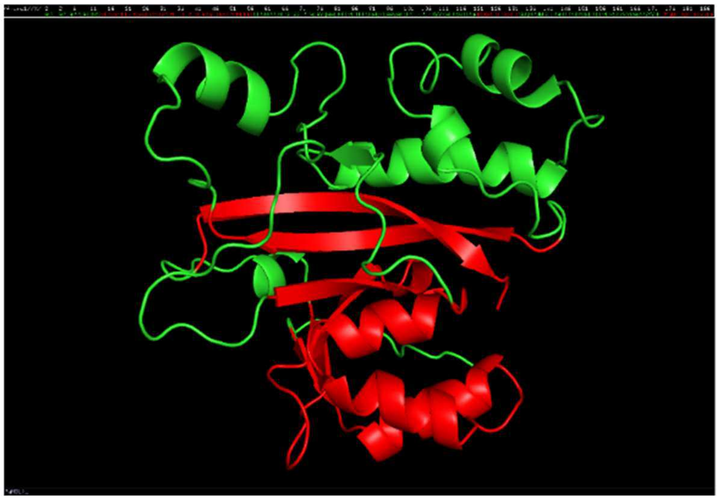


Fig.35: **KalbTGase regions used for the construction of thr Sapphire plot**. Structure of KalbTGase: in red the regions used for the annotation of the Sapphire plot

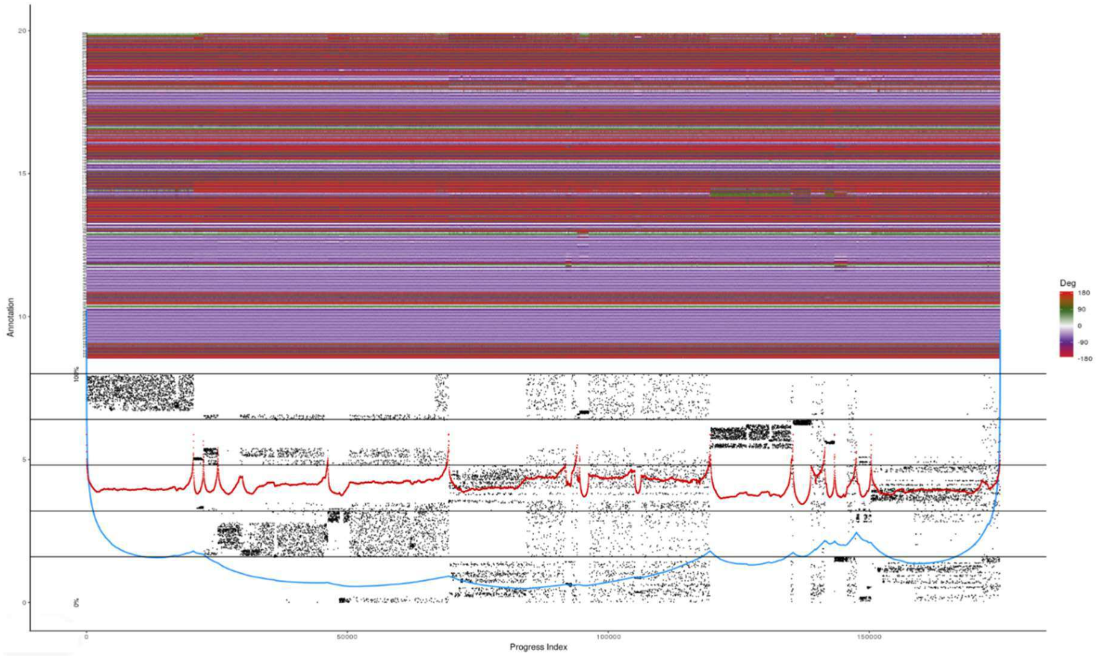


Fig.36: **Sapphire plot of KalbTGase MD simulations at 300K**. Sapphire plot on the dhiedral angles Φ and Ψ for all the connected trajectories obtained by the five different MD simulations. At the top is possible to see the annotation according to the rotations of the dihedral angles, at the bottom the time line (black dots) and the energetic basins (blue/red lines).

In particular for the catalytic Cysteine-Aspartate-Histidine no particular rotation of their Φ and Ψ dhiedral angles is detected, as is possible to see in the annotation of the Sapphire plot in *Fig.37*.

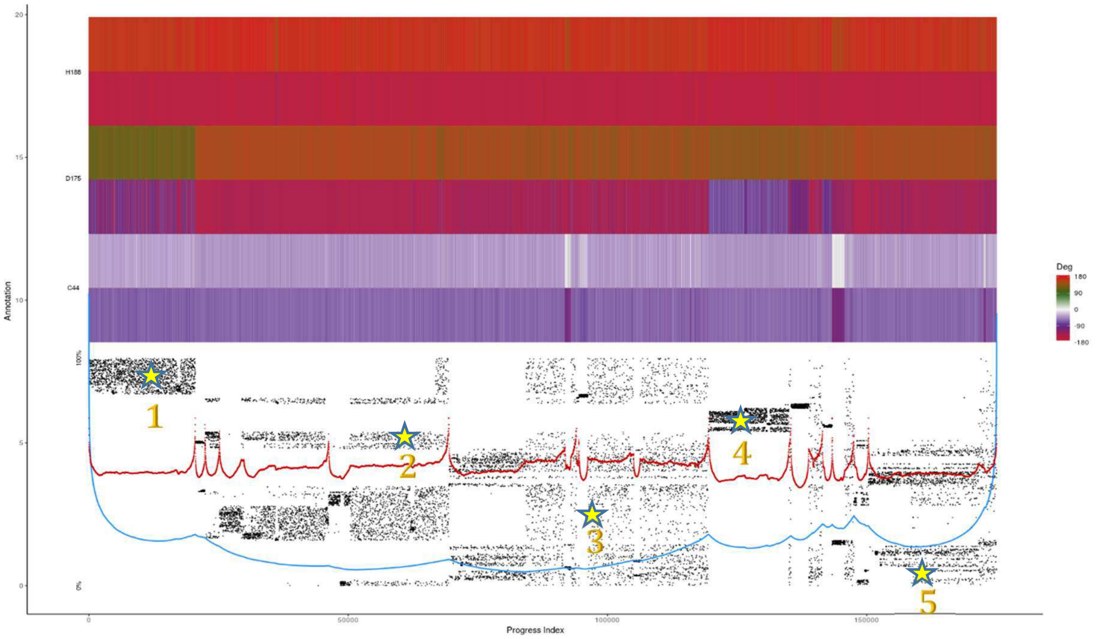


***Fig.37:* Sapphire plot with dhiedral angles annotation of the catalytic residues (KalbTGase MD at 300K).** Sapphire plot annotated with the dihedral angles Φ and Ψ of the catalytic triad Cys, Asp and His. Labelled with a star, the frames from which the structures analyzed have been taken.

Looking at the time line and to the energetic basin in *fig.37* it is possible to see that the protein seems to assume some conformations that are slightly different. Even if it seems to spend most of the MD time in a state, represented by the conformation 3, protein for a specific period visits, at least, also other three states, from each of whom a mean conformation has been taken (labelled with stars in *fig.37*). Therefore, if the selected structure no.2 has taken as reference, analyzing the Sapphire plot it is possible to say that:

- Structure no. 1 differs from no. 2 in the dihedral angles of the aa 173-176, 185, 225-226 (*Fig.38A*)
- Structure no. 3 differs from no. 2 in the dihedral angles of the aa 129-131, 173, 123-124, 208-212, 224-226 (*Fig.38B*)
- Structure no.4 differs from no. 2 in the dihedral angles of the aa 131, 173-175, 223-226 (*Fig.38C*)
- Structure no.5 differs from no.2 in the dihedral angles of the aa 129-131, 173-174, 208-212, 224-226 (*Fig.38D*)

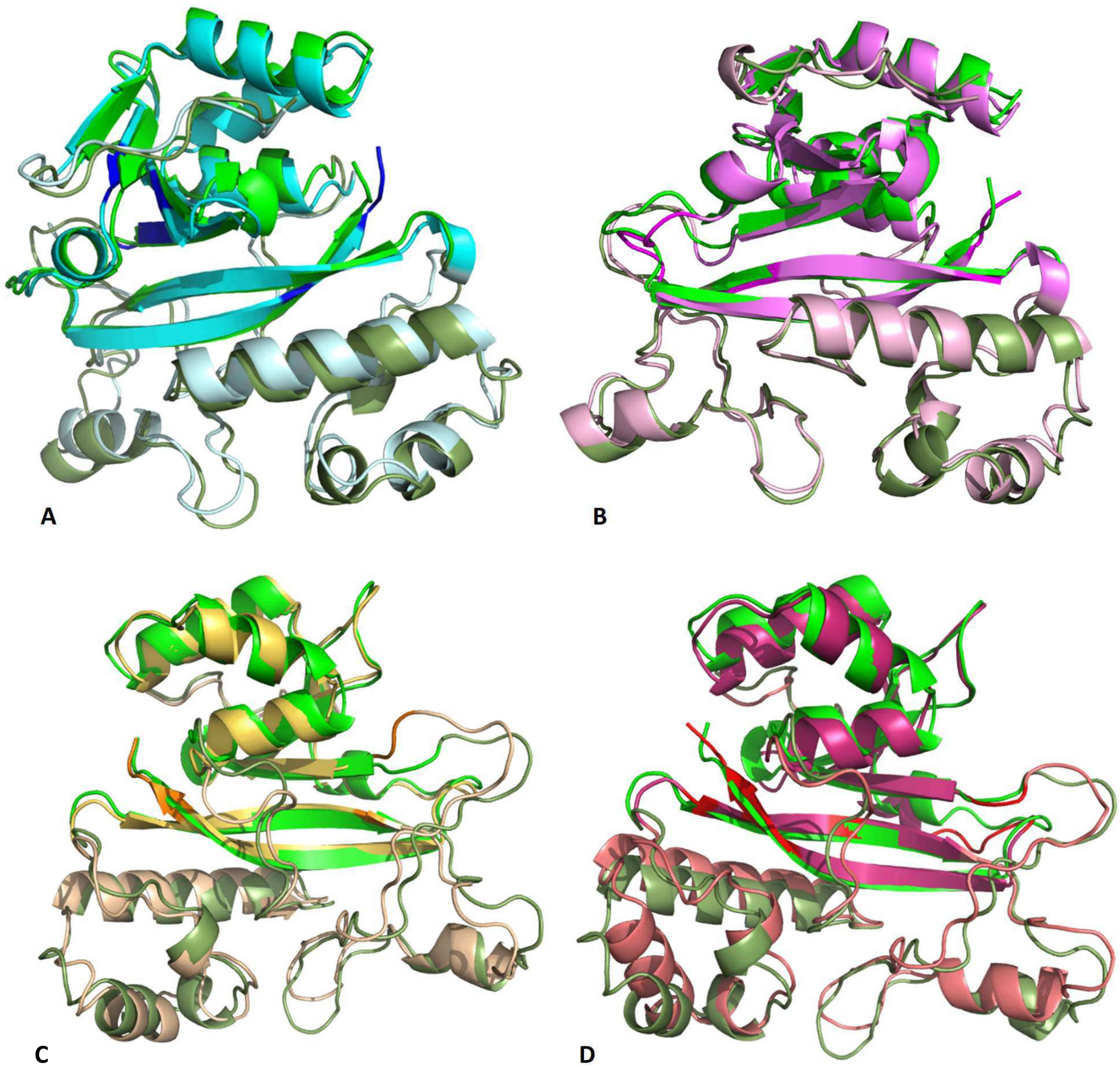


Fig. 38: **Structural analysis of the Sapphire plot (KalbTGase-300K)**. In brighter colors the regions to whom belong the dihedral angles used for the annotation of the Sapphire plot. **A**: in blue, **B**: in violet, **C**: orange and **D**: red respectively the amino acids whose dihedral angles change.

From the comparison of the 5 selected different frames it is possible to say that there isn't a large conformational change; the protein seems to preserve its structure and the catalytic site seems to be the most stable part. This result not only is in agreement with the results obtained by the RMSF plot but also with what was reported by Steffen et colleagues (2017), i.e. the most flexible part seems to be, according to the B-factors, far from the active site.

The same analyses have been performed also for MTGase MD simulations (300K, 350ns). Even in this case, as demonstrated by the **RMSD** plot (*Fig.39*), the five simulations seem to have a very similar evolution and to confirm a stability of the system not underlining particular rearrangements of the structure. Moreover, the variation range of the RMSD value is comparable with the one found for KalbTGase, making the system even better comparable.

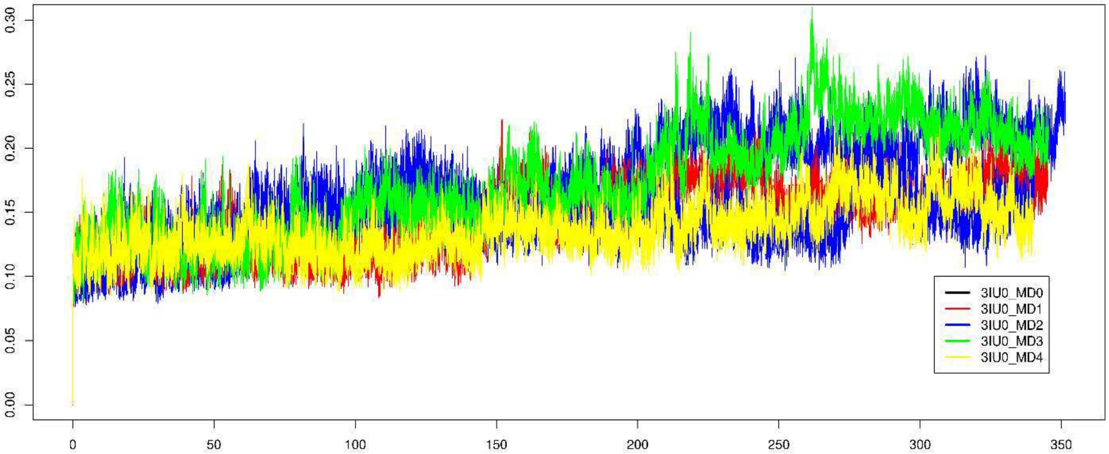


Fig.39**: RMSD plot of all the five different simulations of MTGase performed at 300K.** In the plot it is possible to see on the x-axis the duration of the simulation in ns, on the y-axis the RMSD values in nm. Different lines have been used to represent RMSD values of different simulations.

An analogue result is found also in this case, as in KalbTGase, in the analysis of the **RMSF** and, because of that, the plot related to the last 100ns of only one simulation is shown.

By the analysis of this plot (*Fig.40*) is possible to see the most flexible parts of the protein and, in particular, which one of these parts has a lower, higher or medium level of flexibility (*Fig.41*). Again, the most flexible regions are predominantly the peripherical loops, but in this case MTGase loops are bigger; moreover the detected fluctuations reach higher peak. These results suggest a major flexibility of MTGase than KalbTGase, which could explain the different specificity. Probably, being more flexible, MTGase hosts better different substrates.

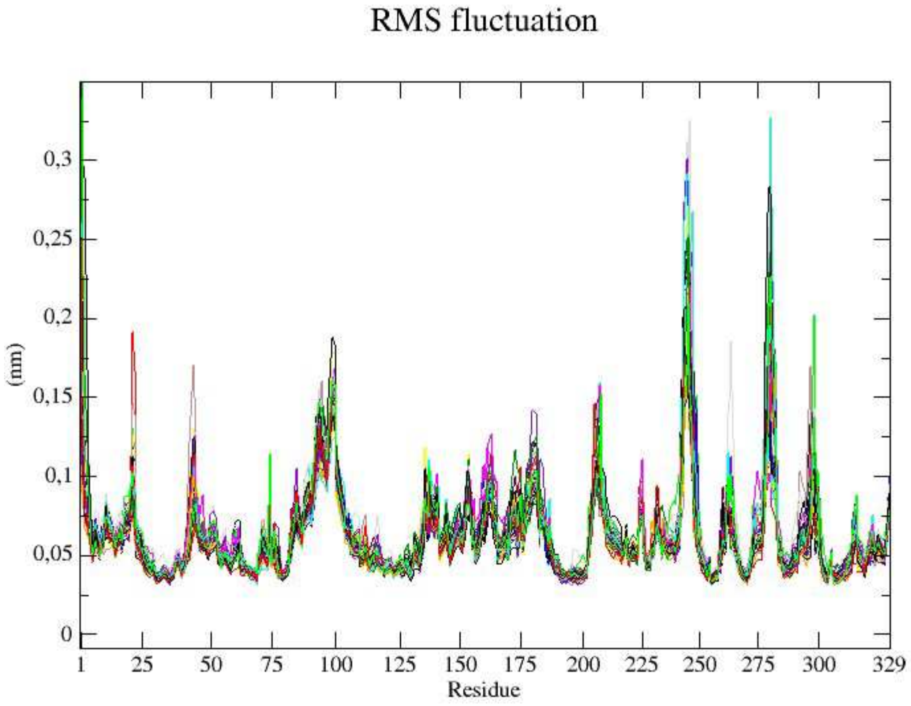


Fig.40: **RMSF plot of the last 100ns of MD simulation for MTGase at 300K**.
On the x-axis the protein residues on the y-axis the RMSF values in nm. Each colored line corresponds to an analyzed time step of the simulation (see paragraph 5.2.a)

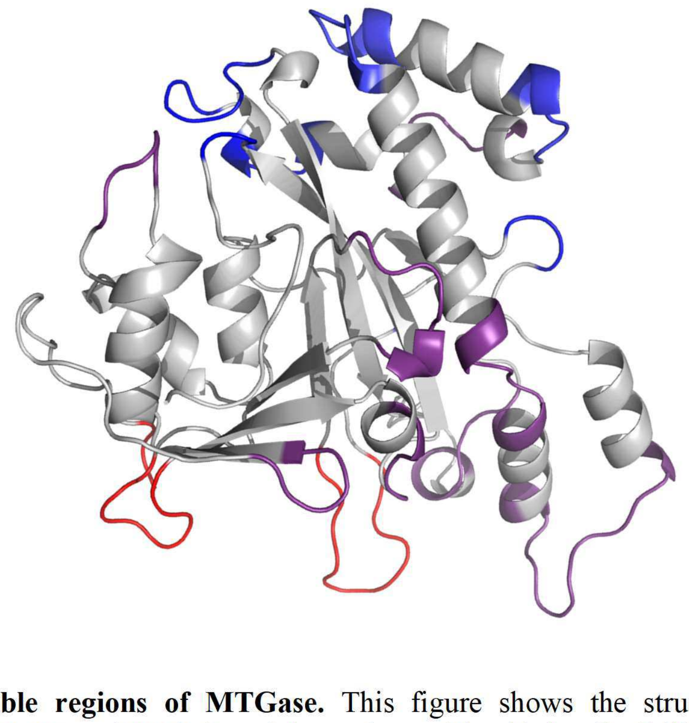


Fig.41: **Flexible regions of MTGase.** This figure shows the structure of *Streptomyces Mobaraensis* MTGase (3IU0): in red the regions with a higher flexibility, in violet the regions with a medium flexibility and in blue the regions with a lower flexibility. In all the other regions, in gray, there is no significant fluctuation.

The **Sapphire plot** (*Fig.43*), built using the dihedral angles of the active site and the regions closest to it (*Fig.42*) allows to see the dihedral angles that during the simulation change their orientation.

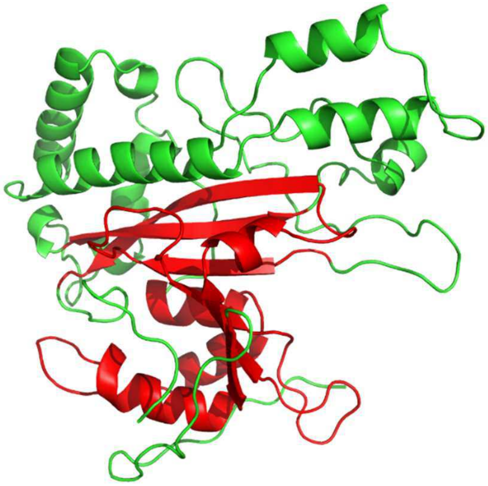


Fig.42: **MTGase regions used for the construction of thr Sapphire plot**. Structure of MTGase: in red the regions used for the annotation of the Sapphire plot

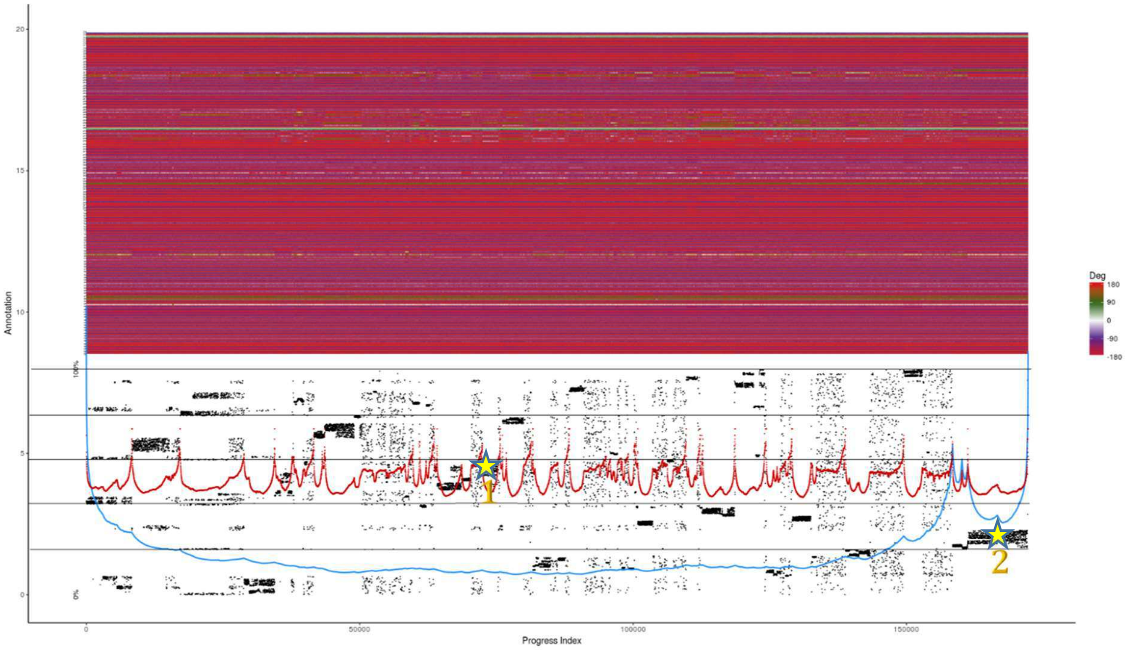


Fig.43: **Sapphire plot of MTGase MD simulations at 300K**. Sapphire plot on the dhiedral angles Φ and Ψ for all the connected trajectories obtained by the five different MD simulations. At the top it is possible to see the annotation according to the rotations of the dihedral angles, at the bottom the time line (black dots) and the energetic basins (blue/red lines). Labelled with stars, the frames from which the structures analyzed have been taken.

From the Sapphire plot it is possible to see the presence of a big basin that underline the stability of the active site and its related regions. However, it is possible to identify a second small basin explored during the second MD simulation.

Selecting two structures, one from the biggest basin and the other from the smallest one, it is possible to compare them and analyze the differences.

As shown also by the annotation of the Sapphire plot, taken as reference the selected structure no. 1, structure no. 2 differs in the dihedral angles of the amino acids 109, 111, 312-313, 321-326, 328-330, 346-349 (*Fig.44*)

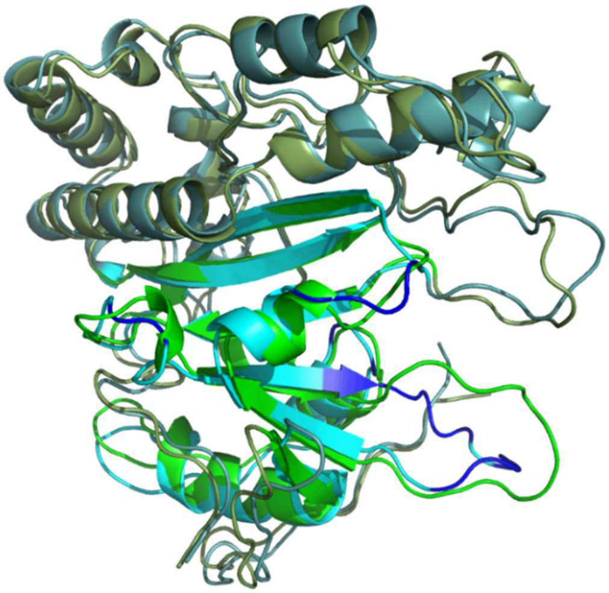


***Fig.44* Structural analysis of the Sapphire plot (MTGase-300K)**. In brighter colors the regions to whom belong the dihedral angles used for the annotation of the Sapphire plot. In blue the amino acids whose dihedral angles change.

The comparison shows that there is not any large rearrangement of the structure and the major differences are related to the long loops present in the molecule.

More in general, in all the MD simulations performed the protein seems to preserve well its folding, in particular the structure of its catalytic domain, and the major flexibility detected regards its long loops. However, even if this flexibility doesn't affect directly the active site but peripherical regions linked to it, it could be one of the reasons of its minor specificity than KalbTGase.

In order to study the adaptability of these two enzymes to different conditions, other four MD simulations have been performed, rising the temperature at 335K at first and at 355K after.

In the next paragraphs, these results are showed and discussed.

9.3 KalbTGase and MTGase: MD simulations at 335K

One simulation at the temperature of 335K, lasting 300 ns was performed for both KalbTGase and MTGase. From the RMSD analysis of the MD simulation on KalbTGase it is possible to see that in comparison to its RMSD at 300K the system shows an evolution. As expected, the RMSD increase with increasing the temperature (*Fig.45*). However, even if this indicates a light structural rearrangement probably in answer to the modified condition, the RMSD, after a first rising of its values, seems to become stable, in fact it starts to variate in a defined little range. Thus, these results suggest that, after a preliminary phase of adaptation, the system is able to reach the stability.

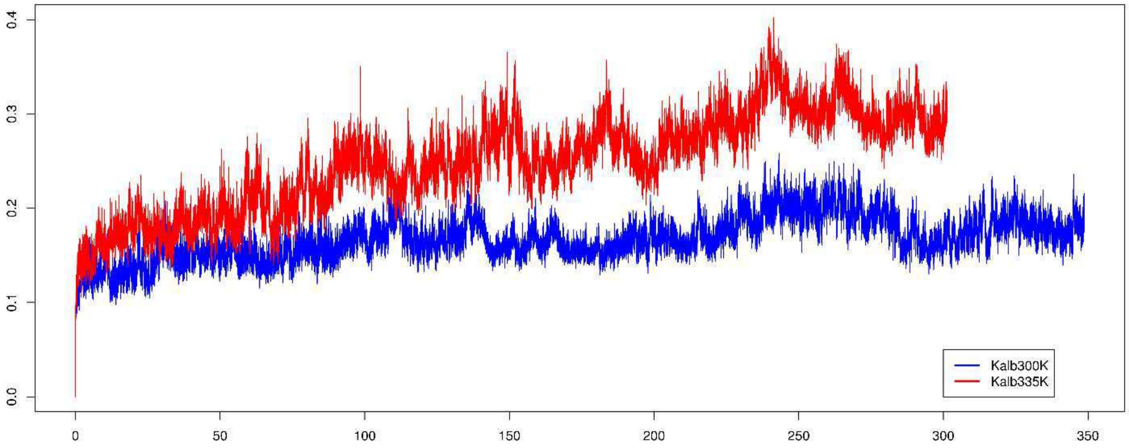


Fig.45: **Comparison between RMSD plot of KalbTGase MD simulation at 300K and at 335K.** RMSD plot of one of the five different simulation performed at 300K (blue line) and the simulation performed at 335K (red line). On the x-axis the duration of the simulation in ns, on the y-axis the RMSD values in nm.

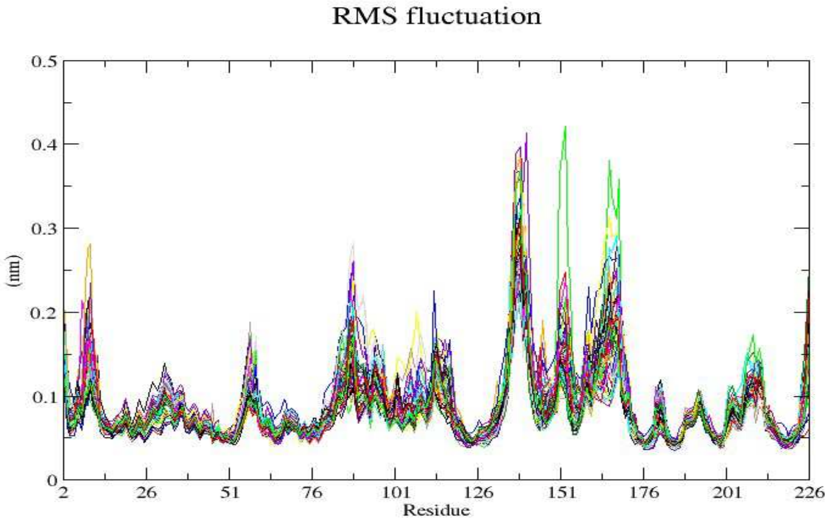


Fig.46: **RMSF plot of the last 100ns of MD simulation for KalbTGase at 335K**.
On the x-axis the residues, which form the protein, numbered consecutively, on the y-axis the RMSF values in nm. Each colored line corresponds to an analyzed time step of the simulation

From the RMSF analysis it is possible to see that the fluctuation peaks are higher (*Fig.46*) but the regions involved in the fluctuations (*Fig47*) are more or less the same found in the simulation at 300K. The highest peaks that in the RMSF plot of KalbTGase at 300K reach a maximum value of 0.25 nm, now reach values among 0.4 nm (0.3 for the peaks judge as medium). For the lowest peaks, instead, the situation is almost the same seen in the previous RMSF plot.

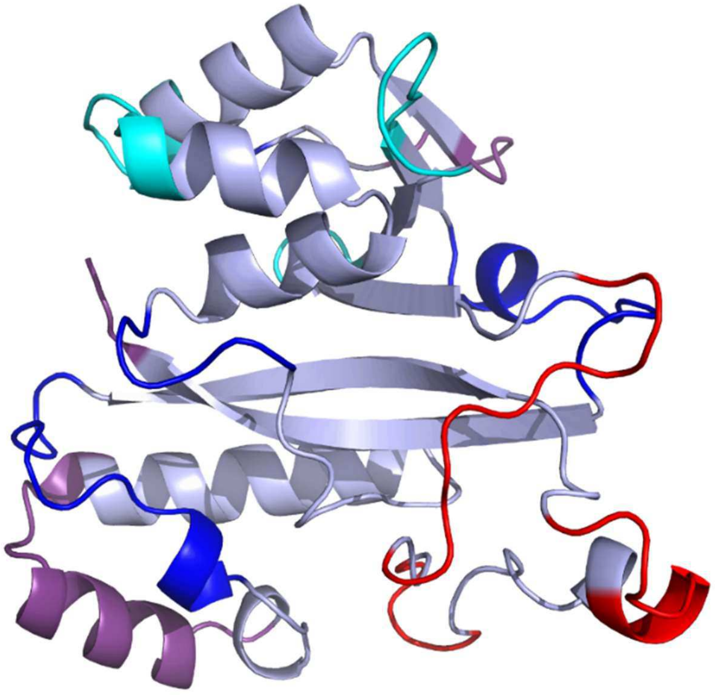


Fig.47: **Flexible regions of KalbTGase (MD simulation at 335K).** This figure shows the structure of Kalb mTGase: in red the regions with a high flexibility, in violet the regions with a medium flexibility, in blue the regions with a low flexibility and in cyan the regions with a very low flexibility. In all the other regions, highlighted in light blue, there is no significant fluctuation.

Also in this case, the **Sapphire plot** (*Fig.48B*) analysis was performed, using the dihedral angles of the active site region and the others closest to it (*Fig.48A*). The protein seems to be stable in its conformation, as demonstrated by the formation of a unique big energetic basin (example structure no. 1). However, it is possible to detect two states that the protein visits for a specific time frame; also in this case a mean structure was extracted from each of the states (pointed by stars 1, 2 and 3 in *fig.48B*) and compared to the others.

A variation of specific dihedral angles is shown by the annotation at the top of *fig.48B*.

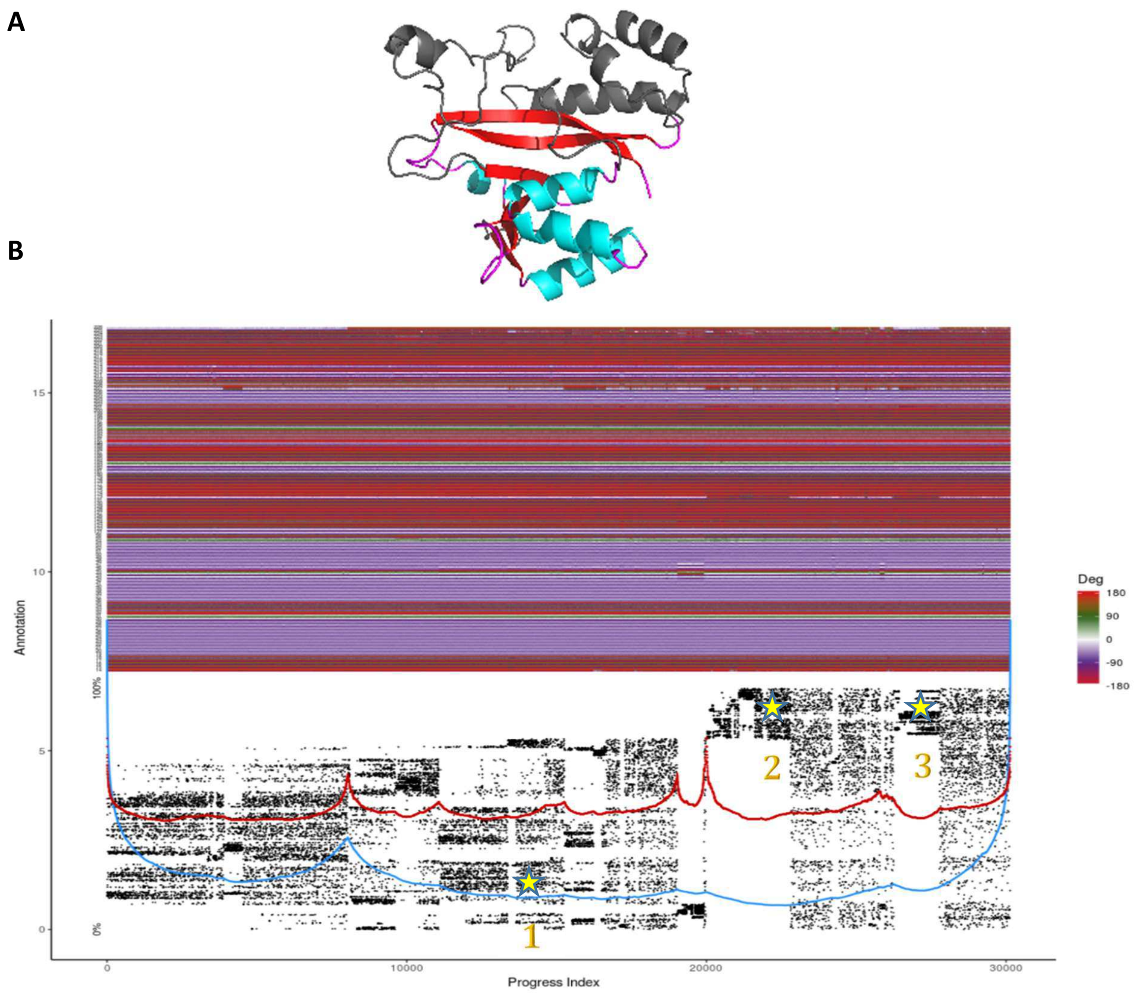


Fig.48: **Sapphire plot of KalbTGase MD simulations at 335K. A**: Structure of Kalb mTGase: in cyan, red and violet the regions used for the annotation of the Sapphire plot. **B**: Sapphire plot on the dihedral angles Φ and Ψ for the trajectory obtained by the MD simulation at 335K. At the top it is possible to see the annotation according to the rotations of the dihedral angles, at the bottom the time line and the energetic basins. Labelled with numbered stars, the frames from which the structures analyzed have been taken.

Taking as reference the selected structure no. 1, from the Sapphire plot analysis it is possible to say that:

- Structure no. 2 differs from no. 1 in the dihedral angles of the amino acids 173-175, 207-209 (*Fig.49A*)

- Structure no. 3 differs from no. 1 in the dihedral angles of the amino acids 173-175, 207-209, 225-226 (*Fig.49B*)

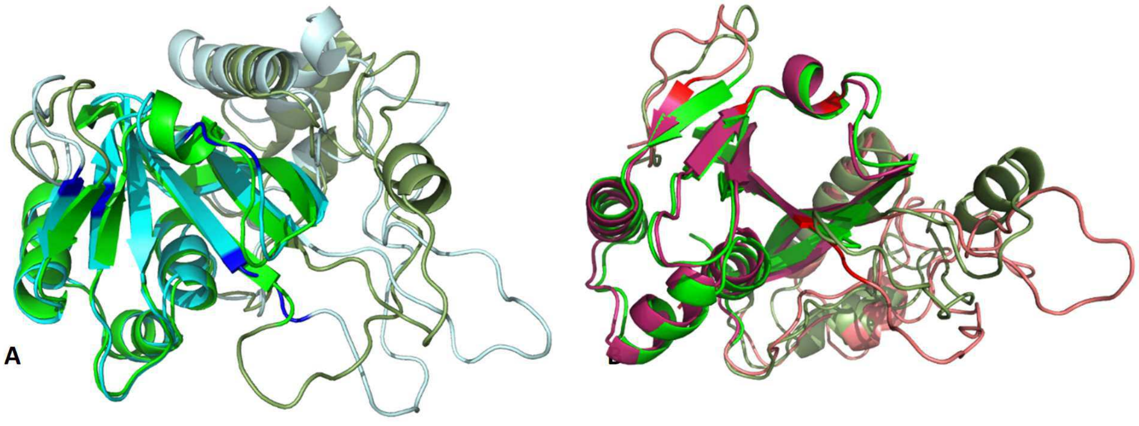


Fig. 49: **Structural analysis of the Sapphire plot (KalbTGase-335K)**. In brighter colors the region to whom belong the dihedral angles used for the annotation of the Sapphire plot. **A**: in blue and **B**: in red respectively the amino acids whose dihedral angles change.

From these selected frames comparison and the general analysis of the MD trajectory it is possible to say that even if there are some conformational rearrangements, there is not a large conformational change in the active site area. The flexible regions are confirmed to be the same showed by the RMSF plot at room temperature and the protein seems to preserve well its structure; moreover, the catalytic site is confirmed to be the most stable part in the protein.

From the **RMSD** (*Fig.50*) obtained from the MD simulation at 335K for 300 ns of MTGase a very surprising result is shown. The rising of the temperature seems do not affect the RMSD. The protein, after a while, reaches the same hypothetical stability profile of the MD simulation of the molecule at 300K.

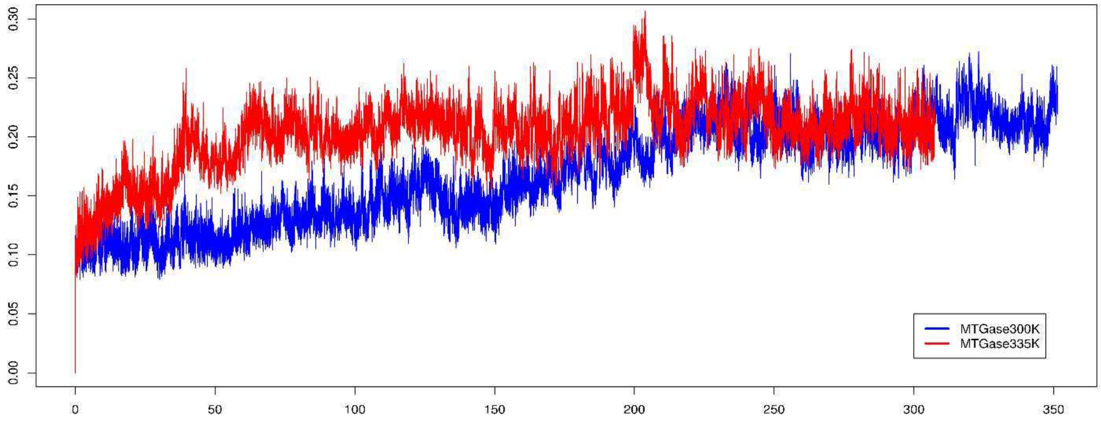


Fig.50: **Comparison between RMSD plot of MTGase MD simulation at 300K and at 335K.** RMSD plot of one of the five different simulation performed at 300K (blue line) and the simulation performed at 335K (red line). On the x-axis the duration of the simulation in ns, on the y-axis the RMSD values in nm.

The RMSF instead seems to be more affected than the RMSD by the rising of the temperature. The fluctuation peaks are higher (*Fig.51*) but the regions involved in the fluctuations are also in this case quite the same (*Fig.52*). The highest peaks, that in the RMSF plot of MTGase at 300K reach a maximum value of 0.35 nm, now reach values among 0.5 nm; however, most of the peaks are judge as medium because they do not overcome 0.25 nm. For the lowest peaks, instead the situation is almost the same seen in the previous RMSF plot.

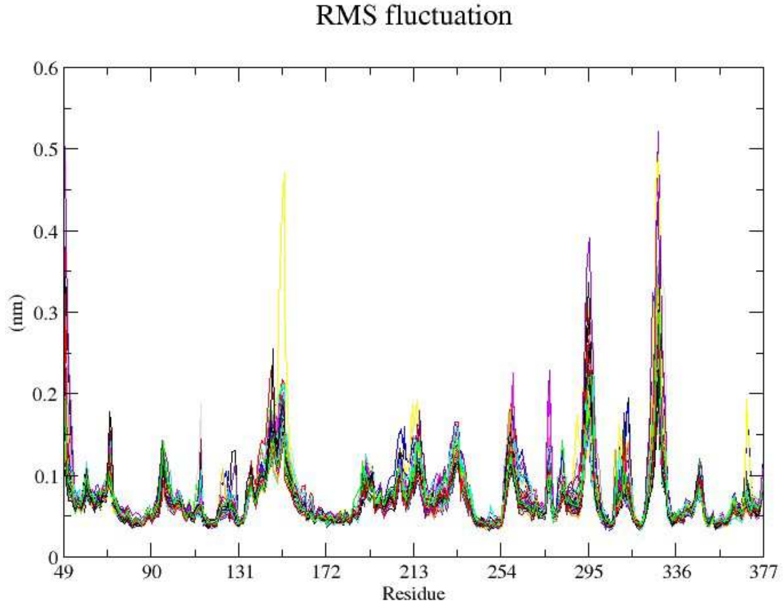


Fig.51: **RMSF plot of the last 100ns of MD simulation for MTGase at 335K**.
On the x-axis the residues, which form the protein, numbered consecutively, on the y-axis the RMSF values in nm. Each colored line corresponds to an analyzed time step of the simulation

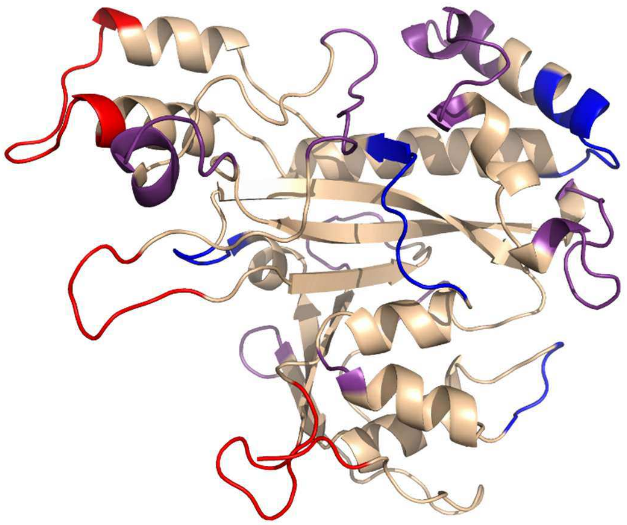


Fig.52: **Flexible regions of MTGase (MD simulation at 335K).** This figure shows the structure of MTGase: in red the regions with a higher flexibility, in violet the regions with a medium flexibility and in blue the regions with a lower flexibility. In all the other regions, highlighted in beige, there are no significant fluctuations.

Due to the fact that the RMSD suggests no significant difference between the two simulations performed at different temperature but the RMSF instead shows a major impact of that one on the molecule, it is possible to conclude that the structure is well preserved at 335K, but its longer loops start to be affected from this rising of temperature so become more flexible and start to fluctuate more. However, at least for what is possible to see from this simulation, this fluctuation doesn't affect the stability of the catalytic core, that remains stable even if some regions close to it show a medium flexibility.

To investigate more in detail the variations that occur in the catalytic pocket, a **Sapphire plot** was built, using the dihedral angles of the active site region and the others closest to it (*Fig.53A*)

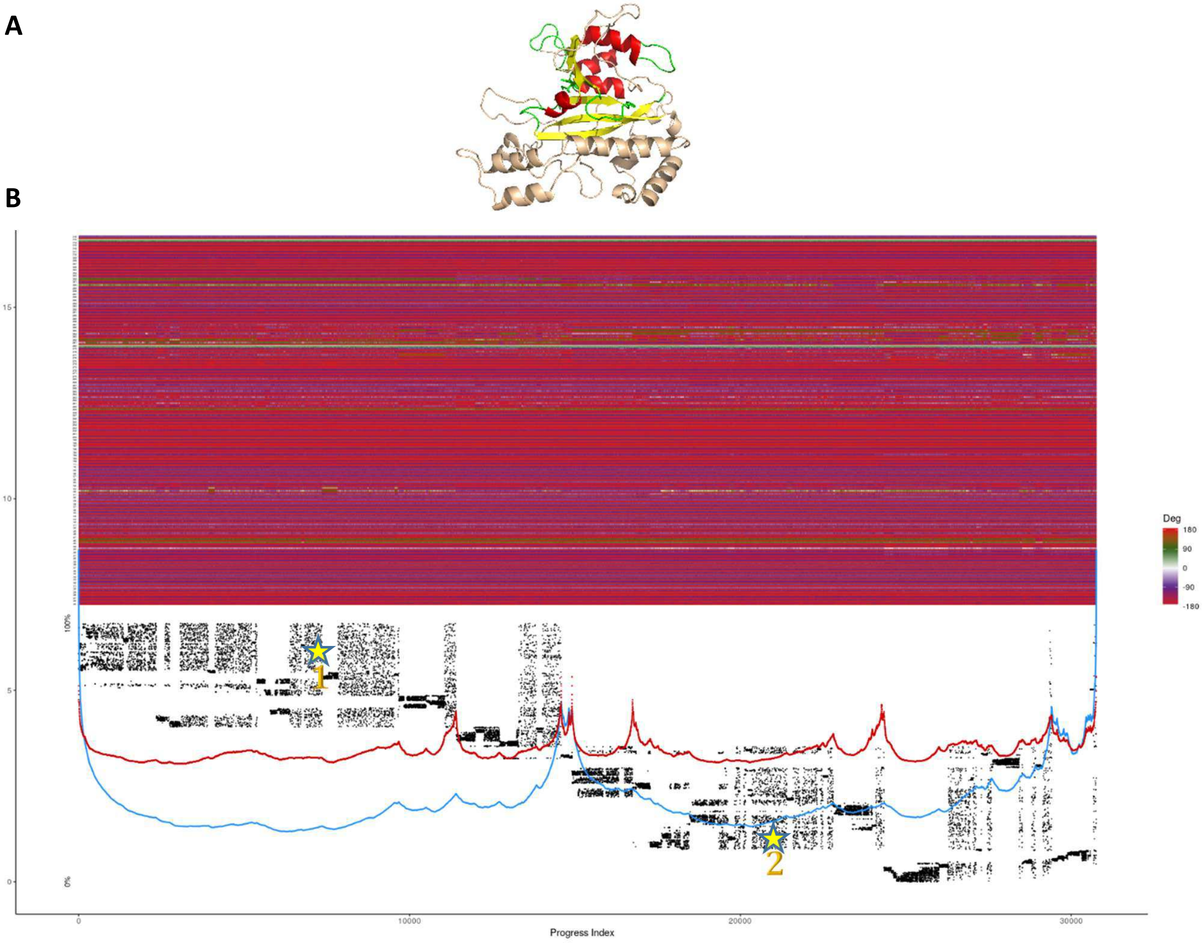


Fig.53: **Sapphire plot of MTGase MD simulations at 335K. A**: Structure of MTGase: in yellow, red and green the regions used for the annotation of the Sapphire plot. **B**: Sapphire plot on the dihedral angles Φ and Ψ for the trajectory obtained by the MD simulation at 335K. At the top it is possible to see the annotation according to the rotations of the dihedral angles, at the bottom the time line and the energetic basins. Labelled with stars, the frames from which the structures analyzed have been taken.

From this plot (*Fig.53B*) it is possible to see that there are two main basins, and an analysis of two structure selected from both of them could be interesting in order to study the main differences between them.

If the selected structure no. 1 is taking as reference, analyzing the Sapphire plot, it is possible to say that structure no. 1 differs from structure no. 2 in the dihedral angles related to the amino acids: 109, 304, 306, 321-335, 349-350 (*Fig.54*).

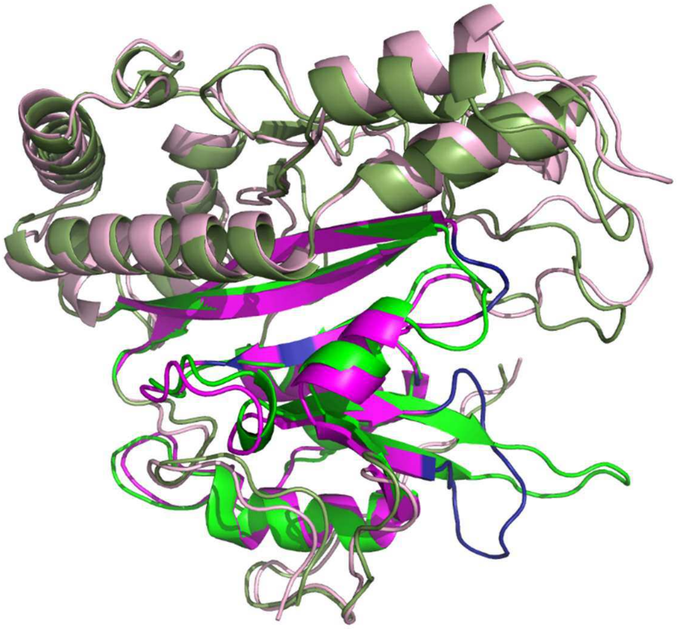


Fig.54: **Structural analysis of the Sapphire plot (MTGase-335K)**. In brighter colors the regions to whom belong the dihedral angles used for the annotation of the Sapphire plot. In blue the amino acids whose dihedral angles change.

From the comparison it is possible to see that the two structures are very similar. Generally, all the results obtained from this simulation underline that there are no structural rearrangements. In fact, the secondary structure of the catalytic domain is well conserved, and the differences notably also from the Sapphire plot are mainly related to the long loops present in the protein, that, as shown by the RMSF plot, are subject to fluctuations that increase as the temperature rises.

9.4 KalbTGase and MTGase: MD simulations at 355K

Other two MD simulation, one for each structure, have been performed at 355K.

The RMSD plot of KalbTGase shows that the RMSD, as expected, increases with increasing the temperature (*Fig.55B*), but in this case it seems do no reach a clear equilibrium (*Fig.55A*). The RMSD profile shows several jumps after which the profile seems to reach a stability, however after this brief steady phase another jump is visible, thus is not possible for certain to say that at the end of the simulation a more stable conformation is reached.

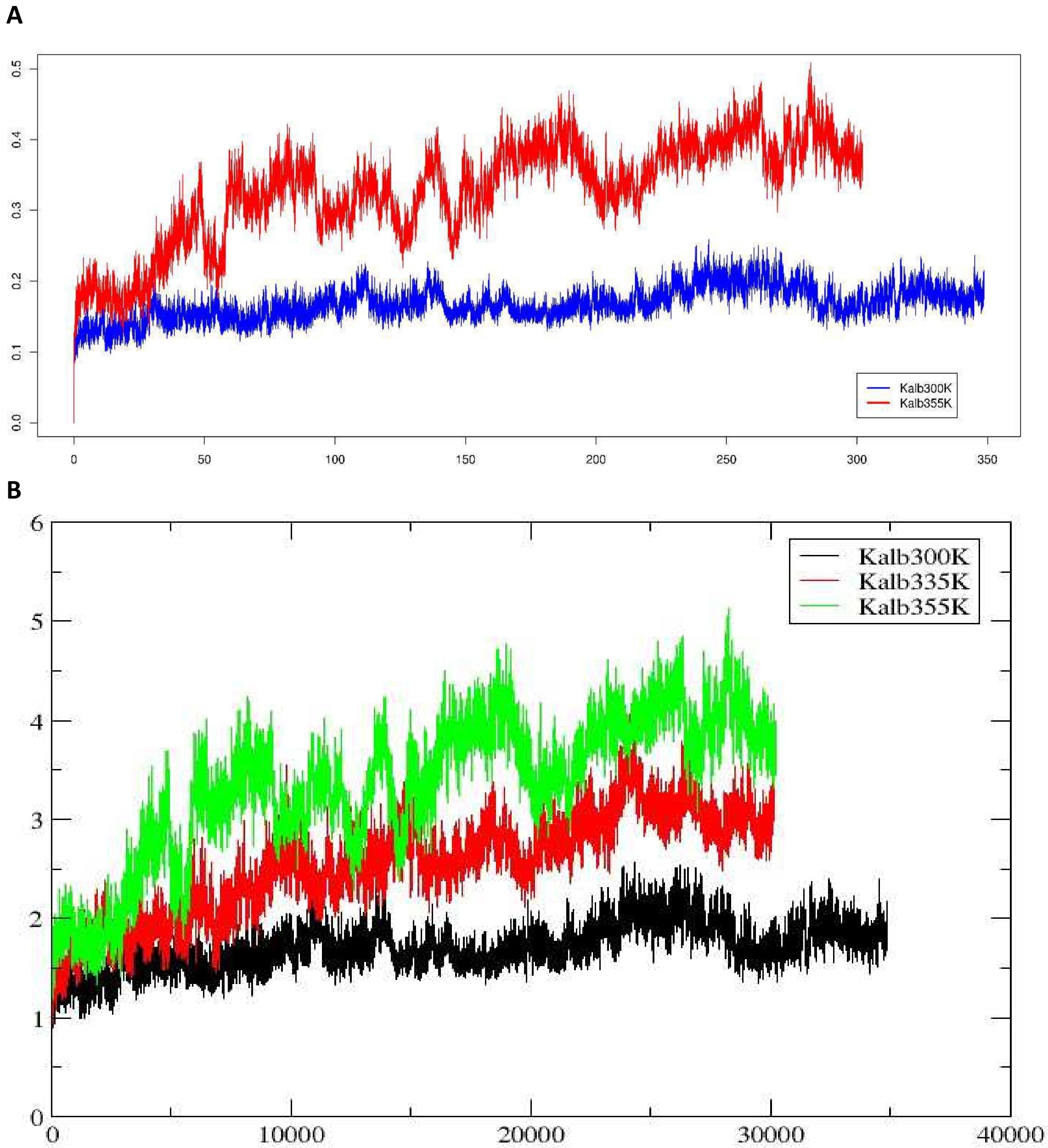


Fig.55: **Comparison between RMSD plot of KalbTGase MD simulations. A**: RMSD plot of one of the five different simulations performed at 300K (blue line) and the simulation performed at 355K (red line). On the x-axis the duration of the simulation in ns, on the y-axis the RMSD values in nm **B**: Comparison of the RMSD of the Cα related to (in black) the MD simulation of KalbTGase at 300K, (in red) the MD simulation of KalbTGase at 335K and (in green) the MD simulation of KalbTGase at 355K. On the x-axis the duration of the simulation in numbers of frames, on the y-axis RMSD in Å. Is possible to see how the RMSD profile and the rising of the temperature correlate.

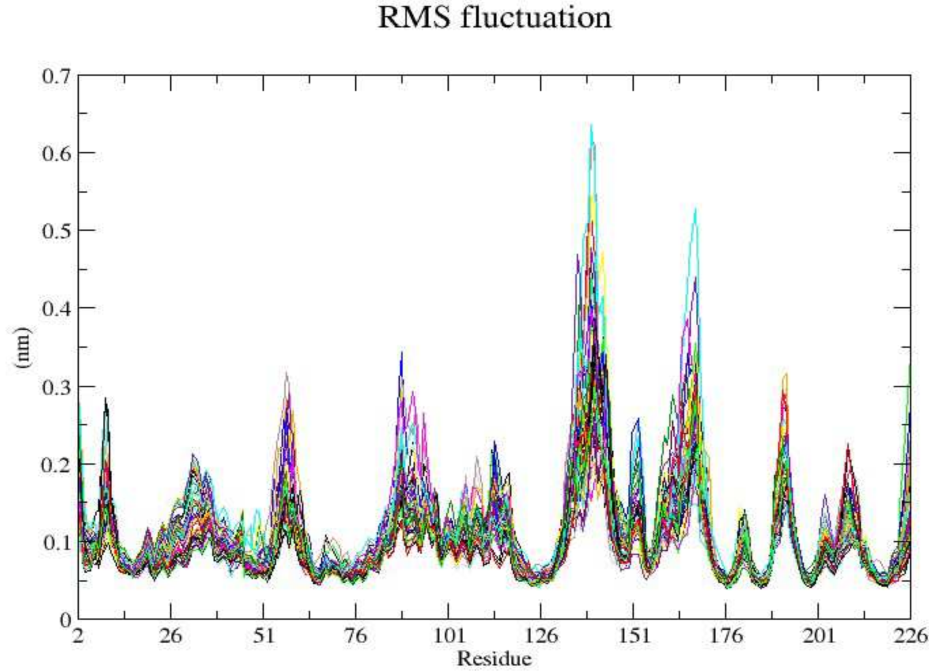


Fig.56: **RMSF plot of the last 100ns of MD simulation for KalbTGase at 355K**.
On the x-axis the residues, which form the protein, numbered consecutively, on the y-axis the RMSF values in nm. Each colored line corresponds to an analyzed time step of the simulation

The RMSF plot shows that, as expected, the fluctuation peaks are higher than the peaks related to the MD simulation at 335K (*Fig.56*) but the regions involved in the fluctuations are the same. This aspect underlines that these parts (*Fig.57*) are the most flexible regions of the molecule. The highest peaks, that in the RMSF plot of KalbTGase at 335K reach a maximum value among 0.4 nm, now reach values among 0.5/0.6 nm (0.3 for the peaks judge as medium). For the lowest peaks (around 0.2nm), instead the situation is almost the same seen in the previous RMSF plot in *fig.46*. These results, in principle, suggest a major flexibility of the molecule.

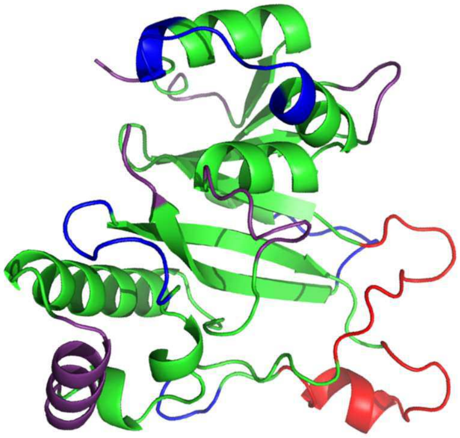


Fig.57: **Flexible regions of KalbTGase (MD simulation at 355K).** This figure shows the structure of KalbTGase: in red the regions with a high flexibility, in violet the regions with a medium flexibility, in blue the regions with a low flexibility. In the green regions, there are no significant fluctuations.

The building of the Sapphire plot (*Fig.58B*) on the active site and the regions immediately close to it (*Fig.58A*), in this case has represented a powerful tool to understand what real happens to this molecule during the simulation performed.

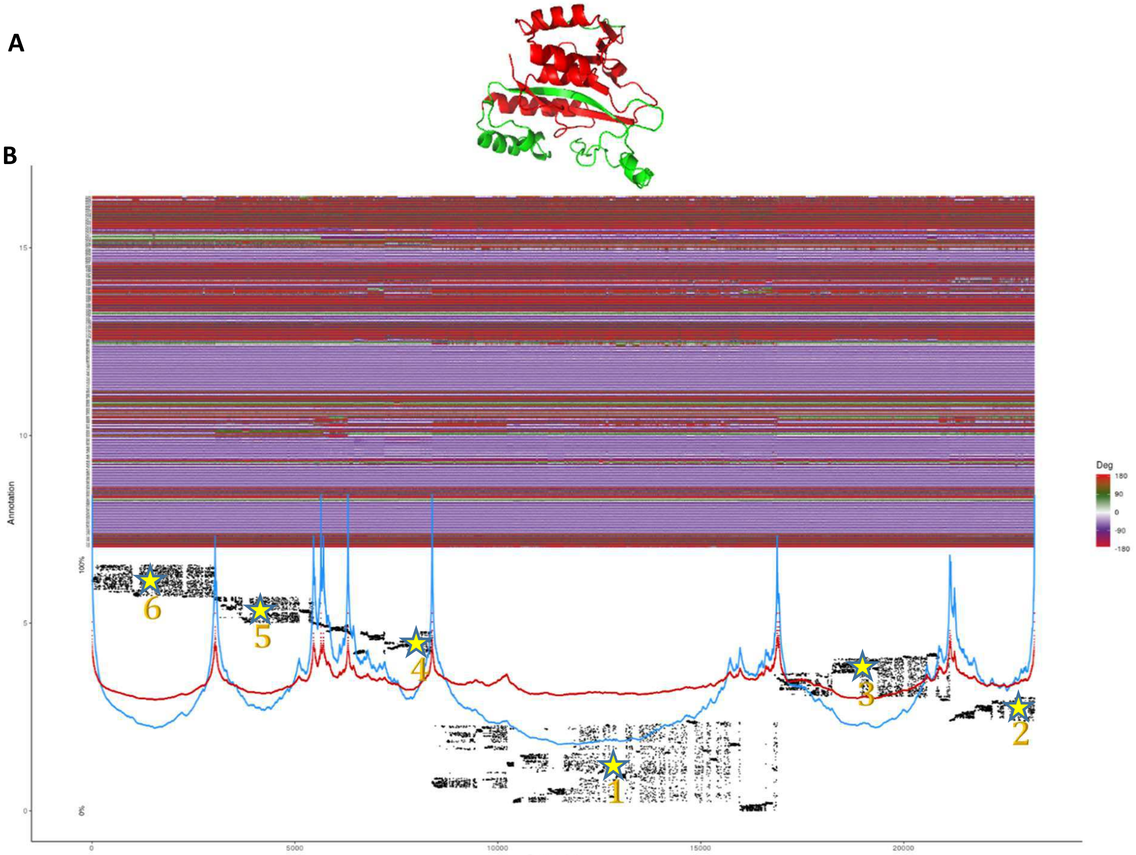


Fig.58: **Sapphire plot of KalbTGase MD simulations at 355K. A**: Structure of KalbTGase: in red the regions used for the annotation of the Sapphire plot. **B**: Sapphire plot on the dihedral angles Φ and Ψ for the trajectory obtained by the MD simulation at 355K. At the top it is possible to see the annotation according to the rotations of the dihedral angles, at the bottom the time line and the energetic basins. Labelled with stars, the frames from which the structures analyzed have been taken.

The analysis of the plot, taking the selected structure no. 1 as reference, shows that:

- Structure no. 1 differs from no. 2 in the dihedral angles of the amino acids 54-59, 174-176, 190-196 (*Fig.59A*)

- Structure no. 3 differs from no. 2 in the dihedral angles of the amino acids 46, 57-60, 190, 203, 211 (*Fig.59B*)

- Structure no. 4 differs from no. 2 in the dihedral angles of the amino acids 51-59, 67-68, 86-88, 174, 189-194, 208-212, 224-225 (*Fig.59C*)

- Structure no. 5 differs from no. 2 in the dihedral angles of the amino acids 53-55, 67-68, 206-213 (*Fig.59D*)

- Structure no. 6 differs from no. 2 in the dihedral angles of the amino acids 52-55, 67-68, 208-212, 224-225, 242(*Fig.59E*)

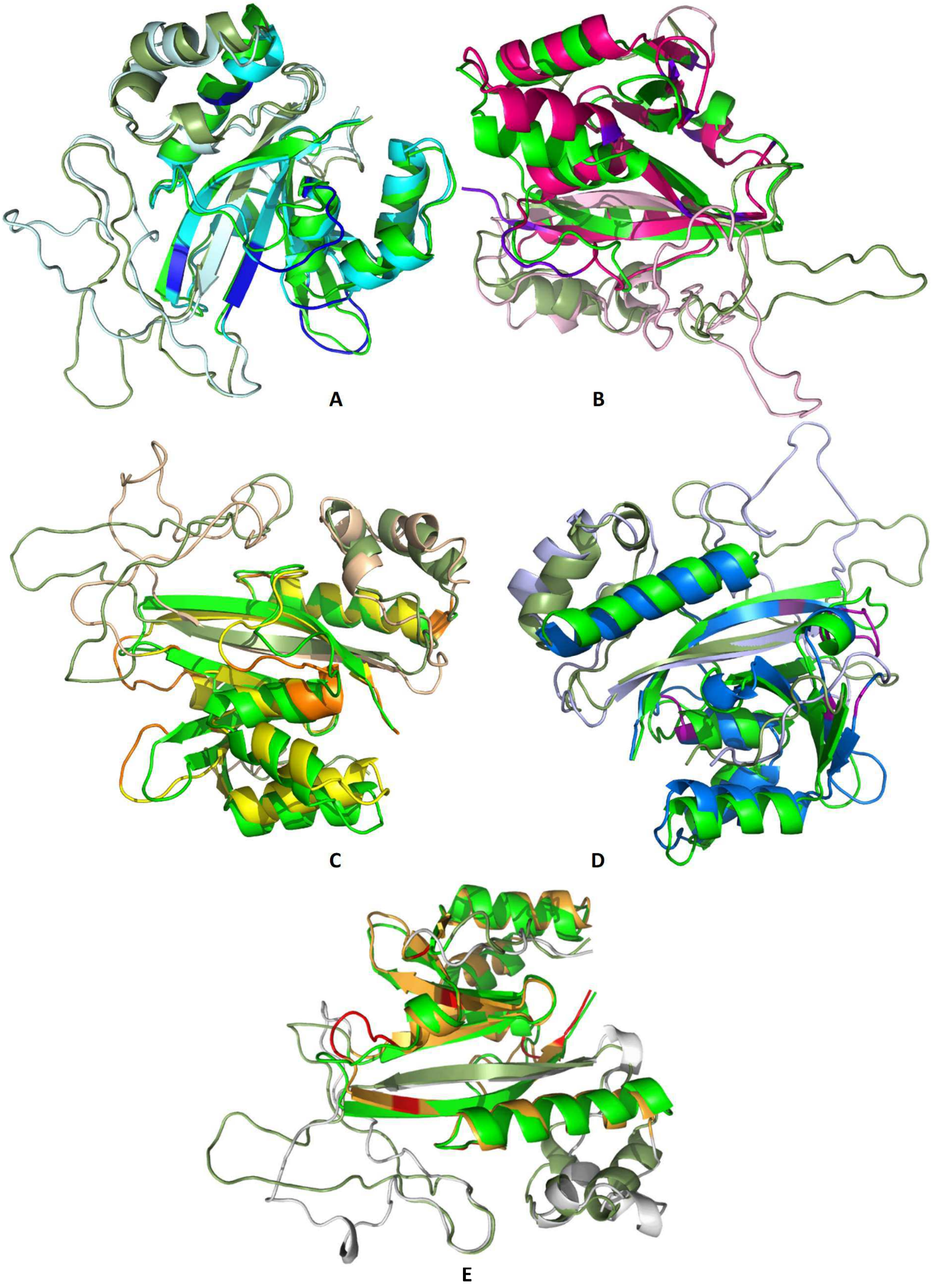


Fig. 59: **Structural analysis of the Sapphire plot (KalbTGase-335K)**. In brighter colors the region to whom belong the dihedral angles used for the annotation of the Sapphire plot. **A**: in blue, **B**: in violet, **C**: in orange, **D**: in violet and **E**: in red respectively the amino acids whose dihedral angles change.

More in general, the dihedral angles that show more differences are that related to the amino acids 54-59, 87-88, 190-195, 205-216. All these amino acids are in loop regions. Moreover, as it is possible to see from the Sapphire plot, all the six structures analyzed seem to be an evolution of the first, and these results are in agreement with the RMSD profile, that shows a structure in evolution.

From this simulation, it is only possible to say that KalbTGase at 355K seems to preserve the stability of its active site but seems to be more flexible. However a simulation of 350ns has not been sufficient to show a rearrangement of the molecule that was kept for all the simulation or probably this molecule simply undergoes several rearrangements during the time that however do not affect its stability.

A similar simulation, which lasted 300ns, was performed also on MTGase.
As expected, the RMSD increases with increasing the temperature (*Fig.60*), but in this simulation the values reached by the RMSD are twice than the RMSD values related to the MD simulations at 335K and 300K respectively (*Fig.61*). Also in this case the RMSD plots seems suggest an evolution of the system; after a first long phase of stability, the system reach at least another stable conformation, where the stability seems to be quite preserved.

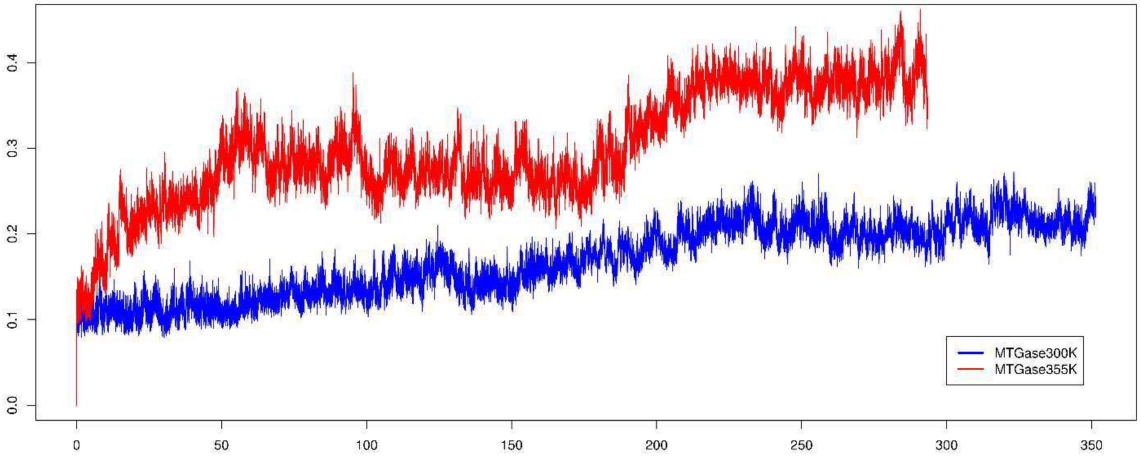


***Fig.60:* Comparison between RMSD plot of MTGase MD simulation at 300K and at 355K.** RMSD plot of one of the five different simulations performed at 300K (blue) and the simulations performed at 355K (red). On the x-axis the duration of the simulation in ns, on the y-axis the RMSD values in nm.

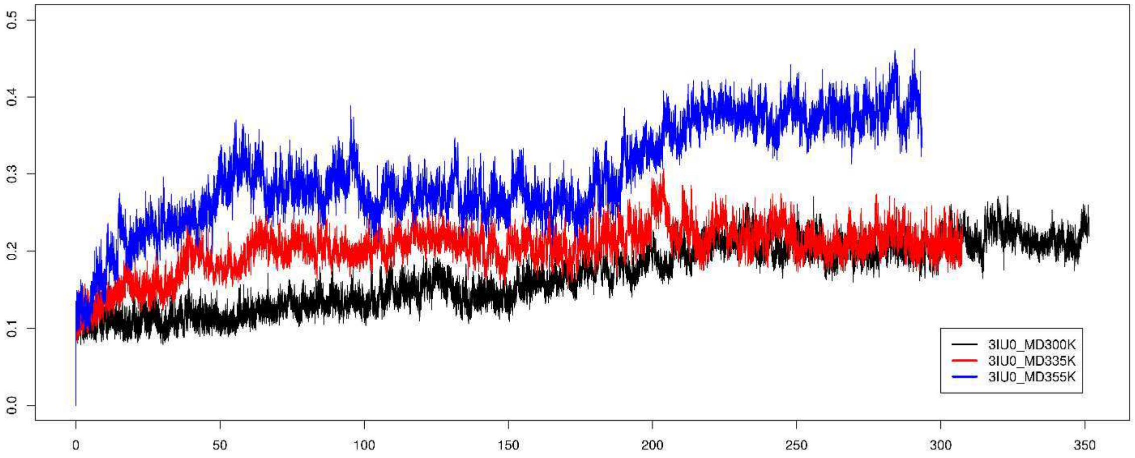


Fig.61: **Comparison between RMSD plot of MTGase MD simulations.** This plot shows a comparison of the RMSD calculated for the Cα related to (in black) the MD simulation of *Streptomyces Mobaraensis* (3IU0) MTGase at 300K, (in red) the MD simulation of MTGase at 335K and (in blue) the MD simulation of MTGase at 355K. On the x-axis the duration of the simulation in numbers of frames, on the y-axis RMSD in Å.

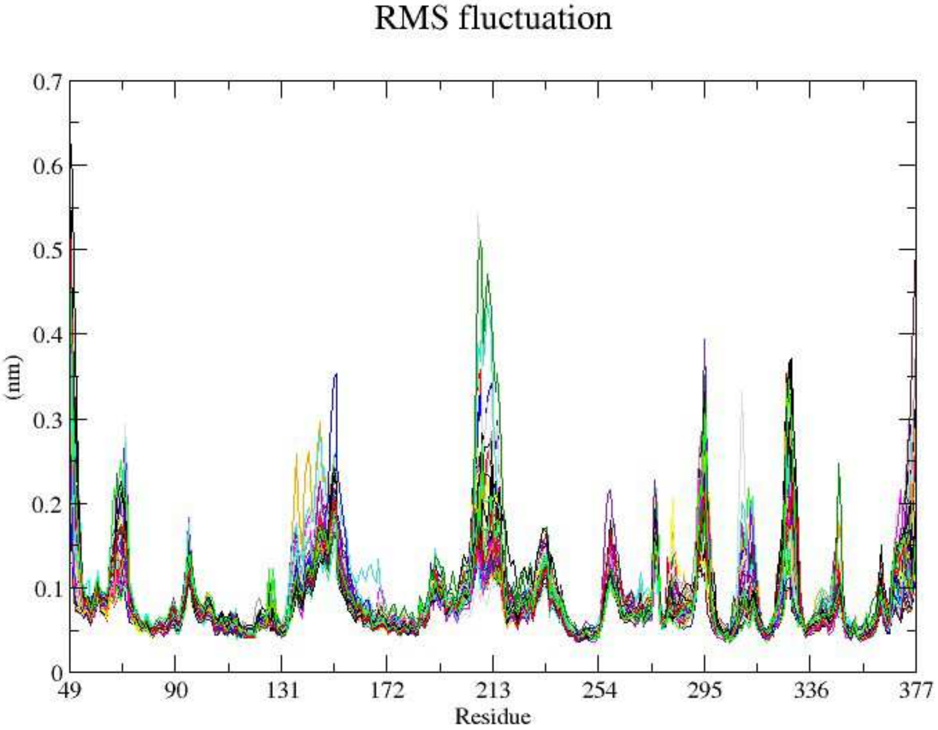


Fig.62 **RMSF plot of the last 100ns of MD simulation for MTGase at 355K**.
On the x-axis the residues, which form the protein, numbered consecutively, on the y-axis the RMSF values in nm. Each colored line corresponds to an analyzed time step of the simulation.

From the RMSF plot (*Fig.62*), it can be deduced that the fluctuation peaks are higher than the peaks related to the MD simulation at 335K but the regions involved in the fluctuations are the same, also in this case. Thus, these parts are the most flexible regions of the molecule (*Fig.63*).

The highest peaks reach almost the same values reached by the corresponding picks in the RMSF plot of MTGase at 335K; however, the peaks judge as medium reach higher values (around 0.4 instead of 0.25 of the previous plot). For the lowest peaks (lower than 0.2nm), the situation is almost the same seen in the RMSF plot at 335K (*Fig.51*).

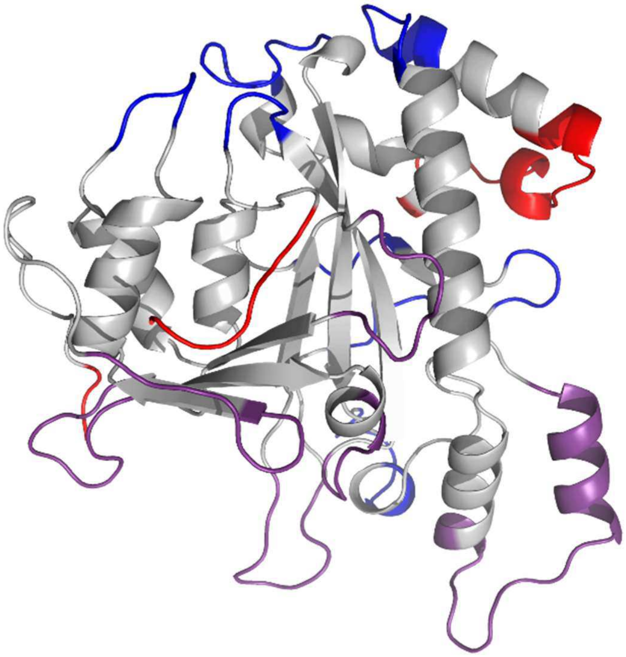


Fig.63: **Flexible regions of MTGase (MD simulation at 355K).** This figure shows the structure of 3IU0 MTGase: in red the regions with high flexibility (RMSF value > 0.5 nm), in violet the regions with medium flexibility (0.2nm < RMSF value < 0.4 nm), in blue the regions with low flexibility (RMSF value < 0.2). In gray, regions without significant fluctuations.

In general, the RMSF plot suggests a more flexibility of those regions of the molecule which for their structure are affected of a major mobility. Actually, the catalytic core does not show any flexibility, however this is not enough to suppose a preservation of the catalytic activity even because the regions involved in these fluctuations are several and in this case the values reach by the fluctuation are high.

Several variations are also visible from the Sapphire plot (*Fig.64*), where the time line and the energetic basins show clearly as after a first phase the system visit at least five different states, suggesting an evolution of the system that start from the first energetic basin visited (reference structure chosen: no. 6).

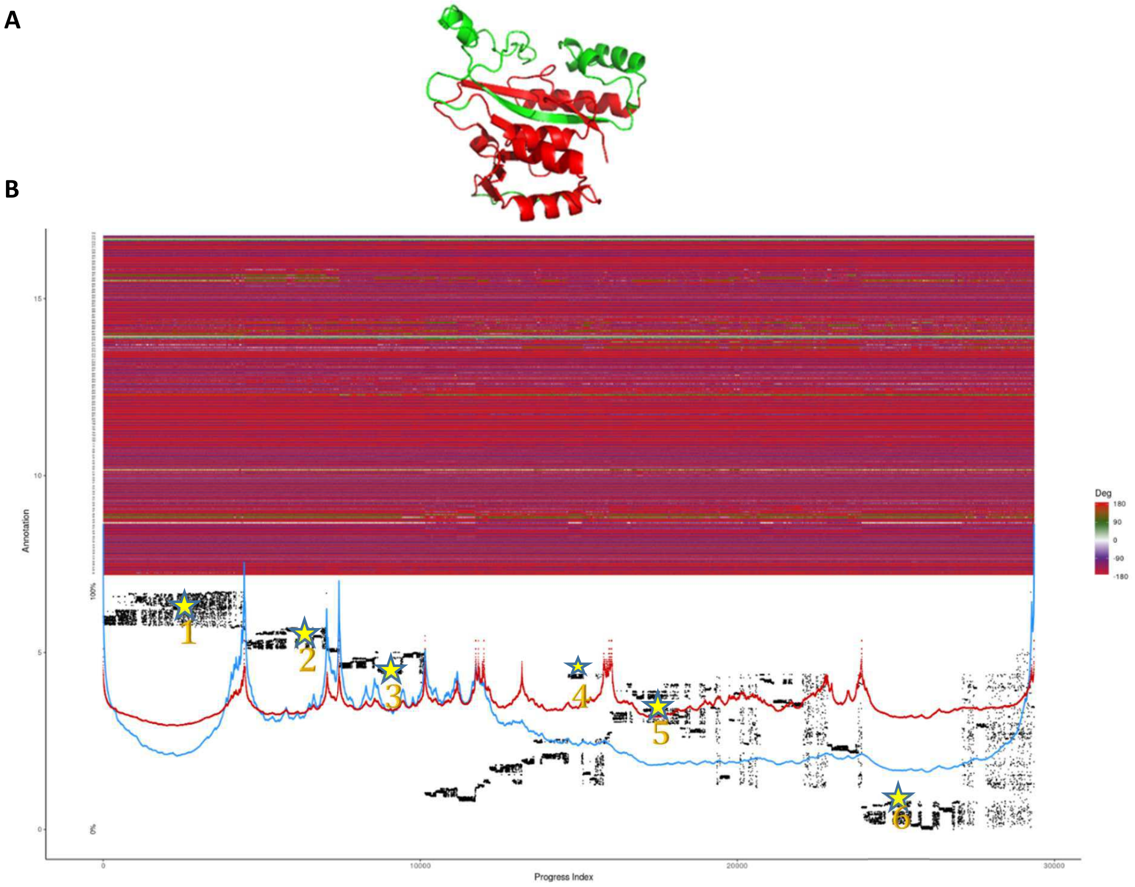


Fig.64: **Sapphire plot of MTGase MD simulations at 355K. A**: Structure of MTGase: in red the regions used for the annotation of the Sapphire plot. **B**: Sapphire plot on the dihedral angles Φ and Ψ for the trajectory obtained by the MD simulation at 355K. At the top it is possible to see the annotation according to the rotations of the dihedral angles, at the bottom the time line and the energetic basins. Labelled with stars, the frames from which the structures analyzed have been taken.

Thus, taking as reference the selected structure no. 6, from the Sapphire plot analysis it is possible to notice that:

- Structure no. 1 differs from no. 6 in the dihedral angles of the amino acids 305-315, 321-323, 327-332, 346-350 (*Fig.65A*)
- Structure no. 2 differs from no. 6 in the dihedral angles of the amino acids 299-300, 304-315, 321-323, 327-332, 346-350 (*Fig.65B*)
- Structure no. 3 differs from no. 6 in the dihedral angles of the amino acids 95-96, 299-300, 304-315, 317, 322, 325-326, 339, 330-332, 346-350 (*Fig.65C*)
- Structure no. 4 differs from no. 6 in the dihedral angles of the amino acids 309-311, 321-323, 327-332, 346-350 (*Fig.65D*)
- Structure no. 5 differs from no. 6 in the dihedral angles of the amino acids 90, 94-98, 304-315, 327-332, 347-350 (*Fig.65E*)

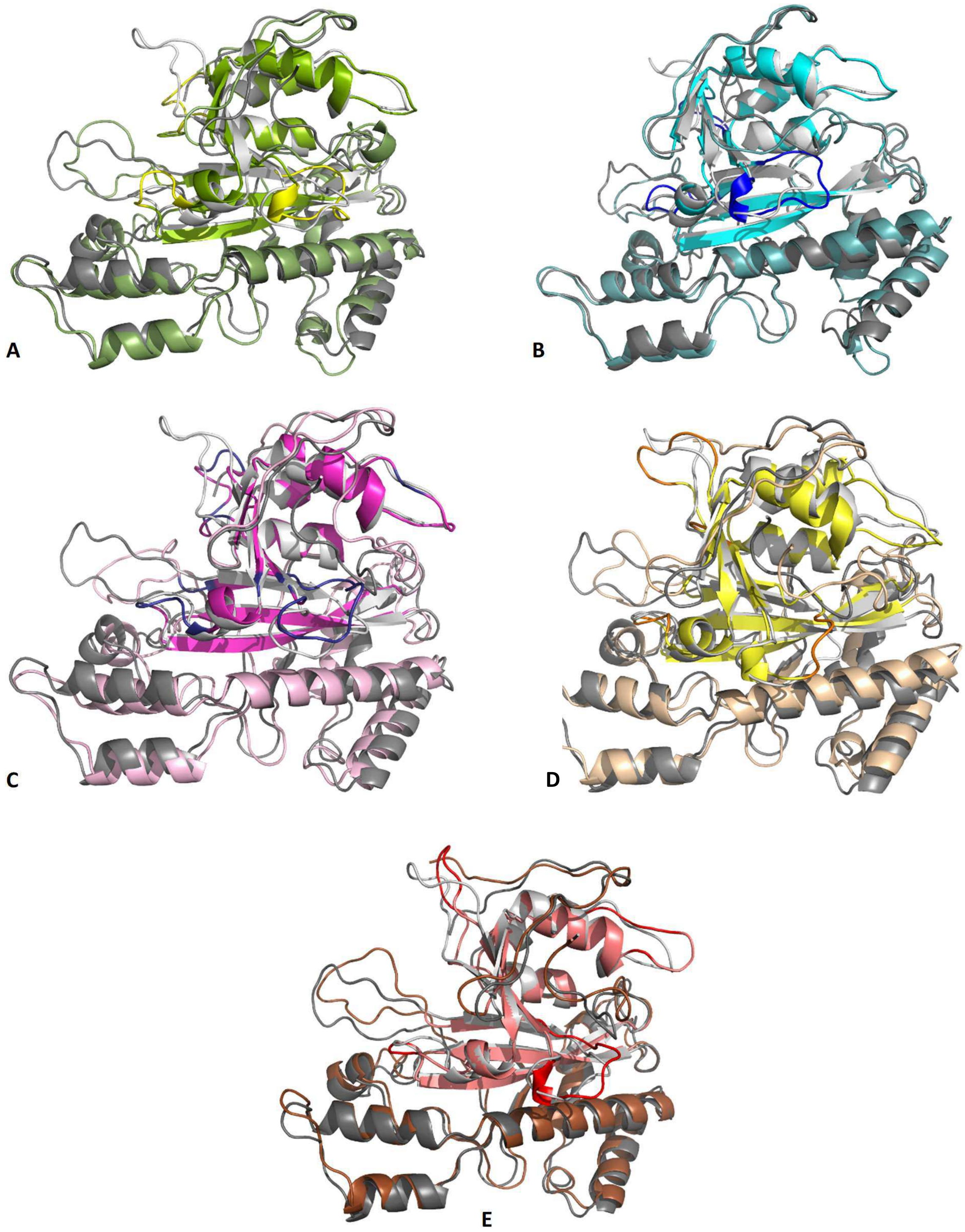


Fig.65: **Structural analysis of the Sapphire plot (MTGase-355K)**. In brighter colors the region to whom belong the dihedral angles used for the annotation of the Sapphire plot. **A**| In blue, **B**: in violet, **C**: in orange, **D**: in violet and **E**: in red respectively the amino acids whose dihedral angles change.

More in general the dihedral angles that show the major differences are that related to the amino acids 94-98, 304-345, 321-323, 327-332, 346-350. However even in this structure these amino acids are all in loop regions.

In conclusion, all the MD simulations performed on the two molecules at 300K give as result a preserved conformation of the catalytic site and its closest areas, and a flexibility of the peripherical loops where, especially in MTGase, structure fluctuation values are higher due to the presence of longer loops than KalbTGase. Thus, from these results appear that KalbTGase at room temperature is a more rigid protein than MTGase.

The MD simulations at higher temperature give a very preliminary result of a possible stability of the molecules also at high temperatures.

Little conformational changes especially in the biggest loops happen, as also little rearrangements of secondary structure in those areas that are far from the active site. In general, MTGase, because of the presence of long loops, shows huge flexibility in these areas, but also lower conformational rearrangement when the temperature rises, above all at 335K, suggesting a major preservation of its structure. KalbTGase is a more compact molecule therefore presents few long loops that show as well high flexibility, but despite of MTGase seems to be generally more adaptable, actually conformational rearrangement far away from the active site domain seems be correlate to the rising of the temperature (*Fig.66*). These structural rearrangements that are more favorite in KalbTGase than in MTGase if from one side could suggest a major adaptability to different substrates not recognized at room temperature, so a loss of specificity, or a preservation of its catalytic activity also at high temperatures, on the other side could also suggest a lack or a reduction of its activity. However, is possible to express the same hypothesis, even if in an opposite way, also for the MTGase: its minor structural rearrangements could suggest a minor adaptability as well as a major preservation of its function. Therefore, experimental assays will be necessary to clarify these aspects.

Another important aspect that is necessary to underline is that these conclusions could be not well supported because at high temperature, only one simulation for each system was performed and this is not enough for giving statistically relevant results.

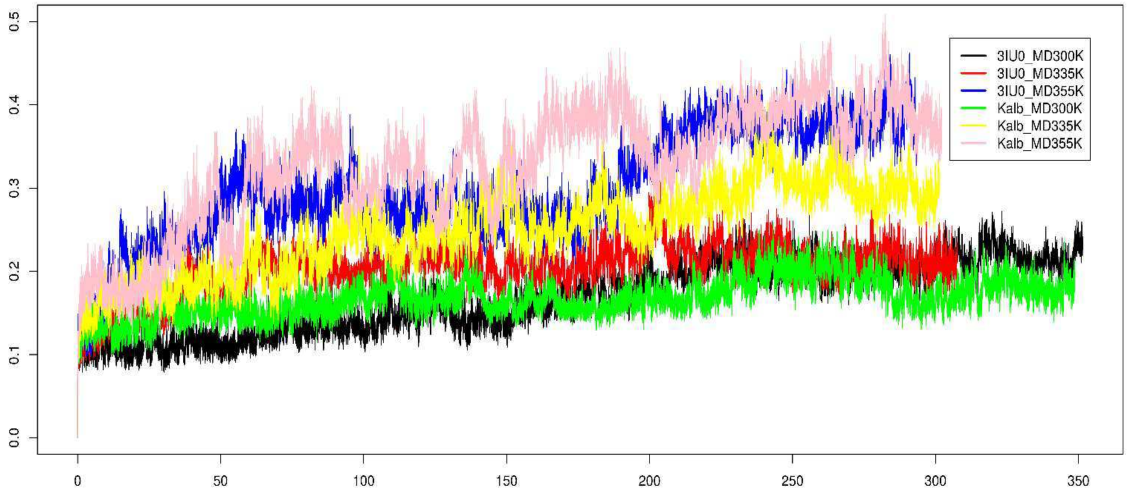


Fig.66: **RMSD comparison between MTGase and KalbTGase MD simulations.** This plot shows a comparison between the RMSD calculated on the Cα related to (in black) the MD simulation of *Streptomyces Mobaraensis* (3IU0) MTGase at 300K, (in red) the MD simulation of 3IU0 MTGase at 335K and (in blue) the MD simulation of 3IU0 MTGase at 355K, (in green) the MD simulation of KalbTGase at 300K, (in yellow) the MD simulation of KalbTGase at 335K and (in pink) the MD simulation of KalbTGase at 355K. On the x-axis lasting of the simulation in numbers of frames, on the y-axis RMSD in Å. It is possible to see how the RMSD profile and the rising of the temperature correlate**.**

9.5 KalbTGase and MTGase: volume pocket analysis

In order to investigate more in the detail the reasons that lead to major specificity of KalbTGase than MTGase a deeper analysis on their active site pocket has been performed. The volume variations that occur during all the MD simulations performed have been analyzed and compared each other in order to find differences and properties.

As reported in paragraph 5.2.d, several tests have been done using different subsampling factors, keeping and discarding the ions. The obtained results, reported in Table 6, show that there is no relevant difference, as expected, between keeping or removing the ions from the trajectories. Actually, in just two runs, only related to MTGase trajectories, the values obtained removing ions are higher than the ones obtained keeping them. Moreover, also the subsampling factor used, strong or weak that was, seems to be not relevant for the volume calculation. However, all the multiple tests carried out on the MD simulations performed at 300K demonstrate that, even if there is a strong similarity between KalbTGase and MTGase active site pockets, the volume of KalbTGase active site is smaller at least of 50 Å^3 than the one of MTGase.

Structure	Conditions	RUN1		RUN2	
		S.20	S.70	S.20	S.70
KalbTGase	IONS	241,7579	243,7756	257,0203	271,1819
	NO IONS	242,603	233,8741	255,9258	270,3378
MTGase	IONS	294,4286	300,9283	331,396	336,7176
	NO IONS	301,7451	302,3015	347,6328	352,0582
		RUN3		RUN4	
KalbTGase	IONS	260,6925	266,0454	279,5574	274,2368
	NO IONS	257,9298	257,9569	278,575	275,8001
MTGase	IONS	359,637	343,9694	250,7454	250,1872
	NO IONS	386,935	377,1677	253,7921	254,6794
			RUN5		Simulations performed at 300K
KalbTGase	IONS		275,399	275,6741	
	NO IONS		275,736	278,9431	
MTGase	IONS		296,965	294,5877	
	NO IONS		308,7204	302,5274	

Table.6: **Pocket volume analysis on MD simulations at 300K.** The table shows the mean values of volume calculated by MDpocket on MTGase and KalbTGase during all the lasting of their 5 simulations performed at 300K. S.20 and S.70 indicate a subsampling used equal to 20 and 70 frames, respectively.

More in the details by MDpocket it was possible to calculate the active site pocket volume for each frame that composes the trajectory, and from these results the mean value and the median value were calculated for each trajectory. The decision to calculate also the median value was due to the big volume fluctuations that were recorded during the trajectory as is possible to see in *fig.67 A* and *B*. Due to the fact that, as seen in table 6, the value obtained with and without ions and using different subsampling were quite the same, a mean value was calculated doing the average of all the mean values obtained for each of the five runs calculated in all the four conditions (subsampling of 20 frames, subsampling of 70 frames, ions in the trajectory, no ions in the trajectory); the same was done for the median value. In this way it was possible to estimate that KalbTGase active site has an average volume equal to 263.65Å^3 (median value = 246.40Å^3), instead MTGase catalytic pocket has an average volume equal to 312.36 Å^3 (median value = 310.50Å^3) (*Fig.67C*).

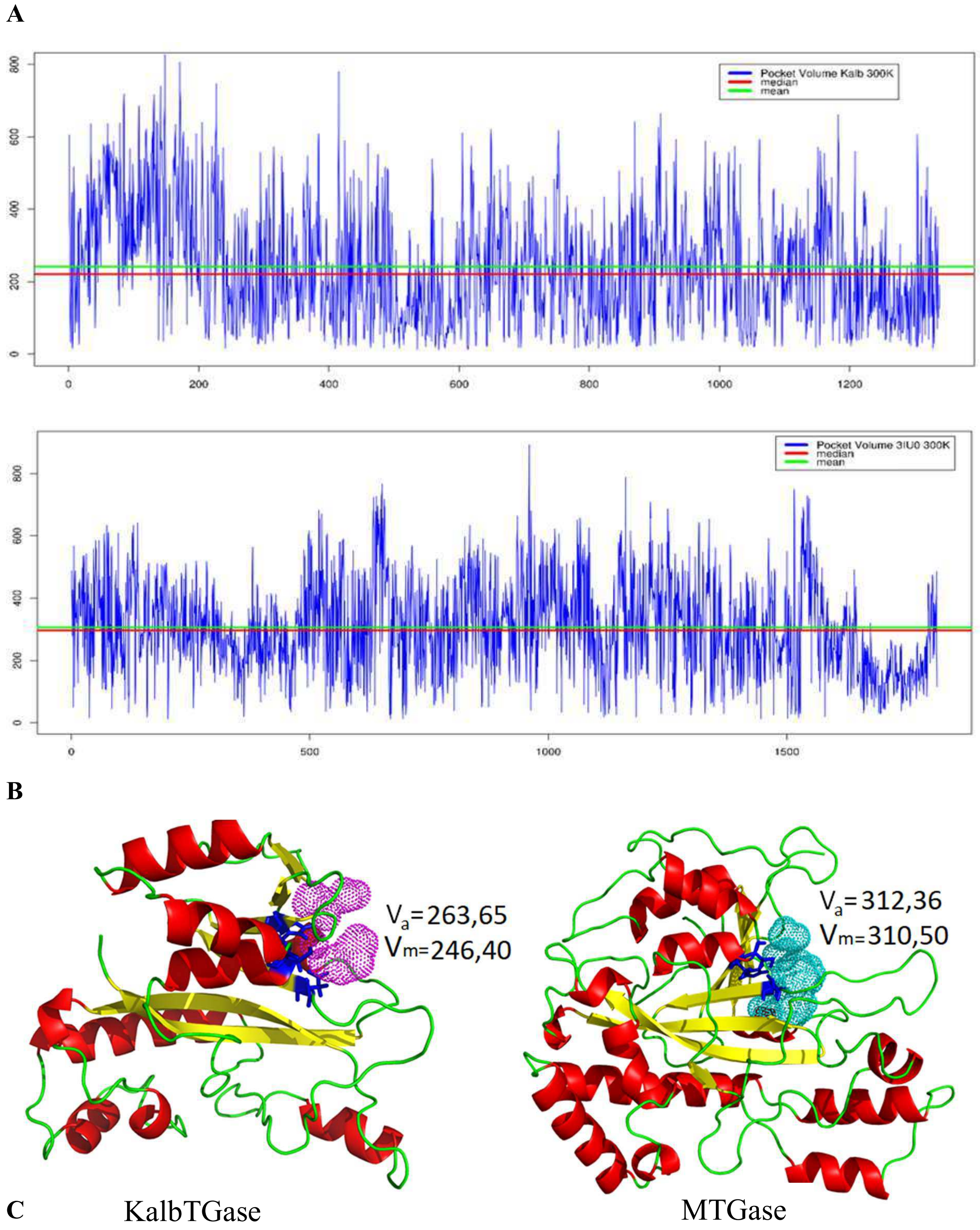


Fig.67: **Volume comparison between KalbTGase and MTGase active sites. A**|Volume variations of KalbTGase active site pocket in the first MD run at 300K **B**: Volume variations of MTGase active site pocket in the first MD run at 300K. In both the plots red and green lines show respectively the mean and the median value. **C**: Representation in cartoon of KalbTGase (on the left) and MTGase structures (on the right), the active site pocket is shown in pink and blue dots respectively. V_a represents the mean value of the pocket volume, V_m the median value.

The discovery that KalbTGase pocket volume is smaller than the one of MTGase could explain the different specificity of these two enzymes. KalbTGase, actually, having a small active site should be more selective in the substrates choice, on the other hand, the larger size of MTGase active site shall ensure a more adaptability to different substrates.

The active site pocket volume analysis has been also performed for the MD simulations carried out at 335K and 355K. Even in this case, no significant differences have been observed among the different conditions used for the analysis. Because of that only the mean and the median value related to the trajectory with ions and with a subsampling equal to 20 (the tightest conditions) will be discussed (*Fig.68*)

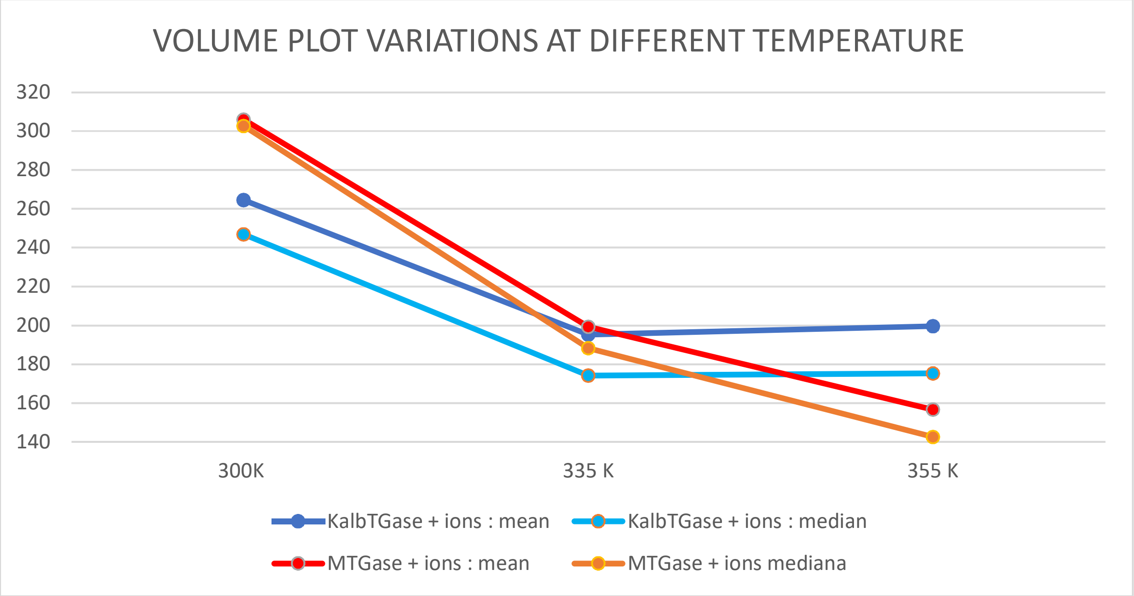


***Fig.68:* Volume plot variations at different temperature.** The variation at the temperature of 300K, 335K and 355K, of the mean and median values of volume are reported for both the enzymes.

From the plot in *fig.68* it is possible to see how for KalbTGase the increase of temperature induces a decrease of the active site pocket volume, that reaches the mean value of 195.28Å^3 at 335K and remains quite the same at 355K, reaching the value of 199.66 Å^3. The calculated median values are lower, being equal to 174.12Å^3 at 335K and to 175.27Å^3 at 355K; however, also the starting value at 300K is lower (the median value at 300K is equal to 246.40Å^3 the average value at the same temperature is instead equal to 263.65Å^3 calculated as). The situation for MTGase, instead, is different; here the volume decreases as the temperature increases. At 335K the mean value calculated for the active site volume is equal to 199.32Å^3 (the median value is quite the same:188.29 Å^3), and at 355K it reaches the value of 156.64Å^3 (median value: 142.54Å^3).

These results are quite in agreement with the other analyses performed on the MD simulations. Actually, from the MD simulations performed at 300K, KalbTGase results less flexible than MTGase and from the active site volume analysis it is possible to see that also its catalytic pocket, during the simulations, is more closed than the one of MTGase. Moreover, the hypothesized less adaptability of MTGase at higher temperature could be in agreement with a reduction of the catalytic pocket and so could suggest also a reduction of its activity. On the other hand, KalbTGase structural rearrangements, that occur when the temperature rises, could explain the reduction and the stabilization of this new shape of the catalytic pocket. However, it is important to remember that the results obtained at high temperature are from only one simulation performed, thus requires further confirmations.

10. Experimental tests

As mentioned in the previous chapters, the enzyme selected for experimental activity assays were two. The first was KalbTGase and the second a hypothetical mTGase from *SaNDy* (organism not disclosed for patent opportunity). After KalbTGase discovery, in agreement with Dr. Steffen and collegues, it was possible to obtain the purified protein and start to test it on food related substrates; in particular, the first substrate tested was the gliadin peptide 56-68, well known for its antigenic property in the celiac patients. However, several bioinformatic screening have also been performed in order to identify which allergenic substrates could be recognized by KalbTGase. In the meantime, studies to clone, express and purify the hypothetical mTGase from *SaNDy* have been performed during my thesis work spent in the Laboratory for Molecular Sensing of Dr. S. D'Auria at the CNR of Avellino under the supervision of Dr. A. Pennacchio.

10.1 KalbTGase: mass spectrometry assays

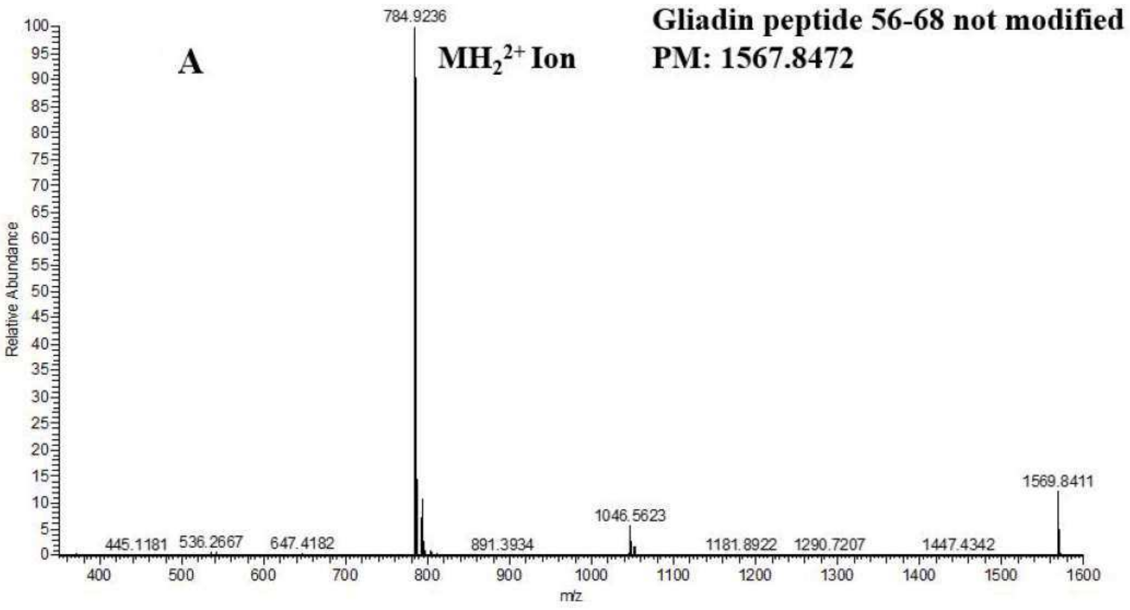
Thanks to the collaboration with Dr. Steffen from Roche Diagnostics, it was possible to obtain a little amount of the purified protein and a control peptide.

Dr. Steffen and colleagues tested thousand sequences as hypothetical KalbTGase substrates, however, in their assays, they did not test sequences showing repetitions in prolines and glutamines. Because of that, it was decided, in collaboration with the spectrometry mass CeSMA-ProBio lab at the CNR of Avellino, that performed the assays, to test the ability of this new enzyme to catalyze transamidation reaction on the gliadin peptide 56-68. After the treatment, the peptide 56-68 and the control peptide have been analyzed by mass spectrometry in order to detect the presence of modifications. As control, the same experiments have been also performed using the MTGase.

The aim was to test the possibility to exploit its ability to form cross-link bonds between glutamine and lysine residues in order to detoxify flour, as done in precedent studies on the MTGase (Mazzeo et al. 2013).

The experimental results confirmed that MTGase is able to catalyze the reaction of transamidation that lead to the formation of a modified peptide 56-68 and a modified control peptide. Actually, the two mass spectra pointed out a difference of 185m/z (corresponding to an addition of a spermine molecule) between the molecular weight of the gliadin peptide in the native form (1567.8472) (*Fig.69A*) and the one modified (1753.0362) (*Fig.69B*). Analogues results are showed for the ROCHE peptide control, for which the molecular weight before the reaction is 1317.5924 and after 1502.7798 because of the addition of the spermine molecule (data not shown).

In the case of KalbTGase, the transamidase reaction modified only the ROCHE control peptide (m/z 1502.7796), whereas the gliadin peptide has not been modified as is possible to see from its molecular weight 1567.8536 corresponding to the native form (*Fig.69C*). These results indicate that the MTG is less specific in the recognition of the peptide substrates, in fact, it is able to recognize both the substrates. KalbTGase is instead more specific and does not modify the gliadin peptide 56-68.



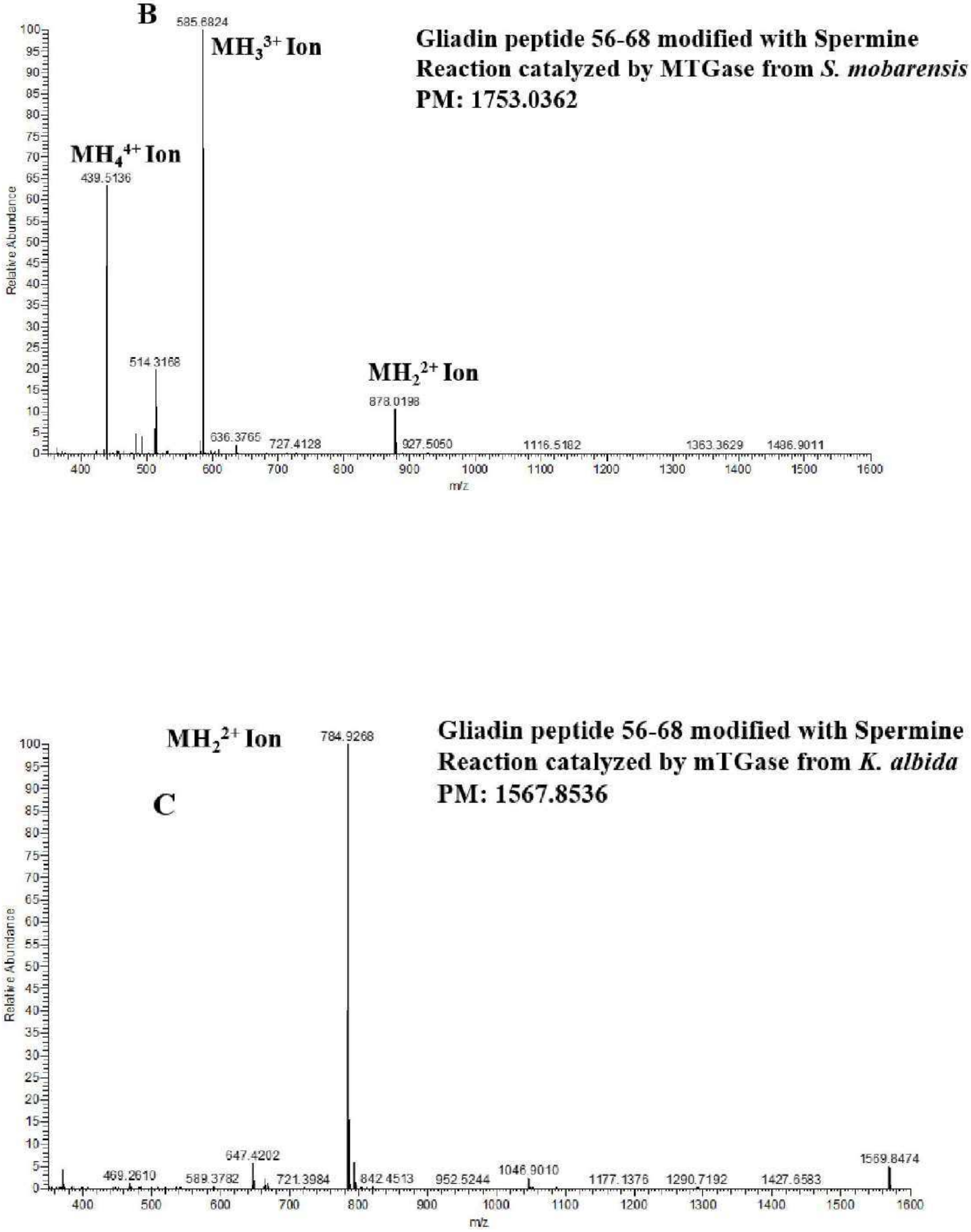
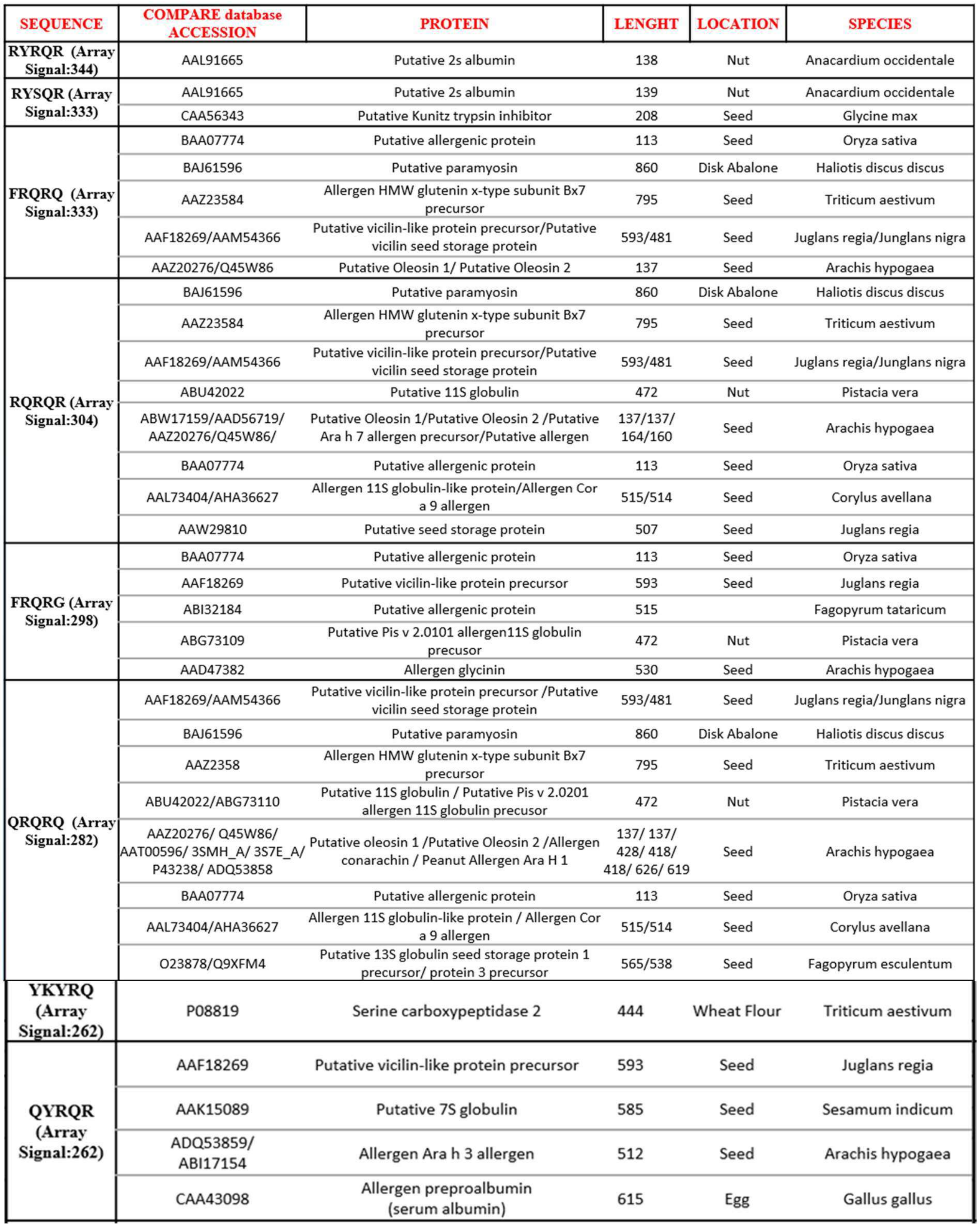


Fig.69: **Mass spectra of the gliadin peptide 56-68 and its reaction mTGase. A**: Mass spectrum of the gliadin peptide 56-68 in its native form, **B**: after incubation with spermine by MTGase **C**: and by KalbTGase.

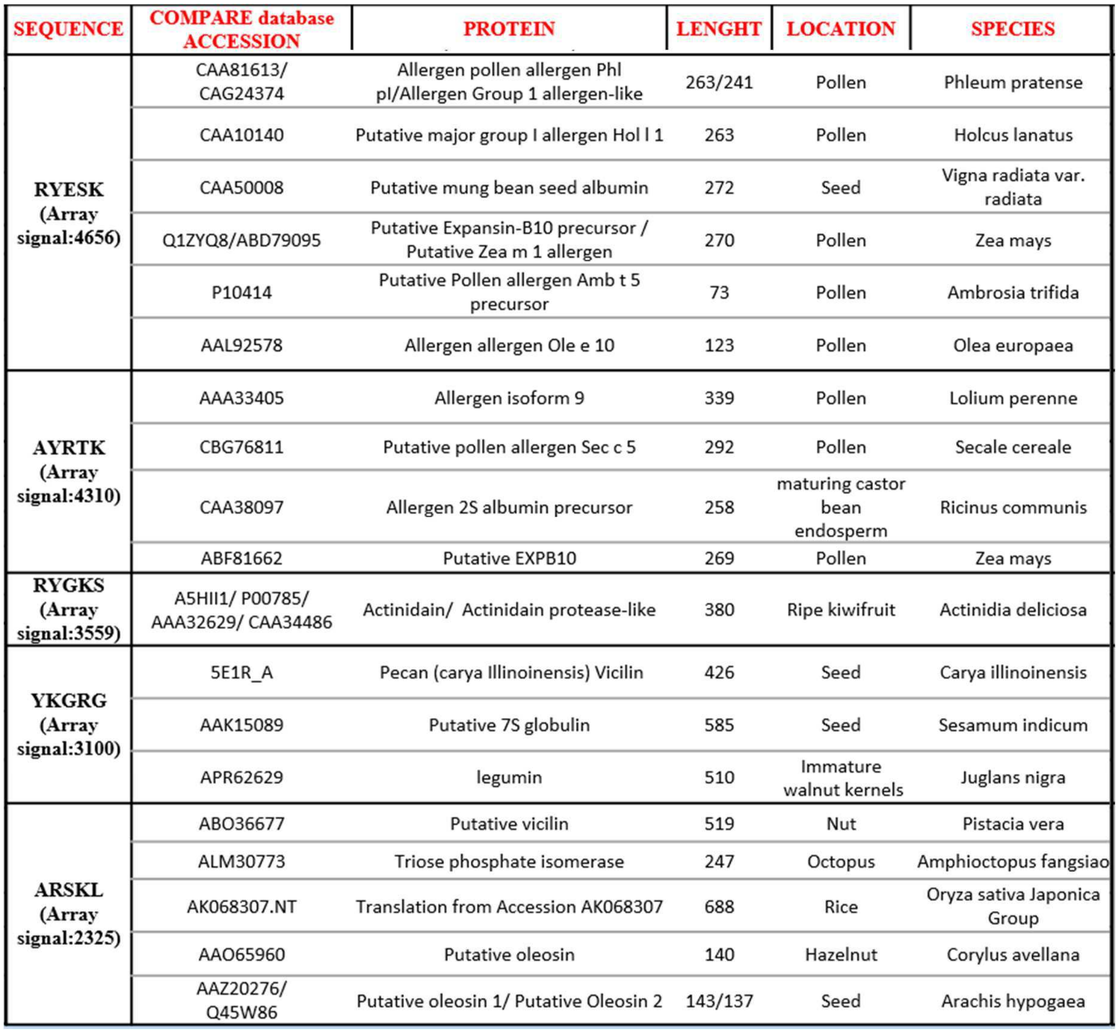
In order to search allergenic epitopes that could react with KalbTGase, COMPARE and IEDB.org databases were used to perform a comparison between the antigenic sequences present in these databases and the Glutamine/Lysine 5-mer best substrates for KalbTGase found by Steffen and colleagues (Steffen et al., 2017). In Table.7A and Table.7B, results related to COMPARE database are shown.

Table.7A



SEQUENCE	COMPARE database ACCESSION	PROTEIN	LENGHT	LOCATION	SPECIES
RYRQR (Array Signal:344)	AAL91665	Putative 2s albumin	138	Nut	Anacardium occidentale
RYSQR (Array Signal:333)	AAL91665	Putative 2s albumin	139	Nut	Anacardium occidentale
	CAA56343	Putative Kunitz trypsin inhibitor	208	Seed	Glycine max
FRQRQ (Array Signal:333)	BAA07774	Putative allergenic protein	113	Seed	Oryza sativa
	BAJ61596	Putative paramyosin	860	Disk Abalone	Haliotis discus discus
	AAZ23584	Allergen HMW glutenin x-type subunit Bx7 precursor	795	Seed	Triticum aestivum
	AAF18269/AAM54366	Putative vicilin-like protein precursor/Putative vicilin seed storage protein	593/481	Seed	Juglans regia/Junglans nigra
	AAZ20276/Q45W86	Putative Oleosin 1/ Putative Oleosin 2	137	Seed	Arachis hypogaea
RQRQR (Array Signal:304)	BAJ61596	Putative paramyosin	860	Disk Abalone	Haliotis discus discus
	AAZ23584	Allergen HMW glutenin x-type subunit Bx7 precursor	795	Seed	Triticum aestivum
	AAF18269/AAM54366	Putative vicilin-like protein precursor/Putative vicilin seed storage protein	593/481	Seed	Juglans regia/Junglans nigra
	ABU42022	Putative 11S globulin	472	Nut	Pistacia vera
	ABW17159/AAD56719/ AAZ20276/Q45W86/	Putative Oleosin 1/Putative Oleosin 2 /Putative Ara h 7 allergen precursor/Putative allergen	137/137/ 164/160	Seed	Arachis hypogaea
	BAA07774	Putative allergenic protein	113	Seed	Oryza sativa
	AAL73404/AHA36627	Allergen 11S globulin-like protein/Allergen Cor a 9 allergen	515/514	Seed	Corylus avellana
	AAW29810	Putative seed storage protein	507	Seed	Juglans regia
FRQRG (Array Signal:298)	BAA07774	Putative allergenic protein	113	Seed	Oryza sativa
	AAF18269	Putative vicilin-like protein precursor	593	Seed	Juglans regia
	ABI32184	Putative allergenic protein	515		Fagopyrum tataricum
	ABG73109	Putative Pis v 2.0101 allergen11S globulin precusor	472	Nut	Pistacia vera
	AAD47382	Allergen glycinin	530	Seed	Arachis hypogaea
QRQRQ (Array Signal:282)	AAF18269/AAM54366	Putative vicilin-like protein precursor /Putative vicilin seed storage protein	593/481	Seed	Juglans regia/Junglans nigra
	BAJ61596	Putative paramyosin	860	Disk Abalone	Haliotis discus discus
	AAZ2358	Allergen HMW glutenin x-type subunit Bx7 precursor	795	Seed	Triticum aestivum
	ABU42022/ABG73110	Putative 11S globulin / Putative Pis v 2.0201 allergen 11S globulin precusor	472	Nut	Pistacia vera
	AAZ20276/ Q45W86/ AAT00596/ 3SMH_A/ 3S7E_A/ P43238/ ADQ53858	Putative oleosin 1 /Putative Oleosin 2 /Allergen conarachin / Peanut Allergen Ara H 1	137/ 137/ 428/ 418/ 418/ 626/ 619	Seed	Arachis hypogaea
	BAA07774	Putative allergenic protein	113	Seed	Oryza sativa
	AAL73404/AHA36627	Allergen 11S globulin-like protein / Allergen Cor a 9 allergen	515/514	Seed	Corylus avellana
	O23878/Q9XFM4	Putative 13S globulin seed storage protein 1 precursor/ protein 3 precursor	565/538	Seed	Fagopyrum esculentum
YKYRQ (Array Signal:262)	P08819	Serine carboxypeptidase 2	444	Wheat Flour	Triticum aestivum
QYRQR (Array Signal:262)	AAF18269	Putative vicilin-like protein precursor	593	Seed	Juglans regia
	AAK15089	Putative 7S globulin	585	Seed	Sesamum indicum
	ADQ53859/ ABI17154	Allergen Ara h 3 allergen	512	Seed	Arachis hypogaea
	CAA43098	Allergen preproalbumin (serum albumin)	615	Egg	Gallus gallus

Table.7B



SEQUENCE	COMPARE database ACCESSION	PROTEIN	LENGHT	LOCATION	SPECIES
RYESK (Array signal:4656)	CAA81613/ CAG24374	Allergen pollen allergen Phl pI/Allergen Group 1 allergen-like	263/241	Pollen	Phleum pratense
	CAA10140	Putative major group I allergen Hol l 1	263	Pollen	Holcus lanatus
	CAA50008	Putative mung bean seed albumin	272	Seed	Vigna radiata var. radiata
	Q1ZYQ8/ABD79095	Putative Expansin-B10 precursor / Putative Zea m 1 allergen	270	Pollen	Zea mays
	P10414	Putative Pollen allergen Amb t 5 precursor	73	Pollen	Ambrosia trifida
	AAL92578	Allergen allergen Ole e 10	123	Pollen	Olea europaea
AYRTK (Array signal:4310)	AAA33405	Allergen isoform 9	339	Pollen	Lolium perenne
	CBG76811	Putative pollen allergen Sec c 5	292	Pollen	Secale cereale
	CAA38097	Allergen 2S albumin precursor	258	maturing castor bean endosperm	Ricinus communis
	ABF81662	Putative EXPB10	269	Pollen	Zea mays
RYGKS (Array signal:3559)	A5HII1/ P00785/ AAA32629/ CAA34486	Actinidain/ Actinidain protease-like	380	Ripe kiwifruit	Actinidia deliciosa
YKGRG (Array signal:3100)	5E1R_A	Pecan (carya Illinoinensis) Vicilin	426	Seed	Carya illinoinensis
	AAK15089	Putative 7S globulin	585	Seed	Sesamum indicum
	APR62629	legumin	510	Immature walnut kernels	Juglans nigra
ARSKL (Array signal:2325)	ABO36677	Putative vicilin	519	Nut	Pistacia vera
	ALM30773	Triose phosphate isomerase	247	Octopus	Amphioctopus fangsiao
	AK068307.NT	Translation from Accession AK068307	688	Rice	Oryza sativa Japonica Group
	AAO65960	Putative oleosin	140	Hazelnut	Corylus avellana
	AAZ20276/ Q45W86	Putative oleosin 1/ Putative Oleosin 2	143/137	Seed	Arachis hypogaea

Table.7 **Allergens that could react with KalbTGase. A|** Allergens which show a similarity with the best Gln 5-mer substrates found by ROCHE **B|** Allergens which show a similarity with the best Lys 5-mer substrates found by ROCHE.

From this screening, it results that KalbTGase could react with many different substrates, however among them, the most interesting from the allergy prevention point of view result nuts, walnuts and peanuts. Also a putative reaction with the kiwi allergens was judge as interesting. Therefore, now further analyses at different conditions are ongoing with gliadin and other different substrates of interest, using both KalbTGase and MTGase.

10.2 A putative mTGase from *SaNDy*: experimental analysis

The hypothetical mTGase sequence from *SaNDy* (SandyTGase) has been available since March 2018. After its discovery the sequence was analyzed, and a good model was created. This sequence is more similar to the MTGase and due to that it was decided to launch immediately experimentally activities to characterize it. It is hoped that maybe with a higher similarity it is possible to find a protein which could be less selective for specific substrates, so it could be possible its application also on the gliadin substrate.

Actually, the protein has been expressed and purified, thus the next step will be to test its activity on different substrates and comparing it with the one of the MTGase.

10.2.a SandyTGase: protein expression results

As described more in details in the paragraph 6.2.a, the gene encoding for this protein was cloned in pET-22b(+), a bacterial vector that encodes a signal sequence for inducible expression of proteins in the periplasm. The vector was expressed in two competent cell lines BL21DE3 and TOP10. Using these cells, several tests of induction and grown at different condition have been performed, however due to insufficient results obtained with the TOP10 cells (data not shown), BL21DE3 cell line was preferred. In order to find the good percentage of grow at which it was better to induce the expression of the vector, a growth curve for the BL21DE3 cell line was performed (*Fig.70*) and the OD equal to 0.7 chosen as the best to start the induction.

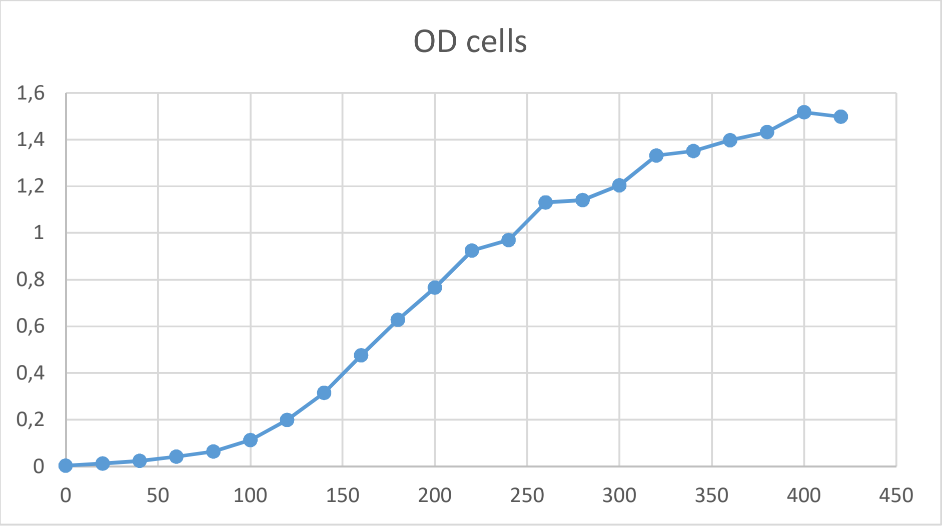


Fig.70: **BL21DE3 growth curve.** On the x-axis the growth time on the y-axis the OD.

Several tests were performed to find the best concentration of IPTG (*Fig.71*), a molecular mimic of allolactose, a lactose metabolite that triggers transcription of the lac operon, and it is therefore used to induce protein expression where the gene is under the control of the lac operator. From the results it was chosen the concentration of 1mM IPTG.

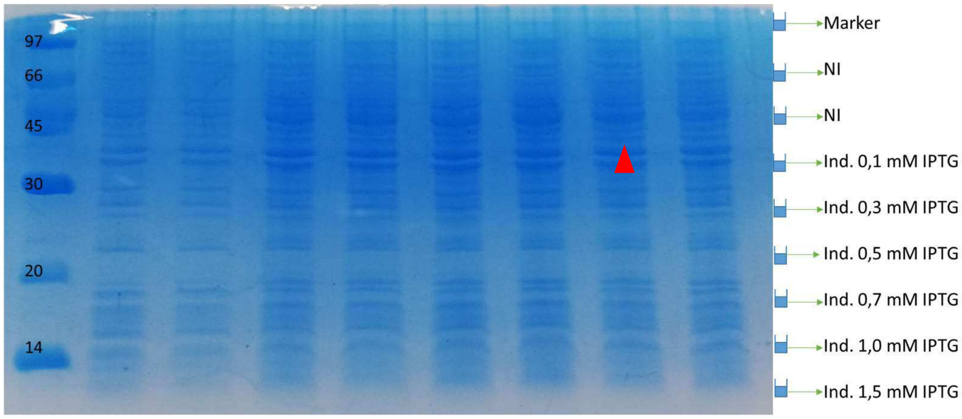


***Fig.71:* SDS PAGE results: tests for IPTG concentration to use**. SDS-PAGE on samples of cells induced with different concentration of IPTG, the wells have been loaded using the sample order indicated at the right of the figure. A specific number labels the lines of the markers according to their weight. The protein of interest has a molecular weight of about 37KDa and is pointed by the red arrow.

10.2.b SandyTGase: purification results

In order to purify the protein, an osmotic shock was performed, and the obtained fractions were purified by anion-exchange chromatography (see paragraph 6.2.b). However, to obtain the best condition of purification, also the best temperature and the best duration of the induction was analyzed.

Osmotic shock procedure was applicated on culture induced at 25°C and 37°C with 1mM IPTG; from the results it was possible to appreciate that the best temperature is 25°C. As it is possible to see from the *fig.72* showing the SDS-PAGE results, at this temperature the band obtained in the sample of the periplasmatic fraction is thicker than the band obtained in the same fraction at 37°C. Similarly, several tests on the best duration of the induction were performed and finally the best conditions used for the induction were: 1mM IPTG, 25°C for 30h (*Fig.73*)

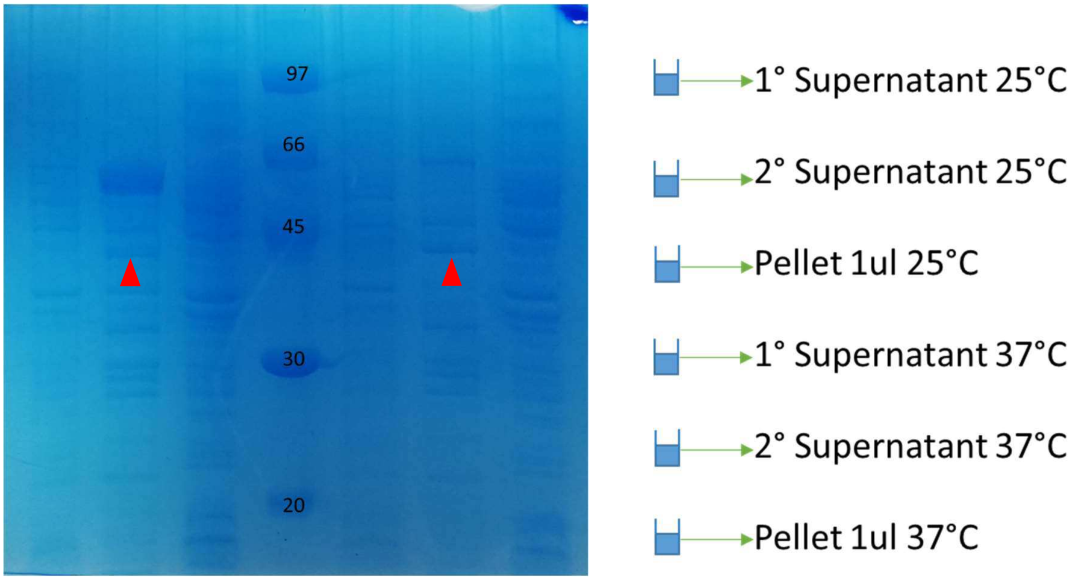


Fig.72: **SDS PAGE results: tests for temperature of induction to use**. SDS-PAGE on samples after purification by osmotic shock of cells induced with 1mM IPTG overnight. The wells have been loaded using the sample order indicated at the right of the figure. A specific number labels the lines of the markers according to their weight. The protein of interest has a molecular weight of about 37KDa and is pointed by the red arrows in the two samples of periplasmatic fraction.

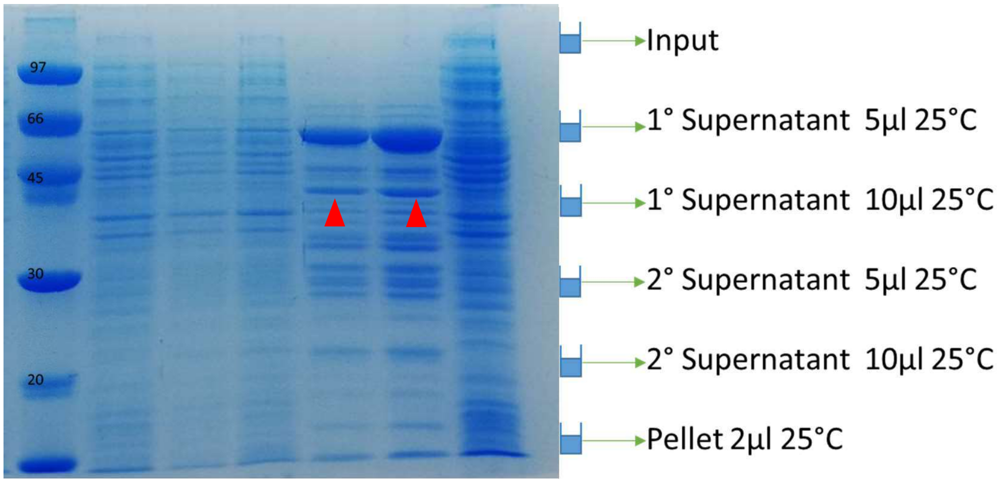


***Fig.73:* SDS PAGE results: protein purification by osmotic shock**. SDS-PAGE on samples after purification by osmotic shock of cells induced with 1mM IPTG at 25°C for 30h. The wells have been loaded using the sample order indicated at the right of the figure. A specific number labels the lines of the marker according to their weight. The protein of interest has a molecular weight of about 37KDa and is pointed by the red arrows in the two samples of periplasmatic fraction.

The periplasmatic fraction so obtained, after verifying to contain the protein of interest (*Fig.73*), was loaded on a HiPrep DEAE column for the anion exchange chromatography. The estimate isoelectric point of our protein was 6.8, so it was expected that it was eluted in the first fractions, nevertheless, it was decided to analyze by SDS-PAGE all the fractions which showed a peak in the OD read in order to do not lose the target protein. From the analysis it was possible to collect some fractions where the protein was present in a pure form (*Fig.75*).

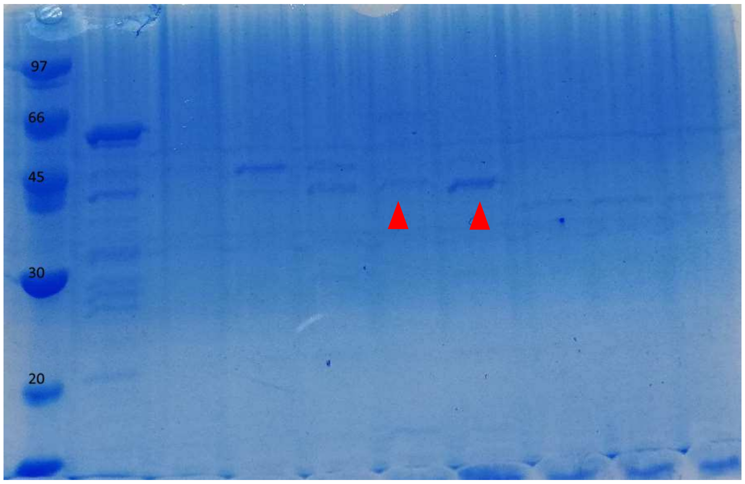


Fig.75: SDS PAGE results: protein purification by anion exchange chromatography. SDS-PAGE on several fractions collected after the purification by DEAE column for the anion exchange chromatography. The sample loaded is the periplasmatic fraction obtained by osmotic shock of BL21DE3 cells induced with 1mM IPTG at 25°C for 30h. A specific number labels the lines of the marker according to their weight. The protein of interest has a molecular weight of about 37KDa and is pointed by the red arrows.

Only the fractions where the protein was pure were collected, the others instead were discarded. Now the protein is in a test phase where is necessary to collect all the data required to analyze its activity.

CONCLUSION AND FUTURE WORKS

11.Work summary, conclusions and related future works.

Searches based on the MTGase and on the hTGase2 sequences evidence a high number of hypothetical mTGases. Among them, it is possible to find different characteristics in terms of length, presence of domains, position of amino acids with expectable functional roles. It should be remarked that most of these sequences are derived from whole-genome sequencing studies and in most cases, there is no experimental evidence of the existence at protein level.

Due to their very high number, a proper classification of all these sequences was necessary in order to divide them on the basis of their features and make easier their characterization.

To do that, an iterative clustering procedure based on phylogenetic trees construction was applied. Thanks to this approach, it was possible to divide all the microbial sequences annotated as hypothetical transglutaminase or reporting as having a putative TGase core in five main groups. The results obtained showed that some of these sequences are very similar to the MTGase, others preserve the catalytic core typical of the hTGase2 or of the *Bacillus* TGl. Others, instead, are completely different and may represent new forms of microbial TGase.

A very extensive motifs research has corroborated this classification and has highlighted the peculiar sequence features of each group, underlining also the fact that different mTGases could exist within the same phylum, and in some cases also in the same organism, as seen in the case of *K. albida* which presents two kind of mTGase: one MTGase-like and the other hTGase2-like.

From the obtained groups, it has been possible to select about 300 sequences, which show many differences in terms of composition of the catalytic pocket, sequence length, and microorganism of origin. By different criteria and bioinformatics tools, some of those sequences have been selected, satisfactorily modelled, and the models were assessed.

Analyzing and comparing the secondary structures obtained for the sequences belonging to groups II, III, and IV (TGl-like, MTGase-like and hTGase2-like respectively) and the ones already present in PDB, it was possible to detect peculiarity of the catalytic core structures for each group. In particular, it was detected that the first catalytic residue (C/S) is always at the N terminal of an α-helix linked to a β-sheet of at least three antiparallel β-strands, with the second and third strand hosting each another catalytic residue. The two catalytic regions, i.e. the α-helix and the β-sheet, are differently linked in the three groups. For group II the linker is composed by few α-helices and several β-strands. For group III, the two regions are linked by a very long portion composed by eight α-helices. For group V the catalytic helix may be linked directly to the catalytic β-sheet or by means of one additional β-strand.

The topology of the three groups open the landscape of new evolutive speculations as the possibility that the mTGase human-like (group V) and MTGase-like (group II), arise from a TGl-like (group II) ancient form of TGl. Moreover, looking at the little spread of the form of MTGase-like along the different phylogenetic phyla, it may be possible that this is a less-specialized form of TGase, that has evolved, perhaps starting from the ancient form previous mentioned, in a different way than the one human-like.

However, because one of the aims of this project was also to characterize by experimental activity assays proteins showing features suggesting that they could become an alternative to the MTGase, among all these sequences two were selected for experimental characterization and deeper structural investigations: the uncharacterized protein MTGase-like from *K. albida* (28% sequence identity) and the hypothetical protein-glutamine gamma-glutamyltransferase from *SaNDy* available since 2018 showing a high similarity with MTGase.

The selected candidate for the experimental characterization, i.e. the TGase from *K. Albida,* was satisfactory modelled as demonstrated by the MD simulation results, confirming the overall good quality of the modelling procedure followed. The discovery by ROCHE of the same enzyme made unnecessary its experimental characterization, so this step was overcome going directly on testing it on other proteins of interest for their involvement in food related pathologies. However, several MD simulations on both the structure MTGase and KalbTGase were performed in order to investigate differences and similarity between these two proteins, in terms of stability of their active site, regions more flexible or more rigid, structural rearrangements and behavior at different condition.

Until now, from the MD simulations performed at 300K it is visible a major flexibility of the MTGase than of KalbTGase; this observation, together with the analyses performed on the active site pocket which show as KalbTGase catalytic pocket is at least 50 Å smaller than the one of MTGase, could explain a lower specificity of the MTGase than KalbTGase.

MD simulations performed at higher temperature show as MTGase is less affected of structural rearrangement above all at 335K where its RMSD looks the same of the one at 300K. However, pocket volume analyses show its active site become smaller as the temperature raising, suggesting that probably this lack of rearrangement may result in a less adaptability and so in a lacking or decreasing of activity. Moreover, the same simulations performed on KalbTGase show as this enzyme is affected of several rearrangements in structures far away from the active site, its RMSD increases with increasing the temperature, but the system after a first phase of adaptation reaches stability, especially at 335K. Actually, the volume analyses of the active site pocket show that the volume of catalytic pocket decreases with increasing the temperature but keeps the same value at both 335K and 355K. These results could suggest a mayor adaptability of this enzyme at higher temperature and probably a preservation, at least in part, of its catalytic activity. However, experimental activity assays will be necessary to analyze this aspect.

Until now this enzyme has been tested in comparison with MTGase, on the gliadin peptide 56-68, but being a very selective mTGase, KalbTGase has not shown any activity at these conditions on this peptide. Further tests at different conditions are necessary.
Screening to find better allergenic food related substrates have been done, and now several tests are ongoing.

Moreover, a new experimentally activity has been launched for the characterization of another novel mTGase, extracted by *SaNDy*, which as mentioned before, is more similar to the MTGase. This because maybe with a higher similarity it is possible to find a protein which could be less selective for specific substrates, so it could be possible its application also on the gliadin substrate. So far, the protein has been cloned, expressed and purified. Now it is necessary to check enzymatic activity and the best activity conditions and start to test it on substrates of interest.

In conclusion the present work has brought to the first classification of the mTGase sequences and to the identification of the most representative features that distinguish the ones from the others. The MD simulations performed on KalbTGase, instead, have suggested some explanations to the major specificity of this enzyme than MTGase, paving the way for novel experimental conditions. Moreover, the substrates screening performed, has allowed to find novel possible substrates, on which this enzyme could be employed for the allergenicity reduction. Last but not least, the present work has pointed out the presence of a putative novel form of mTGase that, even if belonging to another genus, for its similarity could be an alternative to the MTGase actually in use: the one from *SaNDy*.

Site-ography

Clustal Omega: https://www.ebi.ac.uk/Tools/msa/clustalo/

CNN: http://edition.cnn.com/2017/03/01/health/gluten-free-diet-history-explainer/inde

COMPARE database: http://comparedatabase.org/

EBI database BLAST tools: https://www.ebi.ac.uk/Tools/sss/ncbiblast/

ExPASy: https://www.expasy.org/tools/

FigTree: http://tree.bio.ed.ac.uk/software/figtree/

Financial Time: https://www.ft.com/content/4ec0f2f2-2c0a-11e7-9ec8-168383da43b7

Financial Time: https://www.ft.com/content/5348432e-1a13-11e7-bcac-6d03d067f81f

Gluten-Free Products Market Size & Share, Industry Report, 2014-2025: https://www.grandviewresearch.com/industry-analysis/gluten-free-products-market

IEDB.org database: https://www.iedb.org/

MICROBESonline database: http://www.microbesonline.org/

MicroScope database: http://www.genoscope.cns.fr/agc/microscope/home/index.php

MUSCLE: https://www.ebi.ac.uk/Tools/msa/muscle/

NCBI database BLAST tools: https://blast.ncbi.nlm.nih.gov/Blast.cgi

Osservatorio AGR: http://www.osservatorioagr.eu/prodotti-senza-glutine-un-mercato-che-in-italia-cresce-del-30-lanno/

PDB: https://www.rcsb.org/

PFAM database: https://pfam.xfam.org/

PhyML: http://www.atgc-montpellier.fr/phyml/

Piz Daint supercomputer: https://www.cscs.ch/computers/piz-daint/

POCASA 1.1: http://altair.sci.hokudai.ac.jp/g6/service/pocasa/

PROCHECK: https://servicesn.mbi.ucla.edu/PROCHECK/

Pro-Origami: http://munk.csse.unimelb.edu.au/pro-origami/

ProSA-web: https://prosa.services.came.sbg.ac.at/prosa.php

QMEAN: https://swissmodel.expasy.org/qmean/

T-Coffee: http://tcoffee.crg.cat/apps/tcoffee/do:expresso

UniProt: https://www.uniprot.org/

UniProt: https://www.uniprot.org/

BIBLIOGRAPHY

AbdAlla S., Lother H., Langer A., el Faramawy Y., Quitterer U. (2004) Factor XIIIA Transglutaminase Crosslinks AT1 Receptor Dimers of Monocytes at the onset of Atherosclerosis. Cell. 119(3):343-54.

Abd-Rabo F.H.R., El-Dieb S. M., Abd-El-Fattah A.M. and Sakr S.S. (2010): Natural State Changes of Cows' and Buffaloes' Milk Proteins Induced by Microbial Transglutaminase. Journal of American Science. 6(9):612-620

Abraham M.J., Murtolad T., Schulzb R., Pálla S., Smith J. C., Hessa B., Lindahla E. (2015): GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 1–2. 19–25.

Ahlborn G.J., Pike O.A., Hendrix S.B., Hess W.M., Huber C.S. (2005): Sensory, mechanical, and microscopic evaluation of staling in low-protein and gluten-free breads, Cereal Chem. 82:328–335.

Akaike H. (1973): Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F, editors. Second international symposium on information theory. Budapest (Hungary): Akademiai Kiado. p. 267–281

Altschul S.F., Gish W., Miller W., Myers E.W. and Lipman D.J. (1990): Basic local alignment search tool. J. Mol. Biol. 215:403-410.

Ando H., Adachi M., Umeda K., Matsuura A., Nonaka M., Uchio R., Tanaka H. & Motoki M. (1989): Purification and characteristics of a novel transglutaminase derived from microorganisms. Agricultural and Biological Chemistry. 53(10):2613-2617. DOI: 10.1080/00021369.1989.10869735

Anisimova M. and Gascuel O. (2006): Approximate Likelihood-Ratio Test for Branches: A fast, accurate, and powerful alternative. systematic biology. 55(4):539-552.

Anisimova M., Gil M., Dufayard JF., Dessimoz C. and Gascuel O (2011): Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Systematic Biology. 60(5):685-99.

Aune D., Keum N., Giovannucci E., et al., (2016): Whole grain consumption and risk of cardiovascular disease, cancer, and all cause and cause specific mortality: systematic review and dose-response meta-analysis of prospective studies, BMJ. 353:i2716. doi:10.1136/bmj.i2716.

Bailey T. L. and Elkan C. (1994): Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. AAAI Press.28-36

Benkert P., Tosatto S.C.E. and Schomburg D. (2008): QMEAN: A comprehensive scoring function for model quality assessment. Proteins: Structure, Function, and Bioinformatics. 71(1):261-277

Berendsen H. J. C., Postma J. P. M., van Gunsteren W. F., Di Nola A., and Haak J. R. (1984): Molecular dynamics with coupling to an external bath. Journal of Chemical Physics. 81:3684-3690

Berman H., Henrick K., Nakamura H. and Markley J. L. (2007): The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res. 35: D301–D303. DOI:10.1093/nar/gkl971

Berti Rocha Mendes F., Hissa-Elian A., Milanez Morgado de Abreu M. A. (2013): Scaff Gonçalves V., Review: dermatitis herpetiformis. An Bras Dermatol. 88(4):594–599. DOI: 10.1590/abd1806-4841.20131775

Biesiekierski J.R. (2017): "What is gluten?". J Gastroenterol Hepatol (Review). 32 Suppl 1: 78–81. doi:10.1111/jgh.13703. PMID 28244676.

Blöchliger N., Vitalis A. and Caflisch A. (2014): High-resolution visualisation of the states and pathways sampled in molecular dynamics simulations. Scientific Reports. 4:6264. DOI:10.1038/srep06264

Bognar P., Nemeth I., Mayer B., Haluszka D., Wikonkal N., Ostorhazi E., John S., Paulsson M., Smyth N., Pasztoi M., Buzas E. I., Szipocs R., Kolonics A., Temesvari E. and Karpa S. (2014): Reduced inflammatory threshold indicates skin barrier defect in transglutaminase 3 knockout mice. Journal of Investigative Dermatology 134,105–111, DOI:10.1038/jid.2013.307

Camolezi Gaspar A. L., Pedroso de Góes-Favoni S. (2015): Action of microbial transglutaminase (MTGase) in the modification of food proteins: A review. Food Chemistry 171:315–322

Camposa N., Castañón S., Urretab I., Santosa M., Torné J.M. (2013): Rice transglutaminase gene: identification, protein expression, functionality, light dependence and specific cell location. Plant Sci 205–206:97–110

Candi E., Oddi S., Paradisi A., Terrinoni A., Ranalli M., Teofoli P., Citro G., Scarpato S., Puddu P., and Gerry Melino (2002): Expression of Transglutaminase 5 in Normal and Pathologic Human Epidermis. J. Invest. Dermatol. 119, issue 3, 670-677.

Caputo I., Lepretti M., Martucciello S. and Esposito C. (2010): Enzymatic strategies to detoxify gluten: implications for celiac disease. Enzyme Research. Volume 2010, Article ID 174354, 9 pages. DOI:10.4061/2010/174354.

Carvajal P., Gibert J., Campos N., Lopera O., Barbera E., Torne J.M., Santos M. (2011): Activity of maize transglutaminase overexpressed in Escherichia coli inclusion bodies: an alternative to protein refolding. Biotechnol Prog 27(1):232–240

Cho S.-Y., Choi K., Jeon J.-H., Kim C.-W., Shin D.-M., Lee J. B., Lee S. E, Kim C.-S., Park J.-S., Jeong E. M., Jang G.-Y., Song K.-Y., Kim I.-G. (2010): Differential alternative splicing of human transglutaminase 4 in benign prostate hyperplasia and prostate cancer. Experimental & Molecular Medicine 42,310–318.

Cook C. E., Lopez R., Stroe O., Cochrane G., Brooksbank C., Birney E. and Apweiler R. (2018): The European Bioinformatics Institute in 2018: tools, infrastructure and training. Nucleic Acids Research. gky1124. DOI: 10.1093/nar/gky1124

Costabile A., Bergillos-Meca T., Landriscina L., Bevilacqua A., Gonzalez-Salvador I., Corbo M.R., Petruzzi L., Sinigaglia M. and Lamacchia C. (2017): An in vitro fermentation study on the effects of Gluten Friendly ^{TM} bread on microbiota and short chain fatty acids of fecal samples from healthy and celiac subjects. Front. Microbiol. 8:1722. DOI:10.3389/fmicb.2017.01722, Darden T., York D., Pedersen L. (1993): Particle mesh Ewald: An N•log(N) method for Ewald sums in large systems. J. Chem. Phys. 98:10089–10092.

David C. C. and Jacobs D. J. (2014): Principal Component Analysis: a method for determining the essential dynamics of proteins. Methods Mol Biol. 1084:193–226. DOI:10.1007/978-1-62703-658-0_11.

De Melo E. N., McDonald C., Saibil F., Marcon M. A., Mahmud F. H., (2015): FRCP, Celiac Disease and Type 1 Diabetes in Adults: Is This a High-Risk Group for Screening?. Can. J. Diabetes 39: 513e519.

Dean M. D. (2013): Genetic disruption of the copulatory plug in mice leads to severely reduced fertility. PLoS Genet9: e1003185

Deasey S. and Nurminskaya M. (2013): Tissue-Specific Responses to Loss of Transglutaminase 2. Amino Acids. 44(1): 179–187.

Dehal P.S., Joachimiak P. M., Price M. N., Bates J. T., Baumohl J. K., Chivian D., Friedland G. D., Huang K. H., Keller K., Novichkov P. S., Dubchak I. L., Alm E. J. and Arkin A. P. (2010): MicrobesOnline: an integrated portal for comparative and functional genomics. Nucleic Acids Res. 38:D396–D400

Demain A.L., Adrio J.L. (2008): Contributions of microorganisms to industrial biology. Mol Biotechnol 38(1):41–55.

Di Sabatino A. andCorazza G. R. (2009): Coeliac disease. Lancet. 373:1480–1493.

Dorgalaleh A., Tabibian S., Shams M., Tavasoli B., Gheidishahran M., Shamsizadeh M. (2016): Laboratory Diagnosis of Factor XIII Deficiency in Developing Countries: An Iranian Experience. Laboratory Medicine 47(3):220–226.

Drozdetskiy A., Cole C., Procter J., Barton G. J. (2015): JPred4: a protein secondary structure prediction server. Nucleic Acids Research. 43(W1):W389–W394. DOI:10.1093/nar/gkv332

Dubbink H.J., Hoedemaeker R.F., van der Kwast T. H., Schroder F.H., Romijn J.C. (1999): Human prostate-specific transglutaminase: a new prostatic marker with a unique distribution pattern. Lab Invest79:141-50.

Dube M., Schäfer C., Neidhart S., Carle R. (2007): Texturisation and modification of vegetable proteins for food applications using microbial transglutaminase. Eur Food Res Technol 225:287–299. DOI 10.1007/s00217-006-0401-2

Eckert R. L., Kaartinen M. T., Nurminskaya M., Belkin A. M., Colak G., Johnson G. V. W., and Mehta K. (2014): Transglutaminase regulation of cell function. Physiol Rev 94: 383–417, DOI:10.1152/physrev.00019.2013

Edgar R. C. (2004): MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32(5): 1792–1797.

Elli L., Roncoroni L., Hils M., Pasternack R., Barisani D., Terrani C., Vaira V., Ferrero S., Bardella M.T. (2012): Immunological effects of transglutaminase-treated gluten in coeliac disease. Hum. Immunol. 73:992–7.

Fernandes C. G., Plácido D., Lousa D., Brito J. A., Isidro A., Soares C. M., Pohl J., Carrondo M. A., Archer M., Henriques A. O. (2015): Structural and functional characterization of an ancient bacterial transglutaminase sheds light on the minimal requirements for protein cross-linking. biochemistry, 54:5723-5734. DOI: 10.1021/acs.biochem.5b00661

Fesus L., Piacentini M. (2002): Transglutaminase 2: an enigmatic enzyme with diverse functions. Trends Biochem Sci. 27: 534-539

Finn R. D., Bateman A., Clements J., Coggill P., Eberhardt R. Y., Eddy S. R., Heger A., Hetherington K., Holm L., Mistry J., Sonnhammer E. L. L, Tate J. and M. Punta (2014): Pfam: the protein families database. Nucleic Acids Res. 42(Database issue): D222–D230.

Folk J.E., Chung S.I. (1973): Molecular and catalytic properties of transglutaminases. Adv. Enzymol. Relat. Areas Mol. Biol. 38:109–191

Gallagher E., Gormley T.R., Arendt E.K. (2004): Recent advances in the formulation of gluten-free cereal based products, Trends Food Sci. Tech. 15:143–152.

Garnier J., Gibrat J.-F., Robson B. (1996): GOR Method for Predicting Protein Secondary Structure from Amino Acid Sequence Methods in Enzymology R.F. Doolittle Ed. 266:540-553

Gascuel O. (1997): BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 147:685–695.

Gasteiger E., Hoogland C., Gattiker A., Duvaud S., Wilkins M.R., Appel R.D., Bairoch A. (2005): Protein Identification and Analysis Tools on the ExPASy Server. (In) John M. Walker (ed): The Proteomics Protocols Handbook, Humana Press pp. 571-607

Gianfrani C., Siciliano R. A., Facchiano A., Camarca A., Mazzeo M. F., Costantini S., Salvati V. M., Maurano F., Mazzarella G., Iaquinto G., Bergamo P. and Rossi M. (2007): Transamidation of wheat flour inhibits the response to gliadin of intestinal t cells in celiac disease. Gastroenterology.133:780–789

Giordano D. and Facchiano A. (2018): Classification of Microbial Transglutaminases by evaluation of evolution trees, sequence motifs, secondary structure topology and conservation of potential catalytic residues. B.B.R.C. 509(2):506-513. DOI:10.1016/j.bbrc.2018.12.121

Grenard P., Bates M. K., Aeschlimann D. (2001): Evolution of transglutaminase genes: identification of a transglutaminase gene cluster on human chromosome 15q15. Structure of the gene encoding transglutaminase X and a novel gene family member, transglutaminase Z. J Biol Chem.276(35):33066-78.

Griffin M., Casadio R., and Bergamini C. M. (2002): Transglutaminases: nature’s biological glues. Biochem. J. 368, 377−96.

Guandalini S., Assiri A. (2014): Celiac Disease A Review, JAMA Pediatr. 168(3):272–278. DOI:10.1001/jamapediatrics.2013.3858

Guindon S., Dufayard J.F., Lefort V., Anisimova M., Hordijk W., Gascuel O. (2010): New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology. 59(3):307-21.

Gundemir S., Colak G., Tucholski J., Johnson G. V.W. (2012): Transglutaminase 2: A molecular Swiss army knife. Biochimica et Biophysica Acta. 1823:406–419

Ha C. R., and Iuchi I. (1998): Enzyme responsible for egg envelope (chorion) hardening in fish: purification and partial characterization of two transglutaminases associated with their substrate, unfertilized egg chorion, of the rainbow trout, *Oncorhynchus mykiss*. J. Biochem.124, 917-926

Hadjivassiliou M., Sanders D.S., Woodroofe N. et al. (2008): Gluten ataxia Cerebellum. 7: 494. DOI:10.1007/s12311-008-0052-x

Hallert C., Grant C., Grehn S., Grännö C., Hultén S., Midhagen G., Ström M., Svensson H. & Valdimarsson T. (2002): Evidence of poor vitamin status in coeliac patients on a gluten-free diet for 10 years, Aliment Pharmacol Ther. 16:1333–1339.

Hausch F., Shan L., Santiago N.A., Gray G.M., Khosla C. (2002): Intestinal digestive resistance of immunodominant gliadin peptides, Am J Physiol. Gastrointest. Liver Physiol. 283(4):G996-G1003

Heffler E., Nebiolo F., Asero R., Guida G., Badiu I., Pizzimenti S., Marchese C., Amato S., Mistrello G., Canaletti F. and Rolla, G. (2011): Clinical manifestations, co-sensitizations, and immunoblotting profiles of buckwheat-allergic patients. Allergy 66(2):264–70.

Hess B., Bekker H., Berendsen H. J. C., Fraaije J. G. E. M. (1997): LINCS: A linear constraint solver for molecular simulations. J. Comp. Chem. 18:1463–1472.

Hockney R. W., Goel S. P., Eastwood J. (1974): Quiet high resolution computer models of a plasma. J. Comp. Phys. 14:148–158.

Humphrey W., Dalke A., Schulten K. (1996): VMD: Visual molecular dynamics. Journal of Molecular Graphics. 14(1):33-38

Isik K., Chun J., Hah Y. C. & Goodfellow M. (1999): *Nocardia uniformis nom. rev*. Int J Syst Bacteriol. 49:1227–1230.

Jiang W.G., Ablin R., Douglas-Jones A., Mansel R.E.(2003): Expression of transglutaminases in human breast cancer and their possible clinical significance. Oncol Rep 10(6):2039-44.

Kashiwagi T., Yokoyama K., Ishikawa K., Ono K., Ejima D., Matsui H. and Suzuki E. (2002): Crystal structure of microbial transglutaminase from Streptoverticillium mobaraense. J. Biol. Chem. 277:44252−60.

Kato J.Y., Suzuki A., Yamazaki H., Ohnishi Y. and Horinouchi, S. (2002): Control by A-factor of a metalloendopeptidase gene involved in aerial mycelium formation in *Streptomyces griseus*. Journal of Bacteriology. 184:6016-6025.

Kelley A. L., Mezulis S., Yates C. M., Wass M. N. and Sternberg M. J. E. (2015): The Phyre2 web portal for protein modelling, prediction and analysis. Nat Protoc. 10(6): 845–858. DOI:10.1038/nprot.2015.053

Kim H.S., Jung S.H., Lee I.S., Yu T.S. (2000): Production and Characterization of a Novel Microbial Transglutaminase from Actinomadura sp. T-2. Journal of Microbiology and Biotechnology. 10:187-194.

Kinoshita N., Homma Y., Igarashi M., Ikeno S., Hori M. & Hamada M. (2001): *Nocardia vinacea sp. nov.* Actinomycetologica. 15:1–5.

Kishino H., Hasegawa M. (1989): Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. 29:170-179

Kobayashi K., Kumazawa Y., Miwa K., Yamanaka S. (1996): Ɛ-(γ-Glutamyl)lysine cross-links of spore coat proteins and transglutaminase activity in *Bacillus subtilis.* FEMS Microbiology Letters 144:157-160

Korsgren C. and Cohen C. M. (1991): Organization of the gene for human erythrocyte membrane protein 4.2: Structural similarities with the gene for the a subunit of factor XIII. Proc. Natl. Acad. Sci. USA.88:4840-4844

Kuraishi C., Sakamoto J. and Soeda, T. (1996): The usefulness of transglutaminase for food processing' in biotechnology for improved foods and flavors. American Chemical Society. Symposium Series 637:29-38

Kuraishi C., Yamazaki K., and Susa Y. (2001): Transglutaminase: its utilization in the food industry. Food Rev Int. 17: 221-246

Kurppa K., Collin P., Viljamaa M., Haimila K., Saavalainen P., Partanen J., Laurila K., Huhtala H., Paasikivi K., Mäki M., Kaukinen K. (2009): Diagnosing mild enteropathy celiac disease: a randomized, controlled clinical study, Gastroenterology. 136(3):816-23.

Langini C., Caflisch A. and Vitalis A. (2017): The ATAD2 Bromodomain binds different acetylation marks on the histone H4 in similar fuzzy complexes. JBC Papers in Press. Published as Manuscript M117.78635. DOI:10.1074/jbc.M117.786350

Laskowski R. A., MacArthur M. W., Moss D. S. and Thornton J. M. (1993): PROCHECK: a program to check the stereochemical quality of protein structures. J. Appl. Cryst., 26, 283-291.

Lebwohl B., Cao Y., Zong G., Hu F.B., Green P. H. R., Neugut A. I., Rimm E. B., Sampson L., Dougherty L. W., Giovannucci E., Willett W. C., Sun Q., Chan A.T. (2017): Long term gluten consumption in adults without celiac disease and risk of coronary heart disease: prospective cohort study, BMJ 357:j1892. DOI:10.1136/bmj.j1892

Lebwohl B., Ludvigsson J. F., Peter H. R. (2015): Green Celiac disease and non-celiac gluten sensitivity. BMJ 351: h4347. DOI: 10.1136/bmj.h4347

Lee J.-H., Jang, S.-I., Yang J.-M., Markova N. G. and Steinert P.M. (1996): The Proximal Promoter of the Human Transglutaminase 3 Gene. The Journal of Biological Chemistry 271, No. 8, Issue of February 23, pp. 4561–4568.

Lefort V., Longueville J.-E., Gascuel O. (2017): SMS: Smart Model Selection in PhyML. Molecular Biology and Evolution. 34(9):2422–2424. DOI:10.1093/molbev/msx149

Li C., Wang C.-L., Sun Y., Li A.-L., Liu F. and Meng X.-C. (2016): Microencapsulation of *Lactobacillus rhamnosus* GG by transglutaminase cross-linked soy protein isolate to improve survival in simulated gastrointestinal conditions and yoghurt. Journal of Food Science.81(7): M1726-34 DOI: 10.1111/1750-3841.13337

Liu S., Cerione R. A. and Clardy J. (2002): Structural basis for the guanine nucleotide-binding activity of tissue transglutaminase and its regulation of transamidation activity. Proc. Natl. Acad. Sci. USA 99(5):2743-2747

Lombardi E., Bergamo P., Maurano F., Bozzella G., Luongo D., Mazzarella G., Rotondi Aufiero V., Iaquinto G. and Rossi M. (2013): Selective inhibition of the gliadin-specific, cell-mediated immune response by transamidation with microbial transglutaminase. Journal of Leukocyte Biology. 93:479-488.

Lorand L., Graham R. M. (2003) Transglutaminases: crosslinking enzymes with pleiotropic functions. Nature Reviews Molecular Cell Biology volume 4, pages 140–156. DOI:10.1038/nrm1014

Malalasekera V., Cameron F., Grixti E., Thomas M.C. (2009): Potential reno-protective effects of a gluten-free diet in type 1 diabetes. Diabetologia 52:798e800.

Martínez B., Miranda J. M., Franco C. M., Cepeda A. and Vázquez, M. (2011): Evaluation of transglutaminase and caseinate for a novel formulation of beef patties enriched in healthier lipid and dietary fiber. LWT-Food Sci Technol. 44:949-956.

Martins I. M., Matos M., Costa R., Silva F., Pascoal A., Estevinho L. M., Choupina A. B. (2014): Transglutaminases: recent achievements and new sources. Appl Microbiol Biotechnol 98:6957–6964

Matthias T., Jeremias P., Neidhöfer S., Lerner A. (2016): The industrial food additive, microbial transglutaminase, mimics tissue transglutaminase and is immunogenic in celiac disease patients. Autoimmunity Reviews. 15:1111–1119

Mazzeo F. M., Bonavita R, Maurano F., Bergamo P., Siciliano R.A., Rossi M. (2013): Biochemical modifications of gliadins induced by microbial transglutaminase on wheat flour. Biochimica et Biophysica Acta. 1830:5166–5174

Milani A., Vecchietti D.,Rusmini R., (2012): TgpA, a protein with a eukaryotic-like transglutaminase domain, plays a critical role in the viability of Pseudomonas aeruginosa, PLoS One. 7:e50323. DOI:10.1371/journal.pone.0050323.

Moreno M.L., Comino I., Sousa C. (2014): Alternative Grains as Potential Raw Material for Gluten-Free Food Development in The Diet of Celiac and Gluten-Sensitive Patients., Austin. J. Nutri. Food Sci. 2(3): 1016. ISSN: 2381-8980.

Moscaritolo S., Treppiccione L., Ottombrino A. and Rossi M. (2016): Effects of two-step transamidation of wheat semolina on the technological properties of gluten. Foods. 5(49):8 pages. DOI:10.3390/foods5030049.

Motoki M. and Kumazawa Y. (2000): Recent research trends in transglutaminase technology for food processing. Food Sci. Technol. Res. 6(3):151-160.

Motoki M. and Seguro K. (1998): Transglutaminase and its use for food processing. Trends in Food Science & Technology. 9:204-210

NCBI Resource Coordinators (2016): Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 44(Database issue): D7–D19.

Notredame C., Higgins D. G., Heringa J. (2000): T-Coffee: A novel method for multiple sequence alignments. JMB.302:205-217

Nygård O., Vollset S.E., Refsum H., Brattström L., Ueland P.M. (1999): Total homocysteine and cardiovascular disease. J. Intern. Med. 246: 425–4l.

Páll S., Abraham M. J., Kutzner C., Hess B., Lindahl E. (2015): Tackling Exascale Software Challenges in Molecular Dynamics Simulations with GROMACS. Proc. of EASC 2015 LNCS, 8759 3-27

Parrinello M. and Rahman A. (1981): Polymorphic transitions in single crystals: a new molecular dynamics method. J. Appl. Phys. 52:7182–7190.

Pasha I., Saeed F., Tauseef Sultan M., Batool R., Aziz M. & Ahmed W. (2016): Wheat Allergy and Intolerence; Recent Updates and Perspectives, Critical Reviews in Food Science and Nutrition. 56(1):13-24. doi: 10.1080/10408398.2012.659818

Pasternack R., Dorsch S., Otterbach J.T., Robenek I.R., Wolf, S. and Fuchsbauer H.-L. (1998): Bacterial pro-transglutaminase from *Streptoverticillium mobaraense* purification, characterisation and sequence of the zymogen. European Journal of Biochemistry**. 257**:570-576.

Peng X., Zhang Y., Zhang H., Graner S., Williams J. F., Levitt M. L., Lokshin A. (1999): Interaction of tissue transglutaminase with nuclear transport protein importin-α3. FEBS Letters. 446:35-39

Pham-Short A., Donaghue K.C., Ambler G., Garnett S., and Craig M. E. (2016): Quality of Life inType 1 Diabetes and Celiac Disease: Role of the Gluten-Free Diet, J. Pediatr. 179:131-8. DOI:10.1016/j.jpeds.2016.08.105

Pinkas D. M., Strop P., Brunger A. T., Khosla C. (2007): Transglutaminase 2 undergoes a large conformational change upon activation. PLoS Biol. 5(12):e327. DOI: 10.1371/journal.pbio.0050327.

Rashtak S. and Murray J. A. (2012): Review Article: Celiac Disease, New Approaches to Therapy. Aliment. Pharmacol. Ther. 35(7):768–781. DOI:10.1111/j.1365-2036.2012.05013.x.NIH

Rebets Y., Tokovenko B., Lushchyk I., Rückert C., Zaburannyi N., Bechthold A., Kalinowski J. and Luzhetskyy A. (2014): Complete genome sequence of producer of the glycopeptide antibiotic Aculeximycin *Kutzneria albida* DSM 43870T, a representative of minor genus of *Pseudonocardiaceae*. BMC Genomics. 15:885.

Ribeiro M., Nunes-Miranda J. D., Branlard G., Carrillo J. M., Rodriguez-Quijano M. and Igrejas G. (2013): One hundred years of grain omics: identifying the glutens that feed the world, J. Proteome Res. 12:4702−4716 dx.doi.org/10.1021/pr400663t

Rossi Marquez G., Di Pierro P., Mariniello L., Esposito M., Giosafatto C. V. L., Porta R. (2017): Fresh-cut fruit and vegetable coatings by transglutaminase crosslinked whey protein/pectin edible films. LWT - Food Science and Technology. 75:124-130

Rotsch A. (1954): Chemische und backtechnische Untersuchungen an künstlichen Teigen Brot Gebaeck. 8:129–130.

Rubio-Tapia A., Ludvigsson J. F., Brantner T. L., Murray J. A. and. Everhart J. E. (2012): The Prevalence of Celiac Disease in the United States. Am. J. Gastroenterol. 107:1538–1544. DOI: 10.1038/ajg.2012.219

Santhi D., Kalaikannan A., Malairaj P. & Arun Prabhu S. (2017): Application of Microbial Transglutaminase in Meat Foods: A Review. Critical Reviews in Food Science and Nutrition. 57(10):2071-2076 DOI: 10.1080/10408398.2014.945990

Schaal K. P., Lee H. J. (1992): Actinomycete infections in humans – a review. Gene. 115:201–211.

Schmidtke P., Bidon-Chanal A., Luque F. J. and Barril X. (2011): MDpocket: open-source cavity detection and characterization on molecular dynamics trajectories. Bioinformatics. 27(23):3276–3285.

Seravalli Guastaferro E. Ap., Iguti A. M., Santana I. Ap. and Filho F. F. (2011): Effects of application of transglutaminase in wheat proteins during the production of Bread. Procedia Food Science 1:935 – 942

Shimodaira H., Hasegawa M., (1999): Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. 16:1114-1116.

Sievers F., Wilm A., Dineen D., Gibson T. J., Karplus K., Li W., Lopez R., McWilliam H., Remmert M., Söding J., Thompson J. D. and Higgins D. G. (2011): Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011; 7: 539.

Smyth D.J., Plagnol V., Walker N.M., et al. (2008): Shared and distinct genetic variants in type 1 diabetes and celiac disease, N. Engl. J. Med. 359:2767e77.

Soares de Barros L. H., Assmann F. and Ayub1 Záchia M. A. (2003): Purification and properties of a transglutaminase produced by a *Bacillus circulans* strain isolated from the Amazon environment. Biotechnol. Appl. Biochem. 37:295–299

Sollid L. M. (2000): Molecular basis of celiac disease. Annu. Rev. Immunol. 18:53–81

Steffen W., Ko F. C., Patel J., Lyamichev V., Albert T. J., Benz J., Rudolph M. G., Bergmann F., Streidl T., Kratzsch P., Boenitz-Dulat M., Oelschlaegel T., Schraeml M. (2017): Discovery of a microbial transglutaminase enabling highly site-specific labeling of proteins, J. Biol. Chem. 292:15622-15635. DOI: 10.1074/jbc.M117.797811.

Stivala A., Wybrow M., Wirth A., Whisstock J. C., Stuckey P. J. (2011): Automatic generation of protein structure cartoons with Pro-origami. Bioinformatics. 27(23):3315–3316. DOI:10.1093/bioinformatics/btr575

Strop P. (2014): Versatility of microbial transglutaminase, Bioconjug Chem. 25:855-862. DOI: 10.1021/bc500099v.

Sud S., Marcon M., Assor E., et al. (2010): Celiac disease and pediatric type 1 diabetes: Diagnostic and treatment dilemmas, Int. J. Pediatr. Endocrinol. 1e15.161285.

Suzuki S., Izawa Y., Kobayashi K., Eto Y., Yamanaka S., Kubota K. & Yokozeki K. (2000): Purification and characterization of novel transglutaminase from bacillus subtilis spores, Bioscience Biotechnology and Biochemistry, 64(11):2344-2351. DOI: 10.1271/bbb.64.2344

Tamura K., Stecher G., Peterson D., Filipski A. and Kuma S. (2013): MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. Mol. Biol. Evol. 30(12):2725–2729. DOI:10.1093/molbev/mst197

Terazawa S., Mori S., Nakajima H., Yasuda M., Imokawa G. (2015): The UVB-Stimulated Expression of Transglutaminase 1 Is Mediated Predominantly via the NFκB Signaling Pathway: New Evidence of Its Significant Attenuation through the Specific Interruption of the p38/MSK1/NFκBp65 Ser276 Axis. PLoS ONE10(8):e0136311, DOI:10.1371/journal.pone.0136311

Tesfaw A. and Assefa F. (2014): Applications of transglutaminase in textile, wool, and leather processing. International Journal of Textile Science.3(4): 64-69. DOI: 10.5923/j.textile.20140304.02

Thacher S. M., Rice R. H. (1985): Keratinocyte-specific transglutaminase of cultured human epidermal cells: relation to cross-linked envelope formation and terminal differentiation. Cell, 40: 685–695. pmid:2578891

The UniProt Consortium (2017): UniProt: the universal protein knowledgebase. Nucleic Acids Res. 45: D158-D169

Thomas H., Beck K., Adamczyk M., Aeschlimann P., Langley M., Oita R. C., Thiebach L., Hils M., Aeschlimann D. (2013): Transglutaminase 6: a protein associated with central nervous system development and motor function. Amino Acids 44,161–177

Thompson T. (1999): Thiamin, riboflavin, and niacin contents of the gluten-free diet: is there cause for concern?. J. Am. Diet. Assoc. 99:858-862.

Thompson T. (2000): Folate, iron, and dietary fiber contents of the gluten-free diet. J. Am. Diet. Assoc. 100:1389-1396.

Tokay F. G. and Yerlikaya P. (2017): Shelf-Life Extension of Fish Fillets by Spraying with Microbial Transglutaminase. Journal of Aquatic Food Product Technology. 26(8):940-948.

Vallenet D., Calteau A., Cruveiller S., Gachet M., Lajus A., Josso A., Mercier J., Renaux A., Rollin J., Rouy Z., et al. (2017): MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes. Nucleic Acids Res. 45(D1):D517-D528

Vita R., Mahajan S., Overton J.A., Dhanda S.K., Martini S., Cantrell J.R., Wheeler D.K., Sette A., Peters B. (2018): The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. DOI: 10.1093/nar/gky1006. PubMed PMID: 30357391.

Vitalis A., Pappu R. V. (2009): Methods for Monte Carlo Simulations of Biomacromolecules. Annual Reports in Computational Chemistry. 5:49-76.

Wallace R. J., JrBrown B. A., Tsukamura M., Brown J. M. & Onyi G. O. (1991): Clinical and laboratory features of Nocardia nova. J Clin Microbiol. 29:2407–2411.

Watanabe M., Suzuki T., Ikezawa Z. and Arai S. (1994): Controlled enzymatic treatment of wheat proteins for production of hypoallergenic flour. Biosci. Biotech. Biochem. 58(2):388-390.

Webb B. and Sali A. (2016): Comparative Protein Structure Modeling Using Modeller. Current Protocols in Bioinformatics. 54:5.6.1-5.6.37. DOI:10.1002/cpbi.3.

Weiss J., Gibis M., Schuh V., Salminen H. (2010): Advances in ingredient and processing systems for meat and meat products. Meat Science 86:196-213
Wiederstein M.and Sippl M. J. (2007): ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res. 35: W407–W410.

Wieser H. (2007): Chemistry of gluten proteins, Food Microbiol. 24(2):115–119.

Yamaguchi S., Jeenes D.J., Archer D.B. (2001): Protein-glutaminase from Chryseobacterium proteolyticum, an enzyme that deamidates glutaminyl residues in proteins, Eur. J. Biochem. 268:1410–1421. DOI:10.1046/j.1432-1327.2001.02019.x

Yang J., Yan R., Roy A., Xu D., Poisson J., Zhang Y. (2015) : The I-TASSER Suite: Protein structure and function prediction. Nature Methods. 12:7-8

Yu J., Zhou Y., Tanaka I. and Yao M. (2010): Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. Bioinformatics. 26(1):46–52. DOI:10.1093/bioinformatics/btp599

Yuan F., Ahmed I., Lv L., Li Z., Li Z., Lin H., Lin H., Zhao J., Tian S. and Ma J. (2018): Impacts of glycation and transglutaminase catalyzed glycosylation with glucosamine on the conformational structure and allergenicity of bovine β-lactoglobulin. Food Funct.9:3944-3955. DOI: 10.1039/c8fo00909k

Yuan F., Lv L., Li Z., Mi N., Chen H., Lin H. (2017): Effect of transglutaminase-catalyzed glycosylation on the allergenicity and conformational structure of shrimp (*Metapenaeus ensis*) tropomyosin. Food Chemistry. 219:215–222.

Zhang D., Wang M., Wu J., Cui L., Du G. and Chen J. (2008): Two different proteases from *Streptomyces hygroscopicus* are involved in transglutaminase activation. Journal of Agricultural and Food Chemistry. 56:10261-10264

Zhang D., Zhu Y. & Chen J. (2009): Microbial Transglutaminase Production: Understanding the Mechanism, Biotechnology and Genetic Engineering Reviews. 26(1):205-222. DOI: 10.5661/bger-26-205

Zhu Y., Rinzema A., Tramper J. and Bol J. (1995): Microbial transglutaminase – a review of its production and application in food processing. Appl Microbiol Biotechnol.44:277-282

Zilhão R., Isticato R., Martins L. O., Steil L., Völker U., Ricca E., Moran C. P., Jr., and Henriques A. O. (2005): Assembly and Function of a Spore Coat-Associated Transglutaminase of Bacillus subtilis. Journal of Bacteriology, Nov. 2005, p. 7753–7764.

Zong G., Lebwohl B., Hu F., Sampson L., Dougherty L., Willett W., Chan A., Sun Q. (2017): Associations of Gluten Intake With Type 2 Diabetes Risk and Weight Gain in Three Large Prospective Cohort Studies of US Men and Women, American Heart Association Meeting Report Presentation. 135:A11

Zotzel J., Keller P. and Fuchsbauer H.-L. (2003): Transglutaminase from *Streptomyces mobaraensis* is activated by an endogenous metalloprotease. European Journal of Biochemistry. 270:3214-3222.

Appendix

List of scientific publications:

Article:

Giordano D. and Facchiano A. (2018): Classification of Microbial Transglutaminases by evaluation of evolution trees, sequence motifs, secondary structure topology and conservation of potential catalytic residues. B.B.R.C. 509(2):506-513. DOI:10.1016/j.bbrc.2018.12.121

Oral presentation:

Giordano D. and Facchiano A. (2018): Microbial transglutaminases: a deep analysis of PFAM sequences. BBCC2018 Meeting – International Conference on Bioinformatics and Computational Biology. Naples 19-21 November 2018.

Poster presentation:

- Giordano D., Facchiano A. (2018) Microbial Transglutaminases' structure and their evolution. NETTAB Workshop 2018. Genoa, 22-24 October, 2018.
- Giordano D., Facchiano A. (2018) Microbial transglutaminases 3D structures and evolution. BITS2018 Meeting – Annual Meeting of the Bioinformatics Italian Society. Turin 27-29 June 2018.
- Giordano D., Facchiano A. (2017) Homology modelling based study of structural properties of Microbial Transglutaminases. BBCC2017 Meeting – International Conference on Bioinformatics and Computational Biology. Naples 18-20 December 2017.
- Giordano D, Facchiano A. (2017) Evolutionary relationships of microbial transglutaminases. PeerJ Preprints 5:e3320v1. NETTAB Workshops 2017. Palermo, 16-18 October, 2017
- Giordano D., Facchiano A. (2017) Sequence analysis and evolutionary relationships of Microbial Transglutaminases. ISMB/ECCB 2017 Conference - Annual International Conference on Intelligent Systems for Molecular Biology (ISMB), Prague, 21-25 July, 2017.
- Giordano D., Facchiano A. (2017) A preliminary classification of Microbial Transglutaminases. BITS2017 Meeting – Annual Meeting of the Bioinformatics Italian Society. Cagliari 5-7 July, 2017.

- Giordano D., Facchiano A. (2016) Microbial Transglutaminases investigations for selecting putative forms of industrial interest. PeerJ Preprints 4:e2260v1. BITS2016 Meeting - Annual Meeting of the Bioinformatics Italian Society. University of Salerno, June 15-17.

I would like to thank:

Professor Caflisch, group leader of the Computational Structural Biology Research group at the University of Zurich, for giving me the opportunity to spend my compulsory PhD period abroad in his laboratory and to perform all the MD simulations required for this thesis. In particular I would like to thank Cassiano Langini and Davide Garolini for their excellent scientific support, Carmen Esposito for all her suggestions and help and Pawel Sledz my office mate and friend.

The spectrometry mass CeSMA-ProBio lab at the CNR of Avellino for their collaboration for the experimental activity assay performed on gliadin peptide

The Laboratory for Molecular Sensing at the CNR of Avellino, for giving me the possibility to purify SaNDyTGase; in particular Dr. **Angela Pennacchio**, a teacher, a collaborator but also a friend.

A special thank-you to my tutors:

Dr. S. D'Auria, for all his support and **Dr. A. Facchiano** for the help, the suggestions, the patience, the support, the encouragement.

A big thank-you to all my lab colleagues of these years and **to my PhD course colleagues**, in particular:

Dina, Imma, Serena and Eugenio for making this period spent together so special, for the laughs, the suggestions and the support.

Antonella and Teresa for having shared laughs, doubt and worries, for their support and their friendship.

Last but not least, I would like **to really thank my family** and **my husband** for every single moment spent together, for the support and for making me a better person.

And finally: "Thanks Franco for inspiring my whole life".