THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY (Ph.D.)

CHARACTERIZATION OF TRANSGLUTAMINASE 2 SUBSTRATE SPECIFICITY USING PHAGE DISPLAY TECHNOLOGY, LOGISTIC REGRESSION ANALYSIS AND INTRINSIC DISORDER EXAMINATION

ÉVA CSŐSZ

Supervisor: Prof. Dr. László Fésüs

DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY
MEDICAL AND HEALTH LIFE SCIENCE CENTER
UNIVERSITY OF DEBRECEN
DEBRECEN
2008

INTRODUCTION

Transglutaminases (E.C. 2.3.2.13.) are a family of structurally and functionally similar enzymes which catalyze Ca^{2+}-dependent posttranslational modification of proteins forming ε(γ-glutamyl)lysine crosslinks between glutamine and lysine residues in proteins and polypeptide chains. In humans, nine transglutaminase genes have been identified and eight of them code active enzymes. The blood coagulation factor XIIIa (FXIIIa), keratinocyte transglutaminase (TG1), tissue transglutaminase (TG2) and epidermal transglutaminase (TG3) are well characterized enzymes while there are less information about the prostate enzyme (TG4), TG5, TG6 and TG7. The erythrocyte band 4.2 protein is catalytically inactive having a structural role in the erythrocyte membrane skeleton.

TG2 is a ubiquitous member of the transglutaminase family found in many tissues and cell types. Inside the cell, it can be present in the nucleus, cytosol, endoplasmic reticulum, mitochondria or associated to plasma membrane but upon externalization it can appear on the cell surface and extracellular matrix as well. The human TG2 protein is a 76 kDa, 688 amino acid containing protein encoded by the ~37 kb TG2 gene found on the chromosome 20q12.

Structure and reaction mechanism of transglutaminase 2

The transamidation reaction catalyzed by TG2 is an acyl transfer reaction in which the active site thiol group of Cys277 attacks the γ-carboxamide group of glutamine residue forming the acylenzyme intermediate. In the next step the ε-amino group of lysine residue attacks the thioester bond and the crosslinked product is formed. When the amine substrate is not available, water can attack the thioester bond resulting in deamidated end-product. The rate limiting step in the catalysis is the acylation step, the formation of the acylenzyme intermediate. Beside the transamidation/deamidation activity, the TG2 has kinase, protein disulphide isomerase and GTP/ATPase activity.

The human TG2 has four domains: the core domain holding the catalytic Cys277-His335-Asp358 triad, an N-terminal β-sandwich domain and two C-terminal β-barrels. The nucleotide binding pocket is formed from side chains of Phe174, Val479, Met483, Arg580, Leu582 and Tyr583 making possible the binding of one molecule of GTP/GDP.

During physiological conditions, the TG2 has two forms: a GTP/GDP bound transamidation inactive (closed) form and a Ca^{2+}- bound active (open) form. The transition from the closed to the open form is accompanied by a large conformational change, the C-terminal β-barrels are displaced by almost 120 Å leading to the appearance of a tunnel where the catalytic Cys277 is located. In the GTP bound closed form, the TG2 acts as a G protein participating in different signaling processes but when the intracellular Ca^{2+}

concentration elevates and the GTP is ablated, the Ca^{2+}-bound enzyme achieves its active conformation leading to the transamidation of the cellular proteins.

Potential biological functions of transglutaminase 2 mediated transglutamination

Transglutaminase 2 is a multifunctional enzyme having diverse physiological functions. In contrast to the other members of the transglutaminase family, which have quiet well-defined specific functions, TG2 has various roles. As a transamidating enzyme, it can modify cytoskeletal proteins (actin, myosin, ROCK2) with a role in cell motility and adhesion. TG2 influences inflammatory cytokine production by cross linking free IκBα leading to NFκB translocation to the nucleus, and by crosslinking and thus enhancing the activity of annexin I. Depending on the cell type, TG2 can exert pro-apoptotic or anti-apoptotic effects. As soon as apoptosis starts and the intracellular Ca^{2+} level rises, the activation of TG2 results in extensive protein cross-linking and formation of detergent insoluble protein scaffolds. It is not clear yet how TG2 influences the energy status of the cell, but it can covalently modify phosphoglycerate dehydrogenase, phosphorylase kinase, mitochondrial aconitase, α-ketoglutarate dehydrogenase and glyceraldehyde-3-phosphate dehydrogenase – these latter three enzymes have reduced activity upon transglutamination. The heat shock protein family members (hsp60, hsp70, hsp90, hsp27, crystallins) and ubiquitin may also act as substrates suggesting a role in defense against misfolded proteins. The nuclear translocation of the enzyme leads to modification of histones, SP1 transcription factor, androgen receptor and retinoblastoma protein, suggesting its transcriptional regulatory effect. TG2 is present on the cell surface promoting cell-matrix interactions by binding to fibronectin and integrins, and by modifying extracellular proteins. It is implicated in extracellular matrix remodeling, tissue repair and wound healing. The extracellular crosslinking activity of TG2 is related to activation/attenuation of signaling pathways as hormones (insulin, glucagon), local mediators (VIP, Substance P, histamine, serotonin) as well as hormone binding proteins (IGFBP-1 and 3, thyroglobulin) and other signaling molecules (ephrinA, midkine) can be modified by the enzyme.

Available information on substrate preference of transglutaminase 2

Physiological and pathological roles of transglutaminase 2 can be fully understood only if we know what the *in situ* TG2 substrates are and how the substrate specificity of the enzyme is determined. To examine the role of different amino acids around glutamine residues on substrate effectiveness the amino acid sequence around substrate glutamine residues was studied extensively. The effect of deletion or substitution of different amino acids from peptides derived from β-casein was examined and the importance of Val at -5, Leu at -2, Lys at +2, Val at +3, Leu at +4 and Pro at +5 positions in determining the

substrate requirement of TG2 was suggested. As another approach, the amino acid sequences surrounding the glutamine residues, which serve as an amine acceptor site in TG2 catalyzed cross-linking reaction, were compared in substrate proteins. There was observed a high proportion of charged and polar amino acids in the vicinity of the substrate glutamines, which suggested a preference for surface location of these substrate sites. Often, two directly adjacent glutamine residues functioned as amine acceptor sites and in the majority of the substrate proteins, the glutamine residue was located close to the N- or C-terminus of the proteins. The study of substance P analogues revealed further insights into the role of different amino acids in the surrounding of substrate glutamines. The proline at +1 position had negative effect but at -1 position, it favored the recognition of the substrate. The presence of asparagine or glycine at either side of the glutamine had favorable effect but the presence of positively charged residues two or four residues away from the glutamine towards the N-terminus seemed to be unimportant for determining the specificity. In another study, the residues, which were missing from the surrounding of the substrate glutamines, were examined. The study was done on crystal structures deposited in PDB; however, the results were presented on sequence level emphasizing the importance of charged residues as discouraging features in the surrounding of glutamine residues.

In the plant storage protein, gliadin, the TG2 recognizes the QxP sequence rather than the QP or QxxP sequences and the resulted deamidated peptide can serve as an anitgen in coeliac disease. The importance of proline in the recognition of substrate glutamine was confirmed by a combinatorial approach as well, administrating phage-displayed random peptide library. The resulted QxPφD(P), QxPφ, and QxxφDP sequences, where φ stands for hydrophobic amino acids, were preferred by TG2.

Considering the lysine substrate preference, fewer earlier studies have been carried out and it was established that the enzyme is less selective toward lysine donor substrates than to the glutamine donor ones and has broader tolerance to structural differences in the lysine donor substrate proteins. To study the amino acid residues influencing the amine donor substrate properties of lysines the sequences around substrate lysine from the modified alpha A-crystallin were studied. The glycine or aspartate before the amine donor lysine had the strongest adverse effects on substrate reactivity while proline, histidine, and tryptophan were found to be less favorable. Valine, arginine, and phenylalanine, and to a lesser extent serine, alanine, leucine, tyrosine, and asparagine had an enhancing effect. The size and charge of arginine exerted a positive effect while a tolerance toward proline was observed.

AIM OF THE STUDIES

1. To collect the so far found TG2 substrate proteins published in the scientific literature.
2. To construct a transglutaminase substrate database.
3. To adapt phage display technology to specifically select the preferred primary structure features around substrate glutamine residues.
4. To develop *in silico* methods for the comparison of sequence contexts around substrate and non substrate residues.
5. To find new methods for the comparison of the spatial environment of substrate and non substrate residues based on their three dimensional structures.
6. To determine the predictor amino acids defining the important features of the spatial shape of substrate proteins necessary for recognition by TG2.
7. To find important factors influencing the substrate recognition in substrate proteins lacking crystal structure.

MATERIALS AND METHODS

Phage biopanning

Peptides were selected from a commercially available 7-mer random library which displayed them on phage M13 with a diversity of $2.8x10^9$. The GST-TG2 was immobilized on glutathione-sepharose 4B beads and the slurry was washed three times, blocked with 0.5% BSA and washed using TBST. A 50 mL reaction mix consisting of 5 mM $CaCl_2$, 10 mM DTT, and 10 mL of the phage library ($2x10^{12}$ PFUs) in TBS was added to the beads and incubated for 1 h with continuous shaking. To remove unbound phages, the slurry was washed 10 times with TBST, and the phage population remaining attached was eluted with shaking in 100 mM glycine-HCl pH 2.0 and was neutralized with. The eluate was amplified in *E. coli* ER2537 cells, purified by precipitation with PEG/NaCl, titrated and used (10^{11} PFUs) in the next selection cycle. In the consecutive biopanning rounds, bound phages were eluted with 5 mM specific amine donor substrate, 5BPA. After the second and third rounds, individual phage clones were isolated, and ssDNA was prepared and sequenced.

Enzyme-linked immunosorbent assay (ELISA)

Phage clones from randomly picked plaques were amplified and purified. The concentration of phage stocks was estimated from the absorbance at 260 nm and dilution series (10^{-4}–10 nM) were prepared in 0.5% BSA containing 10 mM DTT and 5 mM $CaCl_2$. 96-well microtiter plates were coated with 10 mg/mL GST-TG2 in TBS containing 10 mM DTT and blocked with BSA. After three washes with TBST, 50-mL aliquots of the phage solutions were incubated in the plate for 1 h, followed by 10 washes with TBST. Bound phage particles were quantified adding a HRP-conjugated anti-M13 antibody. After washing, the plates were incubated with 1 mg/mL tetramethyl-benzydine for 10 min, the color was developed by adding H_2SO_4, and the absorbance was measured at 450 nm in a Wallac 1420 Victor2 microtiter plate reader.

In vitro TGase assay

To test for TG2 catalyzed transamidation the reaction mixture contained 5 mM $CaCl_2$, 5 mM DTT and 1 mM 5BPA as amine donor in Tris-HCl at pH 8.5. The glutamine donor substrate had different concentrations. The aliquots of amplified phage clones ($5x10^{10}$ PFUs) were dissolved in Tris-HCl containing 15% glycerol. The SGYGQQGQTPYYNQQSPHPQQQQP peptide was dissolved to a final concentration of 200 mM in Tris- HCl containing either 1 mM BKP or 1 mM 5BPA as amine donor substrate. In case of peptides predicted as substrates for TG2 using the identified spatial

amino acid pattern, the reaction mixture contained 0.5 mM glutamine donor substrate peptide (orexin B or neuropeptide Y). In case of exendin 4 (predicted amine donor substrate) the NQEQVSPLTLLK peptide was used as glutamine donor substrate. All reaction was initiated by the addition of recombinant TG2 and incubated for 1 h or 2 h at 37°C and stopped by the addition of 5 mM EDTA. Reaction products were analyzed either by Western blotting or by mass spectrometry.

Immunoblotting

Aliquots of in vitro TG2-labeling reactions containing $5x10^{10}$ phage particles were run on 10% SDS-PAGE. Protein bands were stained or transferred to an Immobilon-P PVDF membrane. Visualization of biotinylated protein species was carried out using HRP-conjugated streptavidin, followed by chemiluminescent staining and detection.

Mass Spectrometry analysis

Liquid chromatography/mass spectrometry (nanoLC/MS) analyses were performed using a QTRAP nanoLC-MS/MS 4000 ion trap mass spectrometer, equipped with a turbo electrospray ion source. The eluting system consisted of 2% formic acid, 2% acetonitrile, water (eluent A), and 0.1% formic acid, in 98% acetonitrile (eluent B). The aliquots of neuropeptide Y, orexin B and exendin 4 containing reaction mixture were injected onto a Zorbax 300SB-C18 column and fractionated by performing a linear gradient of eluent B in eluent A, at a flow rate of 0.5 µL/min. The 100-mL aliquots of SGYGQQGQTPYYNQQSPHPQQQQP peptide containing reaction mixture were injected onto an RP-HPLC C18 column Tagra and fractionated by performing a linear gradient of eluent B in eluent A from at a flow rate of 6 mL/min. The resulting mass data were elaborated using the Analyst software.

Database search

The heptapeptide sequences obtained from the selected phage clones after the third round of biopanning were examined. The deduced consensus pattern was used to further examine if it is also present in sequences around glutamine residues modified by transglutaminase in known substrate proteins. As a broader approach, peptide GQQQTPY was chosen from modified heptapeptides as a representative substrate sequence, and proteins that contain this sequence were searched from the PIR database using BLAST algorithm at the NCBI server. The group of SWI1/SNF1-related chromatin remodeling factors was chosen for further investigation.

Sequence and structure files

The UniProt sequence data were used for comparative sequence analysis and intrinsic disorder prediction. In the spatial environment studies the crystal structure data files originated from PDB. The surface accessibility of amino acid residues was estimated and an arbitrary threshold of 18 % was established. The amino acids with less than 18% surface accessibility were deleted to investigate only those amino acids which are located on the surface of the protein. The whole structures and the surface accessible structures were used separately during the examination.

Comparison of TG2 substrate sequences

The occurrence of each amino acid in a "window" of five amino acids, at either side of the glutamine residues, was studied with SEQSTAT program. All glutamine residues which were reported in the literature as substrate for TG2 were considered. The results for substrate and non substrate datasets were compared and the significant differences were considered.

Spatial environment analysis

The whole structure and the surface-accessible structure files were used as inputs for ATOMDIST, a computer program that counts the number of amino acid residues at given distances. Parallel evaluations referring to the glutamine donor substrates and the lysine donor substrates of TG2, respectively, were done. The reference point for examinations was the CD of glutamine residues and NZ of lysine residues. The number of amino acid residues present at each angstrom in a 15-Å-radius sphere around CD of glutamine and NZ of lysine residues was counted. Each of the 20 amino acids was identified by one single atom, usually the most distant carbon or heteroatom from the Cα to increase the resolution of the calculation. In this study, we defined the identified effective substrate sites as "substrates", while those residues that had not been used by TG2 at all, as "non-substrates".

Statistical analysis

Statistical analysis was performed on the results of ATOMDIST. The number of amino acid residues counted in a certain distance of the 15 Å radius was recoded using indicator method by emphasizing the total number of amino acid residues at given distances. Each of the recoded variable was entered into a cross tabulation for calculation of the odds ratios with 95% confidence intervals. Predictors with 95% confidence interval of the odds ratio differing from the value 1.0 were selected and entered into a multivariate logistic

regression analysis and those with significant odds ratio were used to construct the final model predicting the substrate and non substrate glutamine and lysine residues. For the internal validation of the prediction model the leave one out cross validation was used.

As an evaluation of the performance of the prediction, the sensitivity and specificity of the parameters used in model construction were measured. To compare prediction models with different independent variables receiver operating characteristics (ROC) curves were plotted and the areas under the curve (AUC) were calculated.

The chi-square test served for the comparison of the SEQSTAT results and in examination of the intrinsic disorder the Mann-Whitney U test was utilized.

Intrinsic disorder prediction

The sequence file of substrate proteins was used as input for intrinsic disorder content prediction. The IUPred (http://iupred.enzim.hu) and PONDR-VSL2 (www.pondr.com) predictors were used and those sequences which turned out to be unstructured with both predictors were accepted as disordered. Next, in each substrate protein where the disorder was present in substrate region and contained substrate and non substrate residues as well, 10 amino acids were considered on both sides of glutamine and lysine residues, respectively. The relative intrinsic disorder and the relative number of disorder promoting amino acids was determined in this “window” and the averages for substrate and non substrate residues in case of each substrate protein were compared.

RESULTS AND DISCUSSION

During the last 40 years, more than 130 proteins have been found susceptible for undergoing TG2 mediated posttranslational modifications in test tubes or cellular experiments. These proteins are localized in variuous cell compartments and often at the cell surface or extracellular matrix. This broad specificity of the enzyme for its targets may provide the flexibility needed to achieve the variety of functions, but also necessitates that the selection of a specific subset of proteins related to a particular biological event must be controlled by additional factors.

To study the substrates of the enzyme first we collected the substrates published in the literature into TRANSDAB Wiki, a publicly available transglutaminase substrate database. Using the extensive structural information deposited in the database we analyzed the substrate specificity of TG2 at sequence and tertiary structure level. Administrating *in silico* and *in vitro* methods we tried to unravel the promiscuous substrate specificity of TG2.

The interactive transglutaminase substrate database - TRANSDAB Wiki

Our aim was to generate a structural database of translgutaminase substrate proteins which provides information about the microenvironment of reactive and non-reactive glutamine and lysine residues. For this reason, we collected the transglutaminase substrate proteins and interaction partners reported in literature and we included them into the TRANSDAB Wiki (http://genomics.dote.hu/wiki) along with as much structural information as possible. The database was constructed on web 2.0 surface to provide the information in an easy to find, user friendly format and utilizes the advantages of wiki platform. Currently TRANSDAB Wiki contains 247 entries about interaction partners and substrate proteins for six transglutaminase types: activated blood coagulation factor XIII, keratinocyte transglutaminase, transglutaminase 2, epidermal transglutaminase, TG5 and microbial transglutaminase. Our studies concentrate mainly on transamidation activity of TG2 but the database, beside the substrates of the transamidation activity, contains the substrate proteins for the deamidation and phosphorylation reactions as well.

Linear sequence determinants of TG2 substrate specificity

To study the favorable primary structures around substrate glutamine residues we have used phage display technique to select glutamine donor substrates from a random heptapeptide containing phage library via binding to recombinant TG2. The heptapeptides exposed on the surface of phage particles specifically bound to the immobilized TG2 and were eluted using low pH in the first round and the specific amine donor substrate 5BPA in

the consecutive rounds. After the second round only 46% of the clones contained one or more glutamines, this percent increased to 75% after the third cycle and 26 glutamine-containing peptides were identified in total. To test whether the resulted glutamine-containing sequences are recognized by TG2 as glutamine-donor substrates, the amplified phage clones were used as possible glutamine donor substrates in an *in vitro* transglutaminase assay. The incorporation of amine-donor 5BPA into phage particles was monitored by Western blot. Two phage clones displaying the MPPPMRS and LMAKPTR peptides were used as negative controls. The peptides GQQQTPY, GLQQASV and WQTPMNS were modified most efficiently and the consensus motif pQx(P,T,S)l was established. This motif was consistent with sequences in identified substrates listed in the TRANSDAB Wiki and previous sequences and features reported in literature. Similar results were achieved by the administration of the phage display system by another group as well. The phage particles exposing dodecapeptides on their surface were introduced as glutamine donor substrates in a TG2 catalyzed reaction and the substrate sequences were specifically labeled by 5BPA. The QxPφD(P), QxPφ, and QxxφDP sequences, where φ stands for hydrophobic amino acids, were preferred by TG2.

Transamidation of GQQQTPY-like motifs within a native peptide

Using the GQQQTPY peptide as a representative example of an efficient TG2 substrate selected from the random phage library, a BLAST search in the PIR database was administrated to find human proteins that contain regions similar to GQQQTPY. The best hit was a group of SWI1/SNF1-related chromatin remodeling factors, which contain two repeats of GQQQTPY-like sequences within one of their two conserved glutamine-proline-rich domains was further analyzed. A 27-mer peptide corresponding to N-terminal part of p270 SWI1/SNF1-related chromatin remodeling factor with the sequence ^{1}SGYGQQGQTPYYNQQSPHPQQQQPPYS^{27} was synthesized and used in a transglutaminase assay followed by mass spectrometry analysis. The recombinant TG2 was able to crosslink the amine-donor peptides to Gln6, Gln8 and Gln22 of the GQQQTPY-like motif of 27-mer peptide suggesting that the heptapeptides identified by a combinatorial approach may have *in situ* relevance as TG2 substrates.

***In silico* study of favorable sequence contexts around TG2 substrate residues**

To examine the presence of favorable residues at given position around substrate residues the amino acid sequences surrounding the glutamine and lysine residues which serve as substrates sites in the transglutaminase catalyzed reaction were compared.

Using the SEQSTAT program the amino acid sequence context of 96 substrate glutamines was compared to the sequence context of 602 non substrate glutamines and the

significant differences were considered. The presence of glutamine residues at -1, -2, +1 positions adjacent to substrate glutamines occurred significantly more often than the presence of any other amino acid. A preference of substrate glutamines for the N-terminal end of the polypeptide chain was observed. The presence of Gly at -1 and Pro at +5, the polar Thr at -3, Gln at +2, Ser at +5, the positively charged Lys at +2 positions and the absence of Leu at -1 and Ser at +3 positions was significantly higher in the sequence context of substrate glutamines.

A similar approach was administrated for lysine residues as well. The amino acid sequence context of 63 substrate lysines was compared to the sequence context of 472 non substrate lysines and the significant differences were evaluated. The preference of substrate lysines for the C-terminal end of the polypeptide chain was observed. In accordance with previous data the presence of Pro at -2, and -3, and positively charged Arg at -4 and +3, Lys at -2 and +2 positions were significant in the context of substrate lysine residues. The presence of Lys at +1 position seems to be an important negative factor and our results did not confirm the importance of residues at -1 position, which was examined in detail by Grootjans et al.

The established pQx(P,T,S)l motif characteristic for TG2 substrate recognition along with the results obtained from the sequence comparisons provide new insights into the substrate recognition of TG2. None of these data could give a full explanation how TG2 glutamine and lysine sites are selected in substrate proteins.

Structural determinants of TG2 substrate specificity

Despite the extensive sequence studies with an attempt to identify a consensus sequence for TG2 modification none of the results could give a full explanation about how TG2 recognizes its substrates. To overcome this problem, our attention turned toward the structure of TG2 substrate proteins and we examined their secondary and tertiary structures to find the spatial shape characteristic for substrate glutamine and lysine residues. Using VMD the position of glutamine and lysine residues in secondary structure elements was examined and a slight preference of TG2 for glutamines situated in turns was observed, while less substrate glutamine residues were situated in beta sheet. The substrate lysine residues were more abundant in turns and in beta sheets than the non substrate ones and slightly more non substrate lysines occurred in coil and helix regions.

Logistic regression analysis based on tertiary structure features

The three dimensional structure of crystallized substrate proteins was examined next and the surface accessible and whole structures were distinguished for their predictive values. Both structure files were used as inputs for ATOMDIST and the output files were

analyzed by logistic regression. The significant differences between the spatial environments of substrate and non substrate residues were those amino acids which might have a role in substrate selection by TG2. Different amino acid residues at different distances from the CD of glutamine or NZ of lysine residues turned out to have either positive or negative effect on substrate selection favoring or reducing the substrate recognition.

Spatial features influencing glutamine substrate specificity

Using only the surface accessible amino acids in the calculations, from the numerous amino acids at different distances one residue of Thr at 5 Å, one Arg, one His and one Leu at 10 Å, one Val at 11 Å and one Arg and Phe at 15 Å from the CD of glutamine residues turned out to be significantly more abundant in the surrounding of substrate glutamines. These residues appeared to have a role in influencing glutamine substrate selection of TG2 and their presence exerted a positive effect on substrate preference of the enzyme. When the whole structures were used in such calculations, the presence of one Asp, one Gly and one Phe at 10 Å, one Ser at 11 Å, one Val at 12 Å and one Asn at 14 Å distance from the CD of glutamine residues acted as discouraging features preventing the glutamine to be used by the enzyme.

Spatial features influencing lysine substrate specificity

Examining only the surface accessible residues Gly, Ser and Asp were found to have a role in discrimination between substrate and non-substrate lysine residues. The presence of one Gly and one Asp at 6 Å, one Ser at 14 Å and one Gly at 15 Å from NZ of lysine residues exerted a positive effect.

When the whole structures were examined, more amino acids were found to have a role in the determination of substrate lysines. Beside one residue of Gly at 6 Å and 15 Å and one Ser at 14 Å, the presence of one His at 5 Å, one Ser at 6 Å, two Gly at 11 Å, one Pro at 12 Å and two Asp at 13 Å appeared to have a positive role in lysine site selection. The results obtained with two parallel calculations using either the surface accessible residues or the whole structures overlap in case of lysine residues but no overlap could be observed in case of glutamine residues.

A limitation of our method is that each amino acid residue is defined as one single atom, so the presence of Asn at 14 Å, for example, means that the C gamma of Asn is situated at 14 Å from C delta of glutamine and does not give information either about the orientation of the side chain or the spatial relation of this amino acid to the examined glutamine. Another limitation is the noisiness of the input data. The residues reported in

the literature as TG2 substrates were used as primary inputs in our examination but very few of them originate from highly accurate mass spectrometry analyses; the majority of the substrate residues were identified by different methods, sometimes with very different sensitivity. Even with these limitations the area under the curve in the ROC statistics in each case was higher than 0.785 indicating a good estimation power of predictor amino acids. The presence of Thr at 5 Å, Arg, His and Leu at 10 Å, Val at 11 Å, Arg and Phe at 15 Å and the absence of Asp, Gly and Phe at 10 Å, Ser at 11 Å, Val at 12 Å and Asn at 14 Å distance from CD of glutamine would favor that glutamine to be used as substrate by TG2. In case of lysine donor substrates, the presence of one His at 5 Å, one Gly, Ser and Asp at 6 Å, two Gly at 11 Å, one Pro at 12 Å, two Asp at 13 Å, one Ser at 14 Å and one Gly at 15 Å act as favorable features increasing the possibility of the lysine residue to be utilized as substrate by TG2.

Identification of novel TG2 substrates based on predictions by logistic regression analysis

It was known that TG2 is able to modify different peptide hormones and neuropeptides like insulin, glucagon, VIP, Substance P, ACTH and beta endorphin, so the crystal structure of several neuropeptides was examined. Using the presence or absence of the predictor amino acid residues as criteria in the examination of 17 neuropeptides showed several of them to be possible TG2 substrates. Among them neuropeptide Y, orexin B and exendin 4 were examined and found to be novel substrates for TG2 *in vitro*. The neuropeptide Y and orexin B can be found in CNS and are involved in stimulation of food intake and orexin B, acting on orexin receptors, takes part in modulation of wakefulness. The exendin 4 or exenatide originates from the saliva of the lizard *Gila monster* and is a GIP-1 incretin mimetic having a role in the regulation of blood glucose level. It is used in the medication of type II diabetes as Byetta (Amylin, Lilly). It needs further studies to investigate whether these peptides are *in vivo* substrates as well and if so, what kind of a role TG2 might have in the regulation of their actions. One possibility could be to control the available amount of active (monomeric) neuropeptides.

The role of intrinsic disorder in substrate recognition

It seems that two groups of TG2 substrate proteins could be analyzed: one group of proteins with a well defined crystal structure and the other group of proteins lacking crystal structure or the crystal structure are available but the parts containing the substrate residues are missing. The logistic regression analysis could be used well in the study of amino acids determining the TG2 substrate glutamine and lysine selection in case of proteins bearing crystal structure but in case of the second group a completely new approach was needed.

The occurrence of intrinsic disorder in proteins is a common phenomenon observed in the protein world and correlates mostly with regulatory functions. Intrinsic disorder has an important role in protein-protein interactions and protein binding partner recognition and is involved in posttranslational protein modifications including deacetylation and phosphorylation. It was demonstrated in some cases that the substrate glutamine and lysine residues tend to occur close to the N- or C- terminal end of substrate molecules. In many instances the regions containing the substrate residues are not resolved in crystal structure.

Based on these observations we considered the intrinsic disorder as a possible factor influencing the substrate recognition of TG2 as well. To test this hypothesis we searched the sequences of substrates for the presence of intrinsically disordered regions. The results suggest that the intrinsic disorder may have importance in substrate recognition in case of half of the studied proteins where either the whole protein was intrinsically unstructured (IUP) or the substrate region was situated in intrinsically disordered region (IDR). These data led us to use a more refined prediction. In case of each protein where the intrinsic disorder might have a role in substrate selection and contained substrate and non substrate residues, a 21 amino acid window was examined around glutamine and lysine residues. The relative disorder and the relative number of disorder promoting amino acids in these sequences were predicted. In case of substrate proteins both the relative disorder and the relative number of disorder promoting amino acids was significantly higher in the surrounding of substrate glutamine and lysine residues than in the surrounding of non substrate residues suggesting that the enzyme preferably uses those glutamine and lysine residues which are in intrinsically disordered regions.

The presented data suggest that the substrate recognition of TG2 requires a complex mechanism; beside the linear sequence features information present in the spatial structure and the presence of intrinsic disorder are also needed.

SUMMARY

Transglutaminase 2 (TG2) catalyzes the Ca^{2+}-dependent post-translational modification of proteins via formation of isopeptide bonds between their glutamine and lysine residues. The enzyme has more than 130 reported substrates but the exact mechanism by which its substrates are selected is still an enigma. As a first approach, we collected the known transglutaminase substrates into TRANSDAB Wiki (http://genomics.dote.hu/wiki), the transglutaminase substrate database and using the deposited information we attempted to find out the rules of TG2 substrate selection.

To study the preferred sequences around substrate glutamines we adapted the phage display technique selecting the glutamine donor substrates from a random heptapeptide library via their binding to recombinant TG2. The pQx(P,T,S)l consensus motif around glutamines was established, which is consistent with so far identified substrates. Database searches showed that several proteins contain peptides similar to the phage-selected sequences, and the N-terminal glutamine-rich domain of SWI1/SNF1-related chromatin remodeling protein p270 was chosen for detailed analysis. Mass spectrometry-based studies of a representative part of the SWI1/SNF1-related chromatin remodeling protein indicated that it was modified by TG2. Along with phage display technique *in silico* methods were used to compare the sequence context of substrate and non substrate residues to get a better understanding about principles of substrate selection of TG2. None of the results could give a full explanation how TG2 selects the different substrate glutamine and lysine residues.

Using the structural information on TG2 substrate proteins listed in TRANSDAB Wiki database a slight preference of TG2 for glutamine and lysine residues situated in turns could be observed. When the spatial environment of the favored glutamine and lysine residues were analyzed with logistic regression the presence of specific amino acid patterns were identified. Using the occurrence of the predictor amino acids as selection criteria several polypeptides were predicted and later identified as novel *in vitro* substrates for TG2. Studying the sequence of TG2 substrate proteins lacking available crystal structure the strong favorable influence on substrate selection of the presence of substrate glutamine and lysine residues in intrinsically disordered regions also could be revealed.

The collected sequence and structural data have provided novel understanding of how this versatile enzyme selects its substrates in various cell compartments and tissues and suggest that instead of the strict linear sequences spatial features must be considered as well to explain the complex physico-chemical interaction between TG2 and its substrates. It seems that in case of this enzyme a divergent substrate recognition system has evolved where beside the linear sequences, spatial structural features and the presence of intrinsic

disorder can be significant in substrate selection. This may reflect the unique nature of how transglutaminase 2 works in almost all cellular compartments, including the cell surface and extracellular space. It is capable to perform diverse biochemical reactions, such as signal transduction through its GTPase activity, ATP hydrolysis, protein disulphide isomerase activity, integrin and fibronectin binding, while its major biochemical function is modifying protein bound glutamine residues whenever it becomes feasible. The need of substrate selection for this classical transglutaminase function may arise under very different circumstances making the flexible recognition mechanisms detailed in this work advantageous.

This thesis is built upon the following publicaions:

Csősz, E., Keresztessy, Zs. and Fésüs. L. (2002). Transglutaminase substrates: from test tube experiments to living cells and tissues. Minerva Biotec. 14, 149-153.
IF: 0.217

Keresztessy, Zs.*, Csősz, E.*, Hársfalvi, J., Csomós, K., Gray, J., Lightowlers, R.N., Lakey, J.H., Balajthy, Z. and Fésüs, L. (2006). Phage display selection of efficient glutamine-donor substrate peptides for transglutaminase 2. Protein Sci. 15, 2466-2480. (* contributed equally)
IF: 3.46

Csősz, E., Meskó, B. and Fésüs, L. (2008). Transdab wiki: the interactive transglutaminase substrate database on web 2.0 surface. Amino Acids. Jul 2. [Epub ahead of print]
IF: 2.78

Csősz, E., Bagossi, P., Nagy, Z., Dosztányi, Zs., Simon, I. and Fésüs, L. (2008). Substrate preference of transglutaminase 2 revealed by logistic regression analysis and intrinsic disorder examination. J Mol Biol. Accepted for publication.
IF: 4.89

Other publications:

Nemes, Z., Csősz, É., Petrovski, G. and Fésüs, L. (2005). Structure-function relationship of transglutaminases – a contemporary view. Prog Exp Tumor Res. 38, 19-36.
IF: 4.214

Vecsei Z, Király R, Bagossi P, Tóth B, Csősz É, Sblattero D, Marzari R, Mäki M, Fésüs L, Korponay-Szabó IR. Coeliac autoantibodies recognize a composite main epitope on transglutaminase 2 involving amino acids from 3 domains. (manuscript).

Kiraly R, Csősz É, Kurtan T, Antus S, Szigeti K, Vecsei Z, Korponay-Szabo, IR, Keresztessy Z, Fesüs L. Functional significance of five non-canonical Ca^{2+}-binding sites of transglutaminase 2 characterised by site directed mutagenesis. (manuscript).

Posters:

First author of posters on the following meetings:

Csősz É., Hársfalvi J., Keresztessy Zs. and Fésüs L.: A humán szöveti transzglutamináz szubsztrátpreferenciájának vizsgálata random heptapeptid fágbemutató könyvtár szűrésével. 8th Conference of Hungarian Biochemical Society, Keszthely, 2002.

Csősz É. and Fésüs L.: A TRANSDAB – transzglutamináz szubsztrát adatbázis – ismertetése. 9th Conference of Hungarian Biochemical Society, Sopron, 2004.

Csősz É., Bagossi P. and Fésüs L An in silico study of substrate preference of transglutaminase 2. 30th FEBS Congress and 9th IUBMB Conference, Budapest, Hungary, 2005. Abstract in FEBS Journal, 272, Supplement 1, 410.

Csősz É., Bagossi P. and Fésüs L.: An *in silico* study of substrate preference for transglutaminase 2. 8th International Conference on Protein Crosslinking and Transglutaminases (PCL8), Lubeck, Germany, 2005.

Csősz É., Bagossi P. and Fésüs L.: An *in silico* study of substrate specificity of transglutaminase 2 – a possible role of unstructured conformations in substrate specificity. EMBO/SPINE2 Workshop, Intrinsically Unfolded Proteins: Biophysical Characterization & Biological significance, Budapest, Hungary, 2007.

Csősz É., Bagossi P., Dosztányi Zs., Simon I. and Fésüs L.: A humán szöveti transzglutamináz szubsztrát specificitásának tanulmányozása in silico módszerekkel – a rendezetlen régiók szerepe a szubsztrát felismerésében. Conference of the Hungarian Biochemical Society, Debrecen, 2007.

Csősz É., Bagossi P., Nagy Z., Dosztányi Zs., Simon I. and Fésüs L.: Structural features influencing the transglutaminase 2 substrate selection. 33rd FEBS Congress and 11th IUBMB Conference, Athens, Greece, 2008. Abstract in FEBS Journal; 275, Suppl 1, 215.

Co-author:

Király R, László É, Keresztessy Z, Fésüs L. A humán szöveti transzglutamináz Ca^{2+}-kötő helyeinek felderítése irányított mutagenezis alkalmazásával. Poster presentation on 6th Conference of Hungarian Biochemical Society, Sárospatak, 2001.

Kiraly R, Csősz É, Keresztessy Z, Fesüs L: Identification of Ca^{2+}-binding sites in the human transglutaminase 2 by surface potential engineering using site directed mutagenesis. Poster presentation on 7th International Conference on Transglutaminases and Protein Crosslinking Reactions. Ferrara, Italy, September 14-17, 2002. Abstract in *Minerva Biotec.* 14, 193.

Király R, Csősz É, Keresztessy Z, Fésüs L. A humán szöveti transzglutamináz Ca^{2+}-kötő helyeinek felderítése irányított mutagenezis alkalmazásával. 9th Conference of Hungarian Biochemical Society, Sopron, 2004.

Kiraly R, Csősz É, Keresztessy Z, Fesüs L: An attempt to identify the Ca^{2+}-binding sites of human transglutaminase 2 using site directed mutagenesis. Poster presentation on 8th International Conference on Protein Crosslinking and Transglutaminases, Lübeck, Germany, 2005.

Vecsei Z, Király R, Korponay-Szabó IR, Csősz É, Mäki M, Fésüs L: Calreticulin can mask the coeliac epitopes of transglutaminase 2. Poster presentation on 8th International Conference on Protein Crosslinking and Transglutaminases, Lübeck, Germany, 2005.

Kiraly R, Csősz É, Kurtan T, Keresztessy Z, Fesüs L: Kísérlet a humán szöveti transzglutamináz Ca^{2+}-kötő helyeinek felderítésére irányított mutagenezis alkalmazásával. Oral presentation on Conference of Hungarian Biochemical Society, Pécs, 2006.

Kiraly R, Csősz É, Kurtan T, Keresztessy Z, Fesüs L: Ca^{2+}-binding sites of transglutaminase 2 revealed by site directed mutagenesis. Poster presentation on 32th FEBS Congress, Vienna, Austria, 2007. Abstract in FEBS Journal; 274, Suppl 1, 167.

Kiraly R, Csősz É, Kurtan T, Keresztessy Z, Fesüs L: Ca^{2+}-binding sites of transglutaminase 2 revealed by site directed mutagenesis. Poster presentation on 9th International Conference on Protein Crosslinking and Transglutaminases, Marrakech, Morocco, 2007.

ACKNOWLEDGEMENTS

First of all I would like to thank my supervisor Prof. László Fésüs for the continuous support during my graduate studentship and for giving me the possibility to join his group.

Thanks to my former supervisor, Dr. Zsolt Keresztessy, with who I started to work in the laboratory, to the colleagues with whom I worked together during the years: members of the Fésüs laboratory: Krisztián Csomós, Róbert Király, Zsófia Vecsei, András Mádi, members of the Tőzsér laboratory: Péter Bagossi, Péter Boross, Tamás Sperka, to Bertalan Meskó, István Andreikovics, Ilma Korponay-Szabó. I am grateful to Julika Darainé, Attiláné Klem and Edit Komóczi for the excellent technical assistance. I would also like to thank for collegues György Fenyőfalvi, Goran Petrovski, Mária Punyicki and Kamilla Bereczki, who helped with thoughtful discussions and practical contribution.

And of last, but not least, I would like to thank to my family all the support and encouragement which made possible for me to achieve my dreams and that they believed in me even than when I was loosing my faith.

I would like to thank the Hungarian Scientific Research Fund (OTKA NI 67877) and the EU (MRTN-CT-2006-036032, MRTN-CT 2006-035624, LSHB-CT-2007-037730) for the possibility of spending time in an international scientific environment.