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

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

by

ÉVA CSŐSZ

Supervisor: Prof. László Fésüs, M.D., Ph.D., D.Sc., M.H.A.Sc.

DEPARTMENT OF BIOCHEMISTRY AND MOLECULAR BIOLOGY

MEDICAL AND HEALTH LIFE SCIENCE CENTER

UNIVERSITY OF DEBRECEN

DEBRECEN

2008

CONTENTS

1. Hungarian summary **6**

2. Introduction **8**

2.1. Characteristic features of transglutaminase 2 **9**

2.1.1. Structure and reaction mechanism of transglutaminase 2 **9**

2.1.2. The activation of transglutaminase 2 **11**

2.2. Biological functions of transglutaminase 2 **13**

2.2.1. Potential biological functions of transglutaminase 2 mediated transglutamination **13**

2.2.2. Available information on substrate preference of transglutaminase 2 **16**

3. Aim of the studies **19**

4. Materials and methods **20**

5. Results **28**

5.1. The interactive transglutaminase substrate database - TRANSDAB Wiki **28**

5.1.1. TRANSDAB Wiki database content **29**

5.2. Linear sequence determinants of TG2 substrate specificity **30**

5.2.1. Phage display selection of TG2-specific glutamine-donor peptides **31**

5.2.2. *In silico* study of favorable sequence contexts around TG2 substrate glutamine and lysine residues **37**

5.3. Structural determinants of TG2 substrate specificity **38**

5.3.1. Logistic regression analysis based on tertiary structure features **39**

5.3.2. Identification of novel TG2 substrates based on predictions by logistic regression analysis **45**

5.4. The role of intrinsic disorder in substrate recognition **46**

6. Discussion **50**

6.1. Linear sequence determinants of TG2 substrate specificity **50**

6.2. Structural determinants of TG2 substrate specificity **53**

7. Summary **56**

8. Acknowledgements **58**

9. References **59**

Supplement 1.

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

Supplement 2.

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)

Supplement 3.

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]

Supplement 4.

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.

ABBREVIATIONS

5BPA – 5-biotinamido-penthylamine

ACTH – adenocorticotropine

AUC – area under the curve

BKP - biotinyl-GPAVTAAPKK

CD – delta carbon

CNS – central nervous system,

DLK – dual leucine zipper bearing kinase

GAPDH – glyceraldehyde-3-phosphate dehydrogenase

GIP-1 – glucagon inhibitory peptide,

GRH – growth hormone

GST – glutathione S-transferase

HRP – horse radish preoxidase

IGFBP – insulin-like growth factor binding protein

IκB – inhibitory kappa B protein,

NZ – zeta nitrogen,

PEG – polyethilen glycol

PFU – plaque forming unit

ROC – receiver operating characteristics,

ROCK-2 – Rho associated, coiled coil-containing protein kinase 2

VIP – vasoactive intestinal polypeptide

VMD – visual molecular dynamics

1. ÖSSZEFOGLALÁS

A 2 típusú transzglutamináz (TG2) a fehérjék Ca^{2+}-függő poszttranszlációs módosítását katalizáló enzim, amely izopeptid kötéseket hoz létre a glutamin és lizin oldalláncok között. A TG2 több mint 130 azonosított szubsztráttal rendelkezik, de a pontos szubsztrát felismerési mechanizmus még nem ismert. A szubsztrátpreferencia vizsgálatához összegyűjtöttük az eddig azonosított TG2 szubsztrátokat és létrehoztuk a TRANSDAB Wiki (http://genomics.dote.hu/wiki) transzglutamináz szubsztrát adatbázist.

A TG2 által preferált szekvenciák vizsgálatára fág bemutató rendszert alkalmaztunk és a random heptapeptideket bemutató könyvtárból kiszűrtük a rekombináns TG2-höz kötődő fágokat. A kapott peptidek szekvenciájában leggyakrabban előforduló, a szubsztrát glutaminok környezetére jellemző pQx(P,T,S)l konszenzus motívum összhangban van az eddig azonosított szubsztrátok szekvenciájával. A pQx(P,T,S)l motívum jelenléte alapján szubsztrátnak jósolt SWI1/SNF1 kromatin remodeling faktor N-terminális részletét megszintetizáltuk és bebizonyítottuk, hogy a TG2 felhasználja szubsztrátként. Eredményeink azt sugallják, hogy a kombinatórikus módszerrel meghatározott szekvenciáknak *in situ* relevanciája lehet. A fág bemutató rendszer mellett *in silico* módszereket alkalmazva is megpróbáltuk feltérképezni a szubsztrát glutamin illetve lizin oldalláncok környezetében jelen levő aminosav oldalláncok fontosságát, de sem az *in vitro* sem az *in silico* megközelítés nem adott teljes magyarázatot a TG2 szubsztrátspecificitására.

A TRANSDAB Wiki adatbázisban található szerkezeti információkat felhasználva megkíséreltünk fényt deríteni arra, hogy milyen térbeli kritériumok szerint válogatja ki a TG2 szubsztrátjait. A glutamin illetve lizin oldalláncok térszerkezeti elemzéséből nyert nagy mennyiségű adatot logisztikus regressziós analízis segítségével összehasonlítottuk és

olyan prediktor aminosavakat kerestünk, amelyek meghatározzák, hogy mely oldalláncokat módosítja a TG2. A prediktor aminosavak jelenlétét vagy hiányát, mint szelekciós kritériumokat felhasználva sikerül három új TG2 szubsztrátot azonosítani.

Azokban a fehérjékben, amelyek nem rendelkeznek kristályszerkezettel a TG2 szubsztrátspecifictásának egyik meghatározó tényezője a rendezetlenség jelenléte; a TG2 azokat az oldalláncokat használta fel nagyobb valószínűséggel szubsztrátként, amelyek rendezetlen régióban helyezkedtek el.

Az összegyűjtött szekvencia és szerkezeti adatok új megvilágításba helyezik e változatos enzimaktivitással rendelkező fehérje szubsztrátspecificitását. A sejtben a TG2 számos kompartmentben fordul elő és transzglutaminázként működve változatos szubsztrát fehérjéket módosít, de rendelkezik GTP/ATPáz aktivitással, kinázként, protein diszulfid izomerázként is működhet vagy a sejtek felszínén integrinhez vagy fibronektinhez kapcsolódhat. A szubsztrátok felismeréséhez szükséges komplex fiziko-kémiai interakciók megértéséhez nem elégséges csak a lineáris szekvenciában rejlő információkat használni, hanem egyaránt szükség van a térbeli szerkezetben kódolt információkra és a rendezetlenség jelenlétének vizsgálatára.

2. 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 (Folk and Finlayson 1977). In humans nine transglutaminase genes have been identified and eight of them code active enzymes (Griffin et al. 2002). 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 (Mehta 2005). The erythrocyte band 4.2 protein is catalytically inactive having a structural role in the erythrocyte membrane skeleton (Mehta 2005).

Beside the crosslink formation they also catalyze a) the deamidation of glutamine residues to glutamate in various proteins (Folk 1983), b) the hydrolysis of epsilon (gamma-glutamyl)lysine isopeptides as it was shown in the case of transglutaminase 2 and blood coagulation factor XIIIa (Parameswaran et al. 1997), c) the hydrolysis of certain esters such as p-nitrophenyl-acetate (Folk 1983), and d) the attachment of long-chain omega-hydroxyceramides to proteins by ester bond formation as it was shown in case of keratinocyte transglutaminase (Nemes et al. 2000). TG2 promotes cell-matrix interactions binding to fibronectin and integrins and together with epidermal transglutaminase have GTPase activity; in the presence of GTP TG2 functions as a G protein participating in signaling processes (Fesus and Piacentini 2002). TG2 has protein disulphide-isomerase activity (Hasegawa et al. 2003) and recently it was shown to have kinase activity phosphorilating several proteins (Mishra and Murphy 2004). It can also function as a BH3-

only protein; through its BH3 domain it is able to interact with the pro-apoptotic Bax senzitizing cells towards apoptosis (Rodolfo et al. 2004).

2.1. Characteristic features of transglutaminase 2

Transglutaminase 2 (TG2) is a ubiquitous member of the transglutaminase family found in many tissues and cell types (Iismaa 2002). 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 (Lesort et al. 1998, Iismaa 2002, Griffin et al. 2002). The human TG2 protein is a 76 kDa, 688 amino acid containing protein (Fesus and Piacentini 2002) encoded by the ~37 kbp TG2 gene found on the chromosome 20q12 (Grenard et al. 2001).

2.1.1. Structure and reaction mechanism of transglutaminase 2

Transglutaminases evolved from the papain-like cysteine proteinases (Lorand and Graham 2003) having cysteine in their active site (Folk and Cole 1966) but the unique specificity for amine substrates distinguish them from the evolutionary related papain family of enzymes (Curtis et al. 1974). The catalytic triad consists of Cys-His-Asp or Cys-His-Asn and is conserved through the active transglutaminases (Lorand and Graham 2003, Griffin et al. 2002).

The reaction catalyzed by TG2 is an acyl transfer reaction (Figure 1) in which the active site thiol group of Cys277 (Lee et al. 1993) reacts with the γ-carboxamide group of glutamine residue forming the first oxyanion intermediate. In the acylation step the acylenzyme intermediate is formed and the ammonia is released. In the next step the ε-

amino group of lysine residue attacks the thioester bond (nucleophilic attack) forming the second oxianion intermediate followed by the deacylation step when the enzyme is regenerated and the crosslinked product is formed (Pinkas et al. 2007, Iismaa et al. 2003).

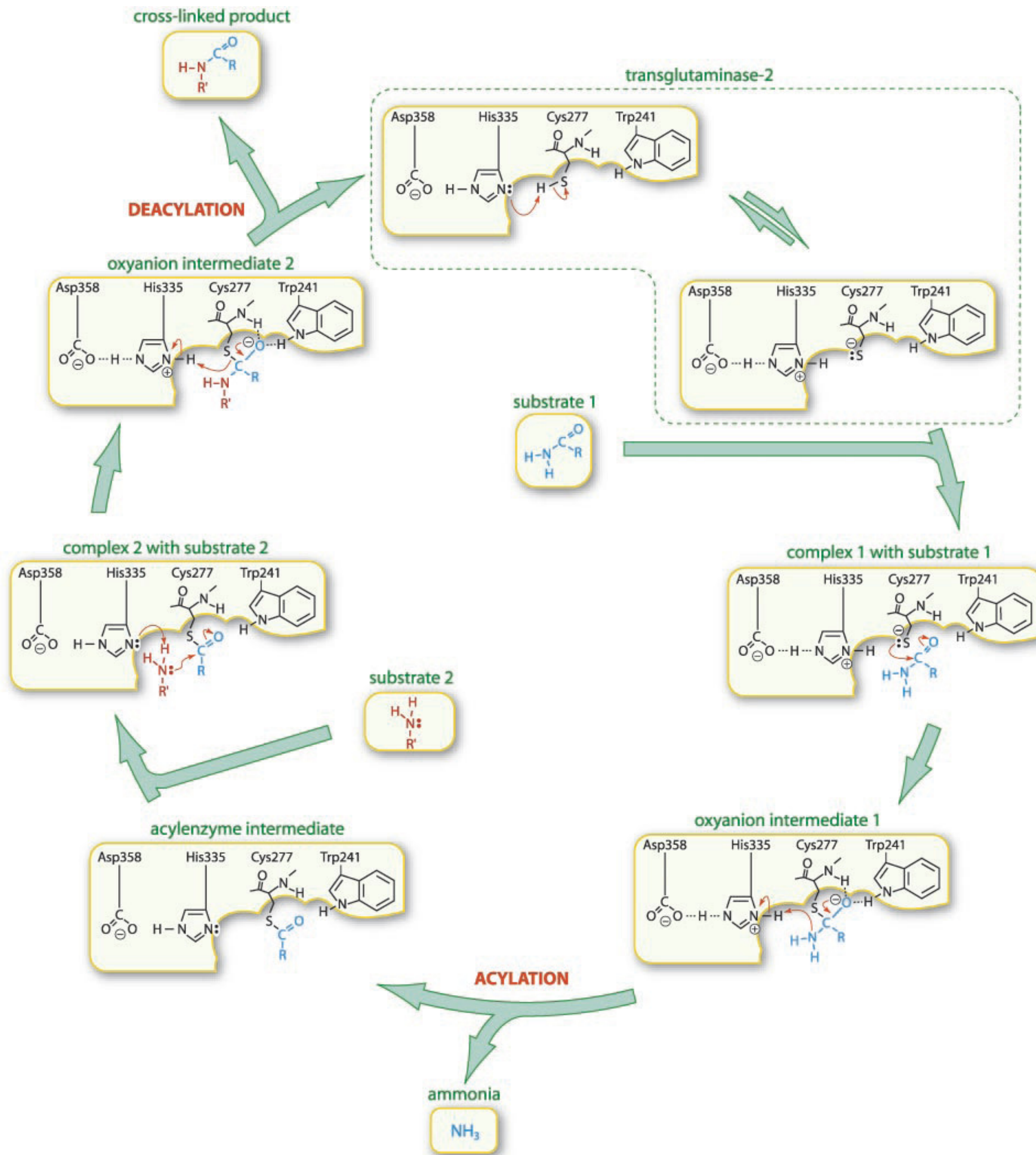


Figure 1. Reaction pathway and a proposed mechanism for TG2-catalyzed transamidations, based on the papain-reaction mechanism (Iismaa et al. 2003).

When the amine substrate is not available, water can attack the thioester bond resulting in deamidated end-product (Folk 1983). The rate limiting step in the catalysis is the acylation step, the formation of the acylenzyme intermediate (Iismaa et al. 2003). Beside the catalytic triad the conserved Trp241 is essential for the catalytic activity having role in the stabilization of the intermediate states (Murthy et al. 2002).

The human TG2 has four domains: the core domain holding the catalytic Cys277-His335-Asp358 triad and the Trp241, an N-terminal β-sandwich domain and two C-terminal β-barrels, belonging to the fibronectin type III CATH superfamily (Pinkas et al. 2007, Liu et al. 2002). 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 (Liu et al. 2002).

2.1.2. The activation of transglutaminase 2

To exert their transamidating activity, the transglutaminases need to be activated. In contrast to the FXIIIa and TG3, which require proteolytic cleavage and Ca-ions for their activation, the TG2 needs only the presence of Ca-ions. 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 (Figure 2) the C-terminal β-barrels are displaced by almost 120 Å (Pinkas et al. 2007) leading to the appearance of a tunnel where the catalytic Cys277 is located accompanied by the two conserved Trp241 and Trp332 residues. In the GTP bound closed form, the TG2 acts as a G protein participating in different signaling processes (Begg et al. 2006) but when the intracellular Ca^{2+} concentration elevates and the GTP is ablated – in case of injury – the Ca^{2+}-bound enzyme

achieves its active conformation leading to the transamidation of the cellular proteins (Begg et al. 2006, Fesus et al. 1989). A disulphide bond between Cys370-Cys371 was observed in the open form of TG2 causing peptide backbone distortion by the adoption of *cis* configuration (Pinkas et al. 2007) and the presence of this disulphide bond can influence the ability of TG2 to adopt the closed conformation upon incubation with GTP (Begg et la. 2006).

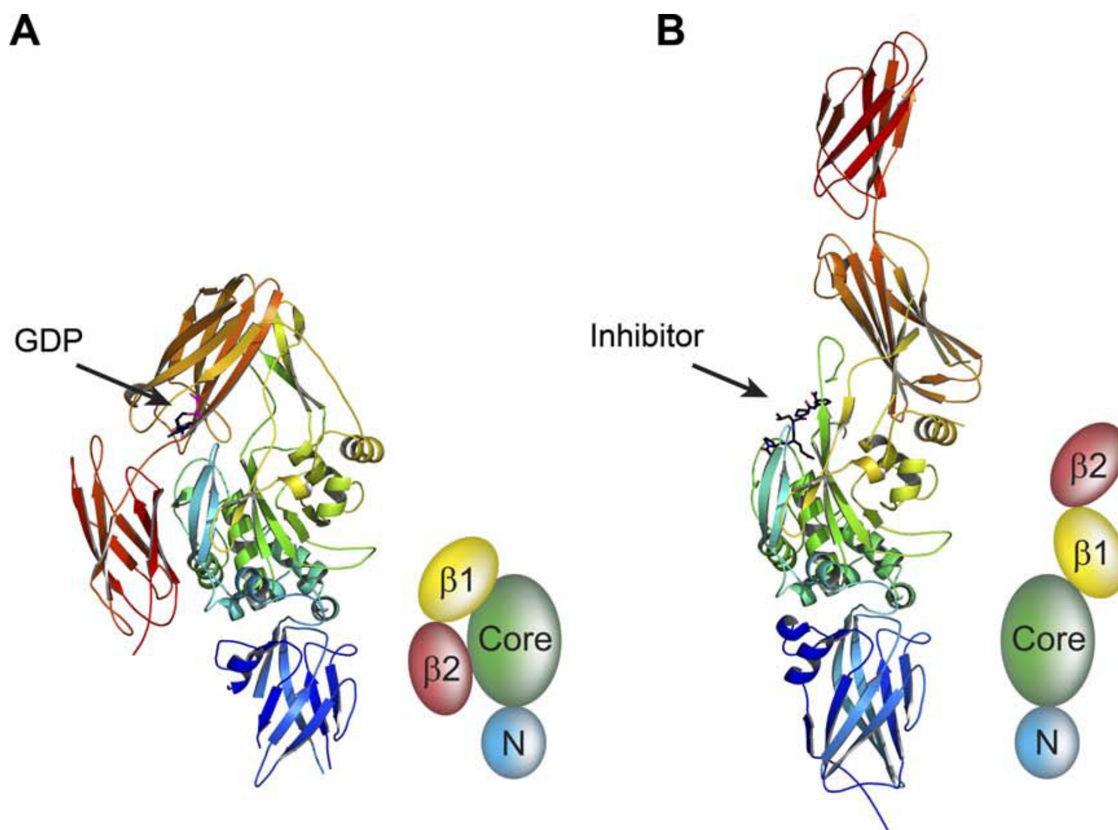


Figure 2. The activation of TG2. In the GDP bound closed conformation (A) the two barrel domains close the entrance to the catalytically active Cys277 on the core domain. Upon activation a large conformational change occurs, the enzyme will adopt an open, catalytically active form (B). N is for the N-terminal β-sandwich domain, β1 and β2 for the C-terminal β-barrel domains and Core for the core domain (Pinkas et al. 2007).

Another possibility for the regulation of TG2 activity is the nitrosilation by nitric oxide (Lai et al. 2001). From the 18 possible cysteines 15 can be nitrosilated and denitrosilated in a Ca^{2+}-dependent manner influencing the transamidating activity of the enzyme (Lai et al. 2001). The interaction of TG2 with sphingosylphosphocholine reduces the Ca^{2+} requirement for the activation of the transglutaminase activity (Lai et al. 1997).

2.2. Biological functions of transglutaminase 2

2.2.1. 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, the TG2 has various roles.

As a transamidating enzyme, it can modify various proteins. In the cytosol, the cytoskeletal protein actin is a physiological TG2 substrate and has a key role in stabilizing cellular morphology, preventing the release of cytoplasmic material from cells and exerting a morphogenic role in programmed cell death (Nemes et al. 1997). The β-tubulin have been shown to be a TG2 substrate (Maccioni and Arechaga 1986) and in some cells β-tubulin probably serves as an anchorage site for binding of TG2 to the cytoskeleton (Piredda et al. 1999). The RhoA, a member of Ras superfamily, is also a substrate for TG2. After deamidation RhoA has decreased GTPase activity functioning as a constitutively active G protein promoting the formation of stress fibers and focal adhesion complexes, while the transamidated form has increased binding to ROCK-2 (Singh et al. 2001). In the heart the transglutaminase activities by modifying troponin I, troponin T and probably

other proteins can act as a switch determining whether apoptosis or necrosis occurs during myocardial ischemia (McDonough et al. 1999). Depending on the cell type, TG2 can exert pro-apoptotic or anti-apoptotic effects (Fesüs and Szondy 2005) and 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 (Piredda et al. 1997).

It is not clear yet how TG2 influences the energy status of the cell, but it can covalently modify fructose 1,6-bisphosphatase, phosphoglycerate dehydrogenase, phosphorylase kinase, mitochondrial aconitase, α-ketoglutarate dehydrogenase and glyceraldehyde-3-phosphate dehydrogenase – these latter three enzymes have reduced activity upon transglutamination (Esposito and Caputo 2005, Kim et al. 2005, Cooper et al. 1997, Ichikawa et al. 2008).

TG2 is implicated in the modulation of inflammatory processes, influences inflammatory cytokine production, by crosslinking free IκBα leading to NFκB translocation to the nucleus (Park et al. 2006), and by crosslinking and increasing the Ca^{2+}-sensitivity to phospholipid binding of annexin I (Ando et al. 1991).

It seems that the TG2 has protective functions; it might be involved in the autophagosome formation as different cytosolic TG2 substrate proteins such as arginase I, betaine-homocysteine S-methyltransferase, glyceraldehyde-3-phosphate dehydrogenase, GST and fructose 1,6-bisphosphatase were found to be enriched in the autophagosomal membranes (Overbye et al. 2007, Ichikawa et al. 2004, Ichikawa et al. 2008). Also the heat shock protein family members (hsp60, hsp70, hsp90), several small heat shock proteins (hsp27, lens crystallins) and ubiquitin may also act as substrates (Esposito and Caputo 2005, Nemes et al. 2004) suggesting a role in defense against misfolded proteins.

Translocation of TG2 to the nucleus (Lesort et al. 1998) might indicate its role in the transcriptional regulation. Core histones, the transcription factor SP1 and the androgen receptor were shown to be substrates for the enzyme (Kim et al. 2002, Han et al. 2000, Mandrusiak et al. 2003). Crosslinking by TG2 may modify the function or activity of substrate proteins as it happens in the case of SP1 transcription factor which has enhanced DNA binding upon transglutamination (Han et al. 2000). In the nucleus the retinoblastoma gene product, pRB, has a pivotal role in regulating cell cycle controlling the G1/S phase transitions. It was shown that pRB is a substrate for TG2 in U937 derived cells undergoing apoptosis and its polymerization might represent a key signal for the initiation of apoptosis (Oliverio et al. 1997).

TG2 is present on the cell surface promoting cell-matrix interactions by binding to fibronectin and integrins (Griffin et al. 2002), and by modifying extracellular proteins such as collagen and osteopontin (Aeschlimann and Thomazy 2000), it is implicated in extracellular matrix remodeling, tissue repair and wound healing. Proteins from the mineralized compartment of bone like osteonectin and SIBLING proteins from teeth may serve as substrates for TG2, linking the transglutaminase activity to bone and teeth development (Aeschlimann and Thomazy 2000, Kaartinen et al. 2002, Kaartinen et al. 2005). 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 (Esposito and Caputo 2005, Alford et al. 2007).

Upon transglutamination the activity of the modified enzymes may change. In case of crosslinked CD38 ADP-ribosyl cylase enzyme the cyclase activity will dominate over the hydrolase activity (Umar et al. 1996) whereas in other cases the activity decreases

(DLK, nuclear GAPDH) or increases (midkine, PLA_2) (Hebert et al. 2000, Cooper et al. 1997, Kojima et al. 1997, Cordella-Miele et al. 1990).

TG2 is implicated in the patophysiology of neurodegenerative diseases such as Huntington's disease, Alzheimer's disease, Parkinson's disease and other neurodegenerative disorders as well as nervous system injuries (Cooper et al. 1997, Ruan and Johnson 2007). The expanded poly-glutamine containing repeats associated by the mutated form of huntingtin, the amyloid beta, alpha synuclein, and different isoforms of tau protein are good substrates for the enzyme (Cooper et al. 1997, Rasmussen et al. 1994, Junn et al. 2003, Murthy et al. 1998). The exact role of the TG2 in the neurodegenerative disorders is not known, it seems to participate in the stabilization of toxic oligomers of the disease-relevant proteins (Ruan and Johnson 2007).

2.2.2. Available information on substrate preference of transglutaminase 2

Physiological and pathological roles of transglutaminase 2 can be understood only if we know what the *in situ* TG2 substrates are and how the substrate specificity of the enzyme is determined. The substrate binding pocket of TG2 is organized in a way which permits the proper orientation of peptide bound glutamine residues while neither the asparagine nor the free glutamine is used by the enzyme; moreover a strong stereospecificity toward the L-isomer was observed (Folk 1983). It was proposed that the TG2 has an extended active site and the interactions made with the oligopeptides influence the catalytic efficiency exerted toward glutamine residues (Folk 1983). To study the role of different amino acids around glutamine residues on substrate effectiveness the amino acid sequence around substrate glutamine residues was studied extensively. Gorman and Folk examined the effect of deletion or substitution of different amino acids from synthetic

pentadecapeptides and decapeptides derived from β-casein and suggested 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 (Gorman and Folk 1981, Gorman and Folk 1984).

The amino acid sequences surrounding the glutamine residues which serve as an amine acceptor site in transglutaminase catalyzed cross-linking reaction were compared in 21 substrate proteins by Aeschlimann et al. (Aeschlimann et al. 1992). They 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 (Aeschlimann et al. 1992).

The study of substance P analogues revealed further insights into the role of different amino acids in the surrounding of substrate glutamines. When the glutamine residue was located at either end of the molecule was not recognized by the TG2, but when it was located in the second or third position from the N-terminus it could be modified efficiently by the enzyme (Pastor et al. 1999). 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 sides 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 (Pastor et al. 1999).

As another approach, the residues which were missing from the surrounding of the substrate glutamines were examined and it has been proposed that the presence of discouraging features would prevent a glutamine to be used as substrate by TG2. The study was done on crystal structures deposited in PDB; however the results were presented on

sequence level emphasizing the importance of charged residues in the surrounding of glutamine residues (Coussons et al. 1992).

The plant storage protein gliadin was found to be an excellent substrate for TG2 being deamidated specifically at Q65 (Arentz-Hansen et al. 2000). In most cases the TG2 recognizes the QxP sequence rather than the QP or QxxP sequences (Vader et al. 2002) in the gliadin peptide and the resulted deamidated peptide can serve as an anitgen in coeliac disease (Arentz-Hansen et al. 2000).

The importance of proline in the recognition of substrate glutamine was confirmed by a fully 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 (Sugimura et al. 2006).

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 (Esposito and Caputo 2004). To study the amino acid residues influencing the amine donor substrate properties of lysines Groenen et al. and Grootjans et al. have modified the native sequence around substrate lysine in alpha A-crystallin (Groenen et al. 1994, Grootjans et al. 1995). 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 (Groenen et al. 1994, Grootjans et al. 1995).

3. 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.

4. MATERIALS AND METHODS

Reagents and peptides

The neuropeptide Y, exendin 4 and orexin B were purchased from GeneScript Corp. (USA), the NQEQVSPLTLLK peptide was a kind gift of Levente Kárpáti from the Department of Clinical Biochemistry and Molecular Pathology, University of Debrecen, the biotinyl-GPAVTAAPKK (BKP) and SGYGQQGQTPYYNQQSPHPQQQQP peptide (SnQ1) was synthesized by Joe Gray at the Molecular Biology Unit, University of Newcastle upon Tyne. The human recombinant His-tagged TG2 was kind gift from Róbert Király from the Department of Biochemistry and Molecular Biology, University of Debrecen. Recombinant human TG2 was produced as a GST-fusion protein (GST-TG2) in *Escherichia coli* DH5α using pGEX-2T vector construct. All other materials were reagent grade and were purchased from Sigma if not indicated otherwise.

Phage biopanning

Peptides were selected from a commercially available 7-mer random library that is displayed on phage M13 via N-terminal fusion to the minor coat protein, g3p, with a diversity of $2.8x10^9$ (Ph.D.-7 Phage Display Peptide Library Kit, New England Biolabs). The recombinant human GST-TG2 served as the target in the selection process. One microgram of GST-TG2 in 50 mL of TBS was immobilized on glutathione-sepharose 4B beads (50%, 50 mL) (Amersham). The slurry was washed three times with 1 mL of TBS by centrifugation (3000 rpm, 10 min) and decantation, blocked with 100 mL of blocking buffer (0.5% BSA in TBS) for 1 h at 4°C, and washed using TBST (TBS containing 0.1% Tween 20). 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 at

25°C with continuous shaking. To remove unbound phages, the slurry was washed 10 times with 1 mL of TBST, and the phage population remaining attached was eluted with shaking in 100 mL of 0.1 M glycine-HCl pH 2.0 (10 min at 25°C) and was neutralized with 150 mL of 1 M Tris-HCl pH 8.0. The eluate was amplified in *E. coli* ER2537 cells, purified by precipitation with PEG/NaCl, titrated as described in the standard protocol, and used (10^{11} PFUs) in the next selection cycle. In the second and third biopanning rounds, bound phages were eluted with 100 mL of 5 mM 5BPA (Pierce) in TBS by 30 min of shaking at 25°C. After the second and third rounds, individual phage clones were isolated, and ssDNA was prepared according to the standard protocol. The insert sequence was determined using the ‘‘-96’’ M13 sequencing primer and an ABI PRISM 310 Genetic Analyzer (PE Applied Biosystems).

Enzyme-linked immunosorbent assay (ELISA)

Phage clones from randomly picked plaques were amplified and purified by precipitation with PEG/NaCl. The concentration of phage stocks was estimated from the absorbance at 260 nm (when A260 = 1, [phage] = $1.1x10^{-8}$ M) and dilution series (10^{-4}–10 nM) were prepared in blocking buffer containing 10 mM DTT and 5 mM $CaCl_2$. Ninety-six-well microtiter plates were coated with 10 mg/mL GST-TG2 in TBS containing 10 mM DTT (2 h, 25°C), and blocked with 200 mL of blocking buffer (2 h, 25°C). After three washes with TBST, 50-mL aliquots of the phage solutions were incubated in the plate for 1 h at 25°C, followed by 10 washes with TBST. Bound phage particles were quantified adding a HRP-conjugated anti-M13 antibody (Amersham) (1:5000 dilution in 0.25% BSA/TBS) for 1 h at 25°C. After washing, the plates were incubated with 1 mg/mL tetramethyl-benzydine (100 mL/well) for 10 min at 25°C, the color was developed by adding 50 mL of 2M 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 of phage-displayed peptides, aliquots of amplified phage clones ($5x10^{10}$ PFUs) were dissolved in Tris-HCl at pH 8.5, 15% [v/v] glycerol, 10 mM DTT, 1 mM 5BPA as the amine-donor substrate, and 2 mg of GST-TG2. Reactions were initiated by the addition of 5 mM $CaCl_2$, and after 1 h incubation at 37°C were stopped by the addition of 5 mM EDTA. Reaction products were analyzed by Western blotting. The SGYGQQGQTPYYNQQSPHPQQQQP peptide was dissolved to a final concentration of 200 mM in Tris- HCl at pH 8.5, 5 mM DTT containing either 1 mM BKP or 1 mM 5BPA as amine donor substrate, and 5 mM $CaCl_2$, in a final volume of 250 mL. Reactions were initiated by the addition of 5–10 mg of GST-TG2 and incubated for 1 h at 37°C. High molecular mass (>10 kDa) components of the reaction mixtures were partially removed by ultrafiltration (Centricon YM-10), and reaction products were analyzed by mass spectrometry. To analyze the 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), 1 mM of 5BPA as amine donor substrate, 5 mM DTT and 5 mM $CaCl_2$ in Tris-HCl pH 8.5 buffer to a final volume of 20 μl. Reactions were initiated by the addition of 8 μg of His-tagged TG2 and incubated for 2 hours at 37°C, then stopped by the addition of 5 mM EDTA and samples were kept at -20°C until analysis with mass spectrometry. In case of exendin 4 (predicted amine donor substrate) the NQEQVSPLTLLK peptide was used as glutamine donor substrate.

Immunoblotting

Aliquots of in vitro TG2-labeling reactions containing $5x10^{10}$ phage particles were run on 10% SDS-PAGE (Mini-Protean III, Bio-Rad). Protein bands were stained by Coomassie Brilliant Blue R or transferred to an Immobilon-P PVDF membrane (Millipore) in a Hoefer

Semi-Dry Blotting apparatus (Bio-Rad). Visualization of biotinylated protein species was carried out using HRP-conjugated streptavidin (Vectastain Western Blotting Kit, Vector Laboratories, Inc.), followed by chemiluminescent staining (ECL Western Blot Detection Kit, Amersham) and detection in an AlphaImager Gel Documentation System (Alpha Innotech).

Mass Spectrometry analysis

Liquid chromatography/mass spectrometry (nanoLC/MS) analyses were performed using a QTRAP nanoLC-MS/MS 4000 ion trap mass spectrometer (Applied Biosystem MDS Sciex), 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 (Agilent) and fractionated by performing a linear gradient of eluent B in eluent A from 0% to 100% B, at a flow rate of 0.5 μL/min. Spectra were acquired from 300 Da. The 100-mL aliquots of SGYGQQGQTPYYNQQSPHPQQQQP peptide containing reaction mixture was injected onto an RP-HPLC C18 Tagra column (Higgins Analytical, Inc.) and fractionated by performing a linear gradient of eluent B in eluent A from 0% to 100% B in 10 min, at a flow rate of 6 mL/min. Spectra were acquired from 200 to 1700 Da. The resulting mass data were elaborated using the Analyst software (Applied Biosystem MDS Sciex).

Database search

The heptapeptide sequences obtained from the selected phage clones after the third round of biopanning were examined for a consensus sequence for TG2 modification. Because no consensus sequence was observed, we have analyzed whether the distribution of amino

acids is different from that expected in the initial random library. The occurrence of amino acids was determined at individual positions around glutamines, and where adjacent Q residues were present, the most C-terminally located one was chosen arbitrarily as the point of reference. 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 listed in TRANSIT (http://bioinformatica.isa.cnr.it/TRANSIT) and TRANSDAB (http://genomics.dote.hu/wiki) transglutaminase substrate databases. It was also examined whether the phage-derived sequences match any transglutaminase substrate protein listed in the TRANSIT database using the Sitematcher searching tool and considering only those results that match with the reactive glutamine residue. As a broader approach, each heptapeptide was searched against the Non-Redundant protein database using the BLAST algorithm with the ''short nearly exact match'' option at the NCBI server to see whether this technique is suitable to predict novel, yet unidentified potential transglutaminase substrates. To address this question, 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 the abovementioned BLAST algorithm with the ''short nearly exact match'' option 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 (www.expasy.org) were used for comparative sequence analysis and intrinsic disorder prediction. In the spatial environment studies the crystal structure data files were originated from PDB (www.rcsb.org). Substrate proteins where the modification site by TG2 was known were retrieved from TRANSDAB Wiki

(http://genomics.dote.hu/wiki). The surface accessibility of amino acid residues was estimated using the Netasa (http://gibk26.bse.kyutech.ac.jp/~shandar/netasa/asaview) web server and an arbitrary threshold of 18 % was established. The amino acids with less than 18% surface accessibility were deleted from the structure 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 sphere showed a skewed distribution and no transformation made it symmetric, therefore these data were treated as ordinal scale. The number of predictor amino acids 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 (Altman 1992) 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. 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.

Performance evaluation

As an evaluation of the performance of the prediction, the sensitivity and specificity of the parameters used in model construction were measured. Sensitivity was defined as the percentage of correctly predicted substrate cases, and specificity was the percentage of correctly predicted non substrate cases. To compare prediction models with different independent variables receiver operating characteristics (ROC) curves were plotted (sensitivity versus 1-specificity) and the area under the curve (AUC) were calculated.

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 each protein were compared using Mann-Whitney U test.

Software

The SEQSTAT and ATOMDIST programs were written by Péter Bagossi from the Department of Biochemistry and Molecular Biology, University of Debrecen and are available upon request. For statistical analyses, the SPSS.15.0 for Windows was used. Graphical analysis was made on Silicon Graphics Fuel workstation using the Sybyl program package (Tripos, St. Louis, MO) and VMD (Humphrey et al. 1996).

5. RESULTS

From the more than 130 so far identified TG2 substrates around 34% is located in the cytosol, 12% in the nucleus, 12% in the plasma membrane and cornified envelope, 4% is present in the mitochondria while 1.5% can be found in other organelles. A significant amount of the substrates is located outside the cell, 9% in the extracellular matrix and 21% in body fluids. 3% of the identified substrates have viral origin and 2% of them originate from other organisms.

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. TG2 has more than 130 known substrates suggesting that the enzyme developed broad substrate specificity contrary to the enzymes with strict substrate specificity, having only one or few substrates. Administrating *in silico* and *in vitro* methods we tried to unravel the promiscuous substrate specificity of TG2.

5.1. The interactive transglutaminase substrate database - TRANSDAB Wiki

TRANSDAB Wiki (http://genomics.dote.hu/wiki) is a continuously increasing database of transglutaminase substrate proteins and interaction partners. Currently TRANSDAB 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 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 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 uses the advantages of wiki platform. 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.

5.1.1. TRANSDAB Wiki database content

Each entry of TRANSDAB contains the substrate peptide or interaction partner name and its alternative scientific names, synonyms, the determination type (in vitro/in situ), the source organism where the protein was studied and the subcellular localization. The UniProt entry for further information can be accessed via link through the UniProt ID and for reference links to relevant scientific article(s) about determination of transglutaminase substrate proteins are also available through their PubMed ID.

The structural information are represented by a picture gallery of reactive glutamine and lysine residues and video image of the rotating molecule, all drawn with VMD (Humphrey et al. 1996) using the structural information from crystal structure files deposited in Protein Databank (PDB).

The crystal structure is also accessible via link to the Protein Databank (http://www.pdb.org/) through the PDBID. A sequence portion containing the substrate glutamine/lysine residues is also given, with reactive residues in italic, to ease the database searches. In pdf and text formats the surface accessibility data obtained with ASAView web server (Ahmad et al. 2004) for those substrate proteins where the modification site is known and the crystal structure contains these residues are also available. For intrinsic disorder prediction there is a link to one of most widely used predictor IUPred (Dosztanyi et al. 2005). Other relevant information about substrate proteins i.e. the change in activity upon transglutaminase action, are listed in Notes. The entries of the database can be browsed or searched by name and can be edited. Each substrate protein is a part of a category type making the information access very easy.

The scientific community is invited to edit the content of the database providing a new platform for the discussions and information sharing.

5.2. Linear sequence determinants of TG2 substrate specificity

The substrate preference for TG2 was studied extensively examining the linear sequence context around substrate glutamine and lysine residues but no consensus sequence was found. Along with the different *in vitro* methods used to study the importance of individual amino acids around substrate glutamine and lysine residues, respectively, computational methods were administrated to find the most favorable or unfavorable features around substrate residues.

5.2.1. Phage display selection of TG2-specific glutamine-donor peptides

To study the favorable primary structures around substrate glutamine residues we have used a random heptapeptide library displayed on the minor coat protein (g3p) of phage M13.

Recombinant human TG2 was immobilized on glutathione-sepharose beads and was incubated with phages expressing the heptapeptide on their surface. After washing the unbound phages off the matrix, the population remaining attached was nonspecifically eluted by low pH. The eluate was then amplified, and the above procedure was repeated twice except that a synthetic amine-donor TG2 substrate 5BPA was used in the elution steps to specifically release phage clones displaying glutamine-donor substrate peptides. During the course of selection, the relative yield of phage recovery increased from 0.12% (first round) to 6.5% (third round), which is consistent with enrichment in binding clones and individual phage clones from the second and third rounds were amplified. The affinity of each clone for TG2 was tested by ELISA, and those that showed concentration-dependent binding (88%) were then subjected to DNA sequence analysis to identify the displayed peptide. 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 (Table 1). The representation of glutamine within the selected pools increased from the initial 5.1% to 8.9% in the second round and to 16.5% in the third round and the proportion of glutamine-containing sequences increased from the 35.7% initial value to 46% in round two and to 75% in round three. In addition, glutamine occurred more than once in 10 of the peptides and these residues were adjacent in seven of them. The above findings demonstrate that our selection strategy facilitates enrichment of phage particles displaying glutamine-containing peptides.

Phage clone[1]	Sequence[2]	Transamidation[3]
PhQ2.7	G**QQQ**TPY	+
PhQ3.9	**QQQ**HFLV	+
PhQ2.2	GL**QQ**ASV	+
PhQ3.13	**QQ**LHLEA	+
PhQ3.4	**QQ**LSIPP	+
PhQ3.1	**QQ**SPLWH	+
PhQ3.2	**QQ**WITVP	+
PhQ2.10	V**Q**ET**Q**RP	+
PhQ3.17	A**Q**H**Q**LEL	+
PhQ3.18	W**Q**IPM**Q**L	+
PhQ3.11	W**Q**TPMNS	+
PhQ3.3	W**Q**RPYPH	(+)
PhQ3.5	S**Q**LWLLP	+
PhQ3.19	S**Q**LTLLP	+
PhQ3.8	T**Q**FNPRY	(+)
PhQ2.8	GL**Q**ATPA	(+)
PhQ2.4	VT**Q**RLPL	(+)
PhQ2.13	T**Q**DTPRT	+
PhQ2.5	**Q**GRIPTS	(+)
PhQ2.6	**Q**SNEPRR	(+)
PhQ2.21	IAHH**Q**FP	+
PhQ3.7	FLTPT**Q**P	(+)
PhQ3.14	AWP**Q**TSA	(+)
PhQ2.9	IKT**Q**ALM*	(+)
PhQ2.1	EKH**Q**PER*	(+)
PhQ3.20	VPNVH**Q**K*	(+)
Consensus	....pQ.Pl.. T S	

Table 1. Deduced amino acid sequence and glutamine-donor substrate character of peptides selected from a phage display library using transglutaminase 2.

[1] The first numeral in phage clone names refers to the selection round in which it was obtained.

[2] Sequences are aligned at the first glutamine or the last one in the case of consecutive glutamines (**.**). Glutamines are shown in bold face. Consensus features (>55%) identified in the environment of glutamines are highlighted. Discouraging features for transamidation are underlined. The presence of lysine is indicated by an asterisk.

[3] Strong (+ sign) or weak ((+) sign) transamidation by TG2 *in vitro* using 5BPA as the amine-donor substrate.

Based on the sequences resulted from the screening of phage display library the characteristic sequence motif pQx(P,T,S)l was established where x, p and l stand for any, polar and aliphatic amino acids, respectively.

5.2.1.1. Sequence analysis and transamidation of selected glutamine-donor peptides

The occurrence of amino acids was determined at individual positions around glutamines (Figure 3) and significant deviations from random distribution were observed at -1 position, where most of the residues are polar (70%); at +2, where serine, threonine, and proline are over-represented (56%); and at +3, where 75% of the residues are aliphatic.

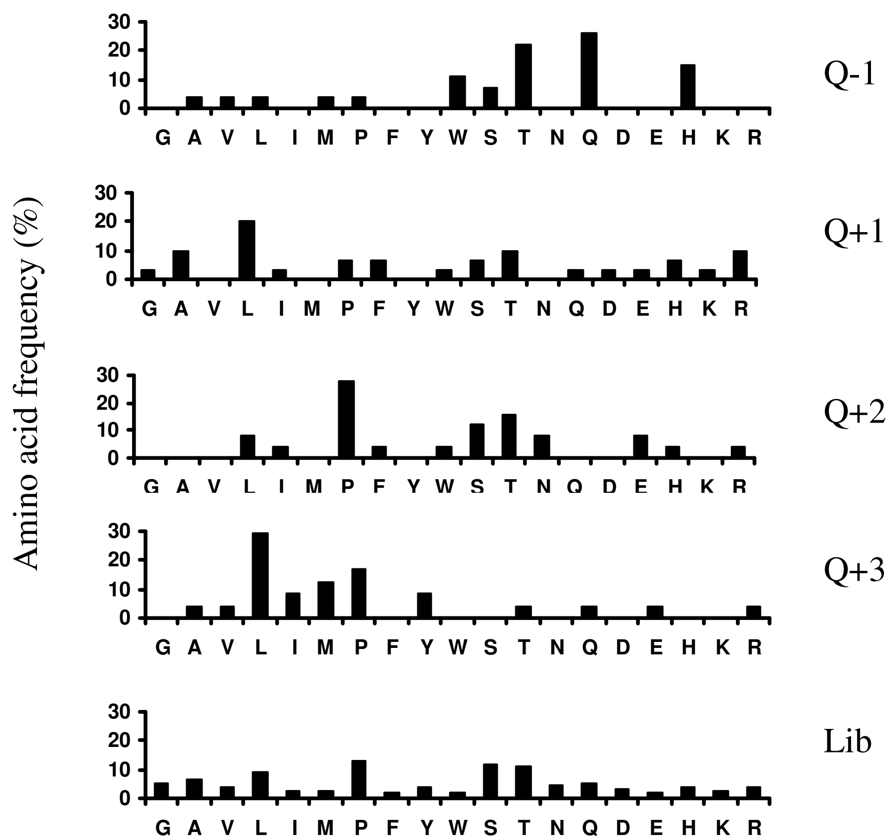


Figure 3. Relative occurrence of amino acids in the environment of glutamines within the selected glutamine-containing peptides compared to those of the initial library. Amino acid frequencies are

shown for relative positions Q -1, Q +1, Q +2, and Q +3 (top diagrams), and the values for the initial library are presented in the bottom diagram.

This is consistent with the dominating sequence pattern pQx(P,T,S,)l identified in the selected peptides.

To test whether the resulted glutamine-containing sequences (Table 1) 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 and were found to be unsusceptible to transamidation (Figure 4) indicating that the phage does not expose efficient TG2-reactive glutamine residues, and therefore transamidation is restricted to the displayed sequence.

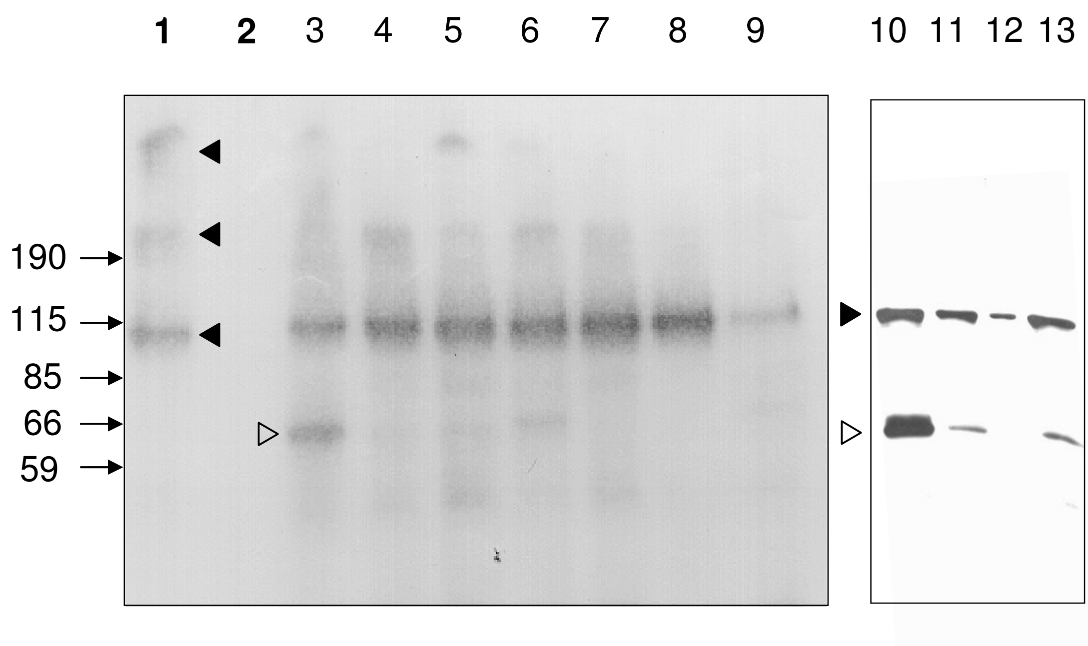


Figure 4. TG2-catalyzed modification of phage-displayed peptides from selection rounds 2 and 3.

1. GST-TG2 incubated with 5BPA, **2.** GST-TG2 incubated with 5BPA and 5 mM EDTA as a negative control; **3.-13.** phage clones incubated with GST-TG2 and 5BPA; **3.** GQQQTPY; **4**. GLQATPA; **5.**

VQETQRP; **6.** GLQQASV; **7.** QGRIPTS; **8.** MPPPMRS; **9.** VTQRLPL; **10.** WQTPMNS; **11.** QQLHLEA; **12.** LMAKPTR; **13.** SQLTLLP. Closed arrowheads mark bands corresponding to GST-TG2 species with incorporated 5BPA. The open arrowhead marks the position of the g3p protein of M13 phage, which runs at an apparent molecular mass of 62 kDa. Lanes 1–9 and lanes 10–13 are from separate experiments.

Strong TG2-mediated 5BPA incorporation was demonstrated for half of the peptides, including those containing two or more adjacent glutamine residues, and the strongest signals were obtained for GQQQTPY, GLQQASV and WQTPMNS.

The established pQx(P,T,S,)l pattern was consistent with sequences in identified substrates listed in the TRANSDAB matching the transamidation-enhancing characteristics and there is a good correlation between the presence of discouraging features (underlined in Table 1) and low efficiency of TG2 catalyzed modification of the corresponding sequences.

5.2.1.2. Transamidation of GQQQTPY-like motifs within a native peptide

Using the GQQQTPY peptide as a representative example of the efficient TG2 substrates 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 (Dallas et al. 2000), which contain two repeats of GQQQTPY-like sequences within one of their two conserved glutamine-proline-rich domains (residues 1–220 and 944–1021 in p270). Therefore, a 27-mer peptide, spanning the S77–S103 segment of p270 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 (5BPA and BKP) to Gln6, Gln8 and Gln22 of the GQQQTPY-like motif of 27-mer peptide (Figure 5)

suggesting that the heptapeptides identified by a combinatorial approach may have *in situ* relevance as TG2 substrates.

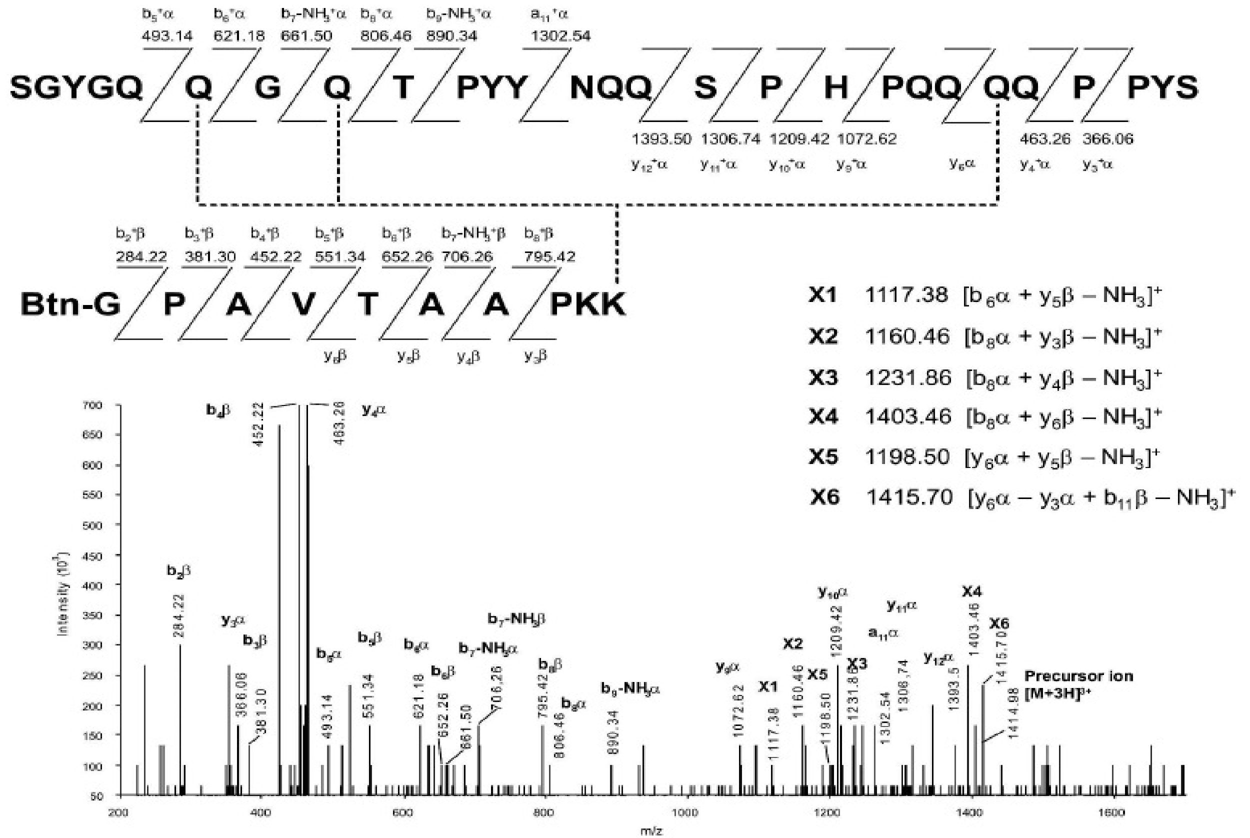


Figure 5. Mass spectrometric identification of TG2-modified glutamine residues in SnQ1. An aliquot of the TG2 reaction mixture of 27-mer peptide and BKP was resolved using LC-ESI/MS, and the triply charged ion at m/z 1414.98, originating from the 27-mer peptide cross-linked to biotinyl-GPAVTAAPKK (BKP), was selected for fragmentation. The MS/MS spectrum obtained from the fragmentation of the above precursor ion is shown below. Peptide sequences and interpretation are given above. Signals X1–X6 are assigned to cross-linked fragments, which are structurally explained on the right. Dashed lines depict the three alternative e(γ-glutamyl)lysine isopeptide bonds between the two peptides as deduced from fragments X1–X6.

5.2.2. *In silico* study of favorable sequence contexts around TG2 substrate glutamine and lysine 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 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 included in Figure 6. 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.

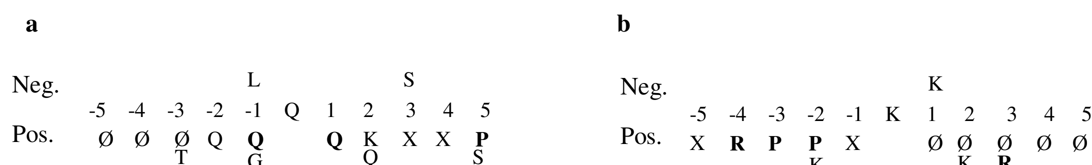


Figure 6. Sequence determinants of TG2 substrate specificity. The results of the examination of substrate and non-substrate sequence contexts with SEQSTAT program were compared with chi square test. Residues present significantly ($p < 0.05$) more often in the sequence context of substrate glutamine (Panel a) and lysine (Panel b) residues are indicated below (Pos.), while the residues present significantly more often around non-substrate residues are indicated above (Neg.) the numbers referring to the position of amino acids. The one letter code for amino acids was used and X and Ø stands for any kind or missing residues, respectively. The Pos and Neg terms refer to the positive or negative effect of the respective amino acid on the substrate selection of TG2 and the highlights refer to the residues mentioned in the literature as important ones in defining TG2 substrate selection.

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 (Figure 6b). 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. (Grootjans et al. 1995).

5.3. 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.

From the 126 examined TG2 substrate proteins, only 23 of glutamine donor and 14 of lysine donor substrate proteins matched the following criteria: 1) the site of modification by TG2 is known, 2) the protein has crystal structure and 3) the structure contains the substrate residues and could be analyzed. 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 (Figure 7a).

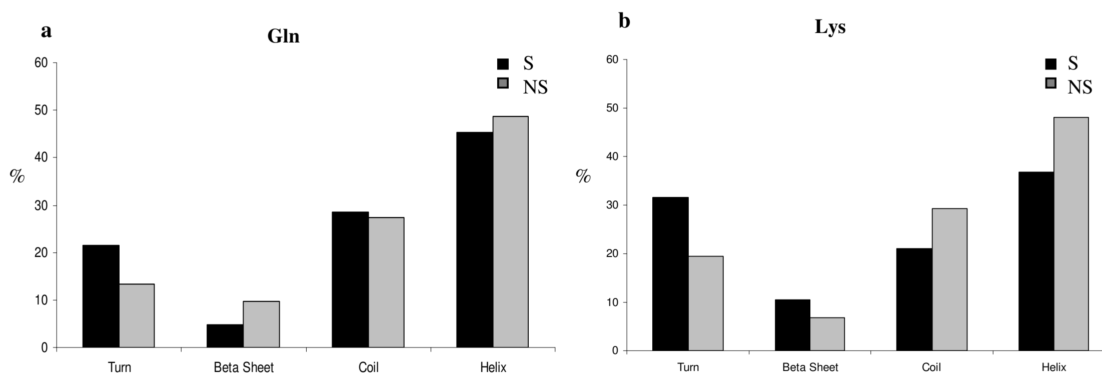


Figure 7. Secondary structure determinants of TG2 substrate specificity. The percent of glutamine (Panel a) and lysine (Panel b) residues situated in the different secondary structures.

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 (Figure 7b).

5.3.1. 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.

5.3.1.1. 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, His and 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 (Figure 8a). 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. To see how good predictors the selected amino acids were, a logistic regression analysis test was performed using the presence of Thr at 5 Å, Arg, His and Leu at 10 Å, Val at 11 Å, Arg and Phe at 15 Å as selection criteria. The sensitivity and specificity of the analysis were calculated and found as 77.6% and 75.4%, respectively. When the whole structures were used in such calculations (Figure 8b), 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. Using these amino acids as selection criteria in the test logistic regression 68.9% sensitivity and 79.4% specificity were observed.

5.3.1.2. 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 (Figure 8c); the sensitivity was 74.8% while the specificity was 74.6% in the logistic regression test.

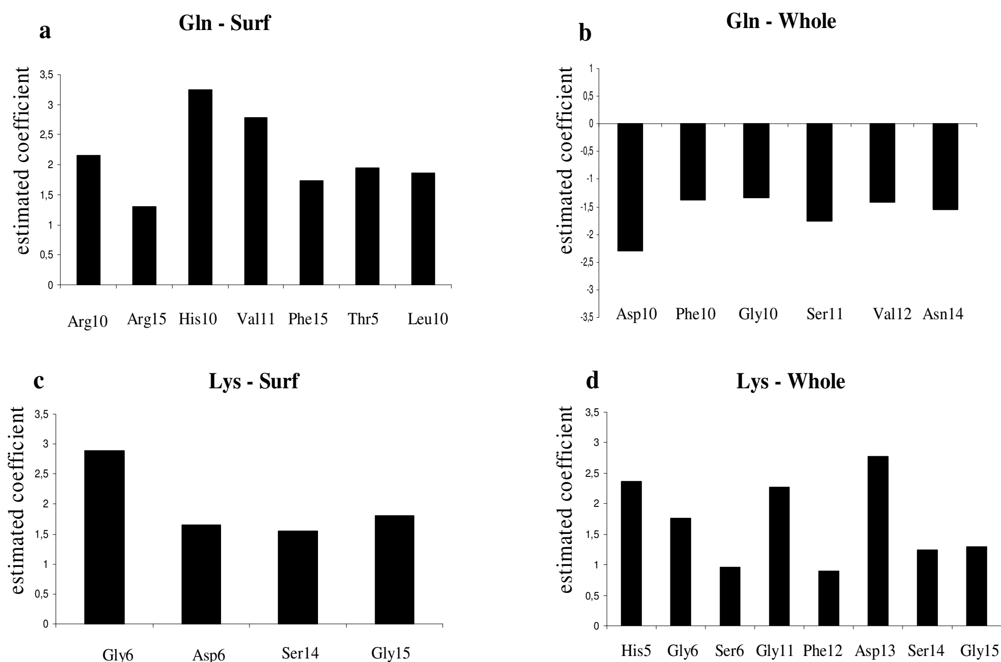


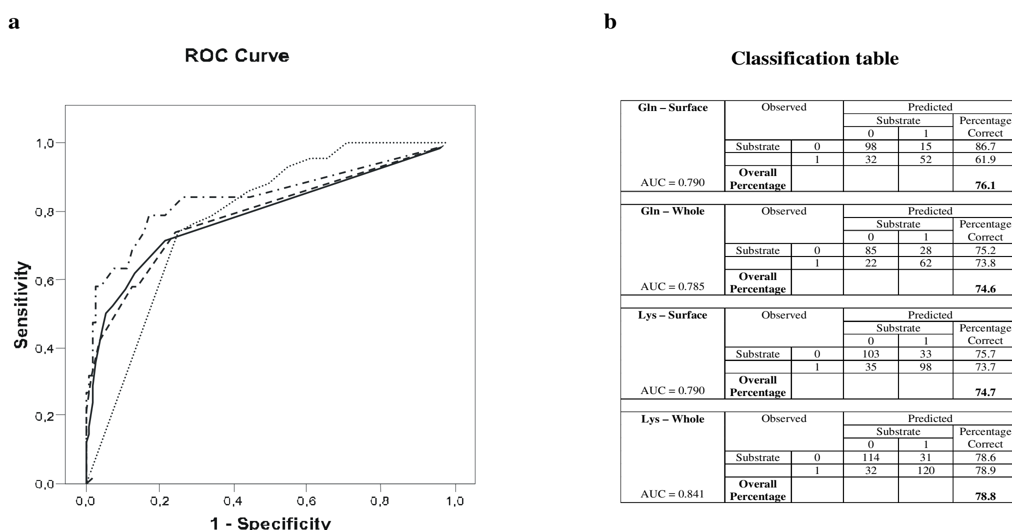
Figure 8. Spatial analysis of the environment of glutamine and lysine residues in substrate proteins and the determination of predictor amino acids by logistic regression analysis. The estimated coefficient (Y axis) for the predictor amino acids (X axis) around substrate glutamine and lysine residues, respectively, were plotted. The positive values indicate the odds with which the presence of the respective amino acid increases the possibility of the glutamine or lysine to be used as substrate by TG2. The terms Arg10, Gly06, etc. refers to the amino acids situated at 10 Å or 6 Å distance, respectively, from the CD of glutamine (in case of glutamine substrates) or NZ of lysine residues (in case of lysine substrates). On **Panel a**, the predictor amino acids on the surface of substrates in the surrounding of glutamines, **Panel b,** predictor amino acids in the whole structure around glutamines, **Panel c,** the surface accessible predictor residues around lysines and **Panel d**, predictor amino acids in the whole structure around lysine residues were examined.

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 (Figure 8d). The sensitivity in the test logistic regression analysis was 79.5% and the specificity was 78%. 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.

5.3.1.3. Validation of the logistic regression analysis results

To test which of the calculations can give the best results, a receiver operating characteristics (ROC) curve analysis was performed and the area under the curve (AUC) was calculated. Values of AUC close to 1.0 suggest a perfectly performing prediction whereas values close to 0.5 suggest that the predictor performs no better than random. The AUC for all predictions were higher than 0.785 (Figure 9), and the most accurate prediction could be done when the surrounding of the lysine residues were examined and the whole structures were used (AUC=0.841).



Classification table

	Observed		Predicted		
Gln – Surface			Substrate		Percentage Correct
			0	1	
	Substrate	0	98	15	86.7
		1	32	52	61.9
AUC = 0.790	**Overall Percentage**				**76.1**
Gln – Whole	Observed		Predicted		
			Substrate		Percentage Correct
			0	1	
	Substrate	0	85	28	75.2
		1	22	62	73.8
AUC = 0.785	**Overall Percentage**				**74.6**
Lys – Surface	Observed		Predicted		
			Substrate		Percentage Correct
			0	1	
	Substrate	0	103	33	75.7
		1	35	98	73.7
AUC = 0.790	**Overall Percentage**				**74.7**
Lys – Whole	Observed		Predicted		
			Substrate		Percentage Correct
			0	1	
	Substrate	0	114	31	78.6
		1	32	120	78.9
AUC = 0.841	**Overall Percentage**				**78.8**

Figure 9. Accuracy estimation using ROC curves. Panel a, The sensitivity was plotted versus (1-specificty) in case of each four examinations. The prediction using the whole structure for lysine (-·-·-·-) was the best with area under the curve: 0.841. The predictions using the surface accessible residues for lysine (------) and glutamine (——) were very similar, with area under the curve: 0.790 for both predictions. The prediction using the whole structure for glutamine (·······) was also good; the area under the curve was 0.785. Panel b, Classification table containing the true positive and true negative matches for each four prediction used in the ROC statistics. 0 refers to non substrate cases and 1 refers to substrate cases.

To check whether the results from the logistic regression analyses are really reflected at the level of protein structures and are not just statistical artifacts, the crystal structure of the substrate proteins was examined. The percent of those substrate and non substrate residues which contained the amino acids having positive or negative effect on substrate selection in their environment was calculated and compared to each other.

When the distribution of amino acids with positive effect in the surrounding of glutamine residues was examined (Figure 10a), it was shown that the percent of substrate glutamines containing His at 10 Å, Val at 11 Å, Arg and Phe at 15 Å in their surrounding was significantly higher than the percent of non substrate residues with such features. In case of Thr at 5 Å and Arg and Leu at 10 Å the tendency was similar but the differences were not significant. The amino acids with negative effect were significantly more abundant in the surrounding of non substrate glutamines (Figure 10b).

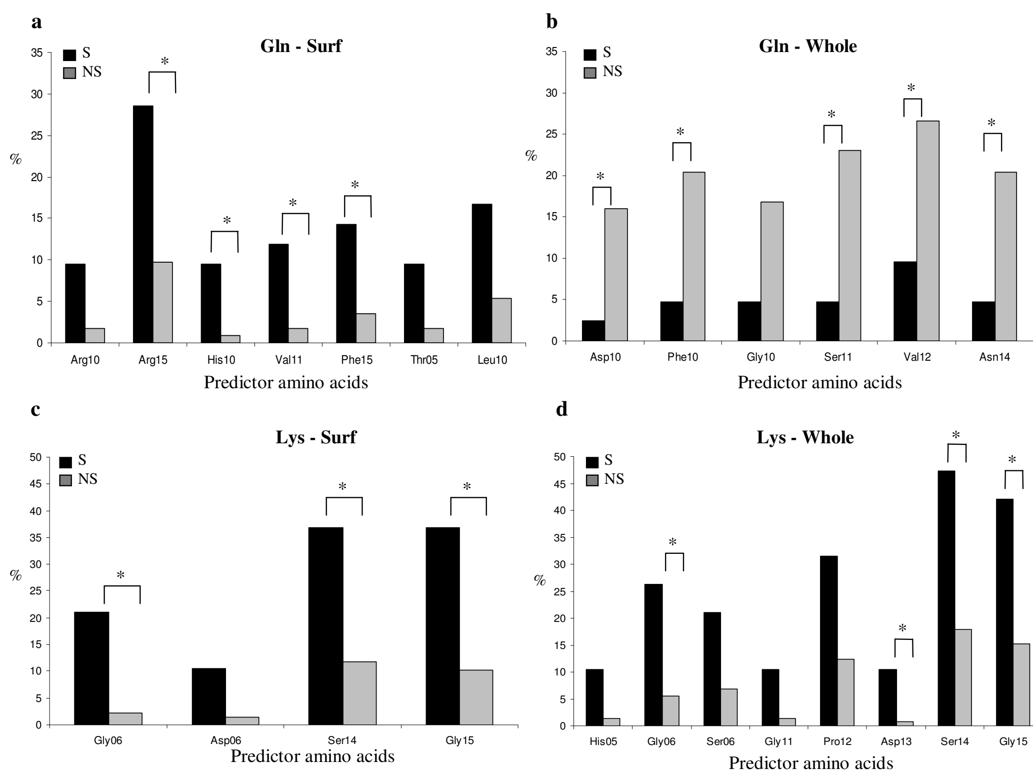


Figure 10. Validation of logistic regression analysis results. The percent of substrate and non-substrate

residues (Y axis) which contain in their surrounding the predictor amino acids (X axis) were plotted. **Panel a,** Predictor surface accessible amino acids around glutamine residues. **Panel b**, Predictor amino acids derived from the examination of whole structure in case of glutamine residues. **Panel c**, Predictor surface accessible amino acids around lysine residues. **Panel d**, Predictor amino acids derived from the examination of whole structure in case of lysine residues.* $p<0.05$.

The percent of those substrate lysines which contained amino acids with positive effect in their surrounding was significantly higher as compared to the non substrate ones (Figure 10c and 10d).

These data suggest that the discrimination between substrate and non substrate residues might be different in case of glutamines and lysines. In case of glutamines it might be based on the presence of amino acids with positive effect and the lack of amino acids with negative effect; while, the presence of amino acids with positive effect alone can probably determine the lysine residues which can be used as substrates for TG2.

To test this further, we checked the presence of amino acids with positive or negative effects in the surrounding of the glutamine and lysine residues in IκB protein. Though its crystal structure is available, this protein was not included in our previous analysis. All glutamines which had amino acids with positive effect in their surrounding but did not contain amino acids with negative effect were considered as possible glutamine donor substrates. The lysine residues having 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 in their surrounding as possible lysine donor substrates. Based on our results, the glutamines Gln123, Gln228, Gln266 and Gln267 were predicted as substrates for TG2, which is in accordance with the data reported in the literature showing that the TG2 uses the Gln266, Gln267 and Gln313 as substrates (Park et al. 2006). The Gln313 was missing from the crystal structure therefore it could not be examined. The Gln249, Gln268 and Gln271 may

also be considered as possible substrate sites as the presence of one amino acid with negative effect might be overcame by the presence of two or three residues with positive effect. In case of lysine donor substrates we could examine only three lysine residues, Lys87, Lys177 and Lys238 from which Lys238 appeared as possible amine donor substrate.

5.3.2. 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 (Esposito and Caputo 2005, Kitahara et al. 1987), hence the crystal structure of several neuropeptides was examined. The presence of the predictor amino acids in the surrounding of glutamine and lysine residues was used to predict possible TG2 modification sites. 40 glutamine and 39 lysine residues of randomly chosen 17 peptides were studied based on our prediction. Gln34 of neuropeptide Y; Lys4 of peptide YY; Lys27 of exendin 4; Gln12 and Gln8 of orexin B; Lys15 and Lys20 of pituitary adenylate cyclase activating polypeptide PACAP21; Gln83, Lys24, Lys58 and Lys80 of neurotrophin III; Gln11 and Lys35 of AGRP; and Gln68, Gln122, Lys38, Lys41, Lys115, Lys145, Lys168 and Lys172 of GRH appeared as possible substrate residues in the TG2 catalyzed reaction. To test whether some of these neuropeptides are indeed substrates for TG2, the neuropeptide Y, exendin 4 and orexin B were used in a transglutaminase assay followed by mass spectrometry analysis. The neuropeptide Y was modified by TG2 on Gln34, the orexin B on Gln8 and Gln12 and the exendin 4 on Lys12 and Lys27. The Lys12 of exendin 4 was used as substrate by TG2 although it does not contain any predictor amino acid in its

surrounding, a similar phenomenon to what was observed in case of Lys177 in IκB, Lys38 in thymosine beta 4, Lys63 in midkine, and Lys80 in alpha synuclein.

5.4. The role of intrinsic disorder in substrate recognition

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. These observations led us to the idea that some substrate residues are situated in flexible, intrinsically disordered regions of proteins. To test this hypothesis we searched the sequences of substrates for the presence of intrinsically disordered regions using IUPred (Dosztanyi et al. 2005) and PONDR-VSL2 (Peng et al. 2006) predictors. Table 2 summarizes the results of intrinsic disorder prediction for glutamine and lysine donor substrates and contains those proteins where both of the predictors indicated the presence of intrinsical disorder. The data suggest that the intrinsic disorder may have importance in substrate recognition in case of 33 glutamine donor proteins where either the whole protein was intrinsically unstructured (IUP) or the substrate region was situated in intrinsically disordered region (IDR) (Table 2). There were 31 proteins where the whole structure was ordered or the substrate residues were in ordered regions. A similar tendency could be observed in case of lysine donor substrate proteins (Table 2), where the substrate lysine was situated in intrinsically disordered region in 18 proteins, while in case of 28 proteins it was situated in ordered environment.

Whole protein disordered (IUP)	Whole protein ordered	Substrate region disordered (IDR)	Disorder is present but the substrate residues are not in disordered regions
Glutamine donor substrates			
1. Salivary acidic proline rich protein	1. Actin, cytoplasmic	1. Alpha-2 plasmin inhibitor*	1. Annexin -1
2. Alpha synuclein*	2. Aldolase A	2. C-CAM*	2. C1 inhibitor
3. Beta casein*	3. Alpha-lactalbumin	3. BetaB3 crystallin*	3. Collagen III
4. Cytochrome C	4. Amyloid bA4	4. BetaB2 crystallin*	4. IkB
5. Gliadin peptide	5. Beta-endorphin	5. BetaA3 crystallin*	5. IGFBP-1
6. HVC core protein	6. Carboxypeptidase B	6. BHMT*	6. MAGP-1
7. Myeline basic protein	7. CD38	7. Band 3 protein*	7. Midkine
8. Osteopontin*	8. Beta 2 microglobulin	8. Collagen VII	8. Nidogen
9. Thymosin beta4*	9. GST	9. Fibrinogen alpha*	9. Osteonectin
	10. Human Clara cell 10 kDa protein	10. Fibrinogen gamma*	10. Plasminogen
	11. Insulin	11. Fibronectin	11. S100A7
	12. Melittin	12. Galectin-3*	
	13. PLA2	13. Glucagon *	
	14. PAI2	14. Core histones*	
	15. S100A10	15. Hsp 27*	
	16. S100A11	16. LTBP-1	
	17. Statherin	17. RAP*	
	18. Substance P	18. RhoA*	
	19. Uteroglobin	19. SVS-IV	
	20. VIP	20. Synapsyn I*	
		21. Tau protein*	
		22. Troponin T*	
		23. Vimentin*	
		24. Vitronectin*	
Lysine donor substrates			
1. Acydic proline rich protein	1. Actin, cytoplasmic	1. Androgen receptor	1. BHMT
2. Alpha synuclein*	2. Alpha lactalbumin	2. AlphaB crystallin*	2. C-CAM
3. HCV core protein	3. Amyloid beta A4	3. Annexin 1*	3. Galectin-3
4. Osteopontin	4. Beta endorphin	4. BetaA3 crystallin*	4. GAPDH
5. Thymosin beta4*	5. Beta2 microglobulin	5. Fibrinogen alpha*	5. Midkine
	6. CD38	6. Fibrinogen gamma*	6. Myosine
	7. Calbindin	7. Histone H2B*	7. Parkin
	8. Carboxypeptidase B	8. Histone H4*	8. Plasminogen
	9. eIF5A*	9. Hsp 27*	9. S100A7
	10. GST	10. IkB*	10. Vimentin
	11. Human Clara cell 10 kDa protein	11. SVS-IV*	
	12. PLA2	12. Tau protein*	
	13. S100A10	13. Troponin I*	
	14. S100A11		
	15. Statherin		
	16. Ubiquitin		
	17. Uteroglobin		
	18. VIP		

Table 2. The intrinsic disorder content of TG2 substrate proteins. The underlined proteins have crystal

structures which contain the substrate residues. * indicates the substrate proteins which were used for the prediction of relative disorder and the relative number of disorder promoting amino acids (see Materials and Methods Intrinsic disorder prediction).

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 (see indicated proteins on Table 2), 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. When the relative disorder was higher than 0.5 the sequence was considered as intrinsically disordered. Using the IUPred predictor (Dosztanyi et al. 2005), in case of glutamine donor substrate proteins both the relative disorder and the relative number of disorder promoting amino acids was significantly higher in the surrounding of substrate glutamines than in the surrounding of non substrate glutamines (Figure 11).

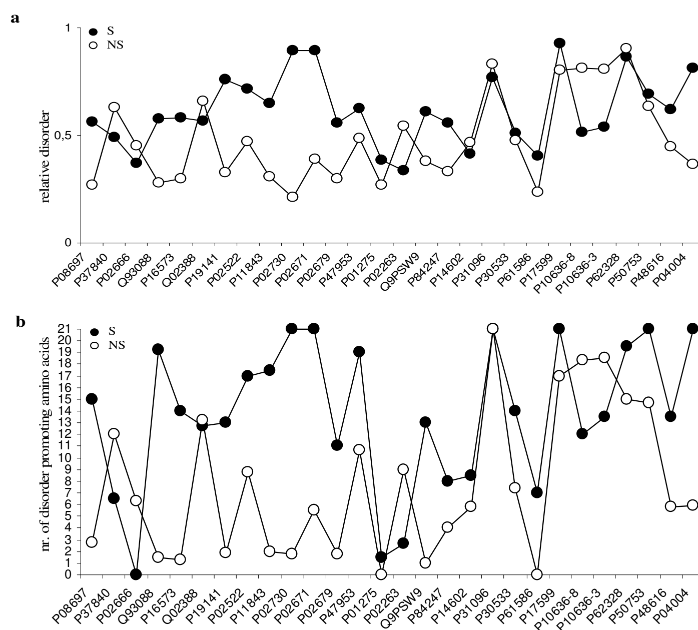


Figure 11. The importance of intrinsically disordered regions in TG2 glutamine substrate selection. The relative intrinsic disorder (Panel a) and the relative number of disorder promoting amino acids (Panel b) was determined with IUPred predictor in case of glutamine donor substrate proteins which contain both substrate and non-substrate residues (see for details Materials and Methods Intrinsic Disorder Prediction).

The proteins are annotated by their UniProt ID. The difference between substrate and non-substrate datasets was significant: for relative disorder $p < 0.005$ and for relative number of disorder promoting amino acids $p < 0.001$.

In case of lysine residues the significant differences between substrate and non substrate datasets were observed as well (Figure 12), regarding the relative disorder and the relative number of disorder promoting amino acids.

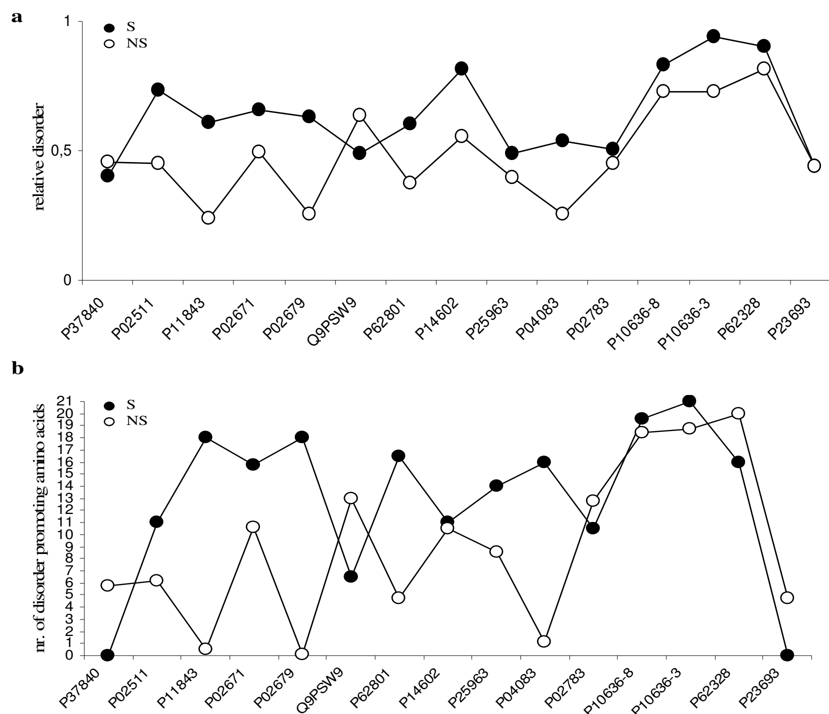


Figure 12. The importance of intrinsically disordered regions in TG2 lysine substrate selection. The relative intrinsic disorder (Panel a) and the relative number of disorder promoting amino acids (Panel b) was determined with IUPred predictor in case of lysine donor substrate proteins which contain both substrate and non-substrate residues (see for details Materials and Methods Intrinsic Disorder Prediction). The proteins are annotated by their UniProt ID. The difference between substrate and non-substrate datasets was significant: in both cases the $p < 0.001$.

When using the PONDR-VSL2 predictor (Peng et al. 2006) similar differences could be observed (data not shown) and the differences were significant in all cases.

6. DISCUSSION

Transglutaminase 2 is a multifunctional enzyme having more than 130 substrates at various locations inside and outside the cells. 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. These factors are numerous, like cell type- and tissue-dependent abundance and localization of the enzyme and its substrates, local availability of Ca^{2+}, the presence of modifying substances like sphyngosylphosphocholine (Lai et al.1997) and nitric oxide (Lai et al. 2001), the absence of inhibitors, and finally, the physical accessibility of modification sites on the individual molecules.

As a starting point for our experiments, the substrates of the TG2 were collected into a publicly available database, TRANSDAB Wiki (http://genomics.dote.hu/wiki) which provides the information about substrate proteins in an easy to find, interactive way making the browsing, information uploading and retrieving very easy.

6.1. 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 the amine donor substrate 5BPA. The peptides GQQQTPY, GLQQASV and WQTPMNS were modified most efficiently and the sequence motif pQx(P,T,S)l was

established. This motif was consistent with sequences in identified substrates listed in the TRANSDAB 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 (Sugimura et al. 2006).

To further investigate whether the technique has a potential to predict TG2-modification sites in native protein sequences, we chose one of the most efficient Q-donor peptides, GQQQTPY, as a representative example. A search in the human protein database revealed that several chromatin remodeling complex proteins and transcription factors contain stretches that are similar to GQQQTPY. The best match was found in the N-terminal glutamine-rich domain of SWI1/SNF1-related proteins (p270, BAF250a, p250, hOsa1), which are involved in unique gene regulatory activities in, i.e., myelocyte differentiation, and contain two conserved glutamine/proline-rich domains with yet unclear function (Dallas et al. 2000). Since existing recombinant versions of the above proteins do not contain the Q-P-rich regions, we decided to synthesize a 27-mer model peptide. Mass spectrometric analysis revealed that the 27-mer peptide could be transamidated by TG2 in vitro, and we identified three modified glutamines of the two GQQQTPY-like motifs (GQQGQTPY and QQQQPPY).

Remarkably, the above domain of SWI1/SNF1-like proteins resembles the sequence of the glutamine-rich plant storage proteins, glutenins and gliadins, and this similarity is manifested in a common spatial arrangement of Q, P, and aromatic amino acids (e.g., common QxPY motifs in SWI1/SNF1 vs. QxPF/PY in glutenins). Importantly, positioning of P relative to Q is an underlining condition in determining TG2 reactivity for gluten

peptides (Vader et al. 2002). Considering the efficient TG2-catalyzed deamidation reported for long gluten derived peptides, the above sequence relationship adds to the hypothesis that the Q-P-rich domains of SWI-SNF related proteins can be modified by the enzyme, although mass spectrometric analysis failed to detect deamidation products when the 27-mer peptide was treated with TG2 in the absence of amine donors

Beside the *in vitro* combinatorial approach, *in silico* method was used to investigate the preferred amino acid residues around substrate glutamines. Using SEQSTAT program, the sequence context of substrate glutamine containing sequences was compared to the sequence context of the non substrate glutamine containing sequences. The results clearly demonstrate the importance in substrate selection of adjacent glutamines, charged residues and the N- terminal position of substrate glutamines, as it was suggested by Aeschlimann et al. (Aeschlimann et al. 1992).

A similar comparison was done with the lysine containing sequences as well. We could demonstrate in our study the importance of C-terminal location of substrate lysine and the presence of proline and the positively charged arginine and lysine as favorable features in the surrounding of substrate lysines. 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. (Grootjans et al. 1995).

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 but none of these data could give a full explanation how TG2 glutamine and lysine sites are selected in substrate proteins.

6.2. Structural determinants of TG2 substrate specificity

The three dimensional structure of the substrate proteins was investigated and the large amount of data derived from the spatial environment of each glutamine and lysine residue of the crystallized substrate proteins could be analyzed with logistic regression. This method turned out to be suitable to identify predictors from a multitude of factors with possible importance in substrate selection of enzymes. In this particular case, we did not use the whole data coded in the 3D structure, only the number of amino acids at different distances from glutamine and lysine residues were evaluated, regardless of their spatial orientation and precise surface exposure. As it was expected, the amino acids in the spatial proximity of glutamine and lysine residues were not in sequential vicinity except for several small proteins with helical structure such as glucagon and thymosine. In this context previous data obtained from sequence studies can not be compared to our findings.

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 AUC 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.

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 (Hanson and Dallman 1995, Sakurai et al. 1998) and orexin B, acting on orexin receptors, takes part in modulation of wakefulness (Siegel 1999). 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 (Heine et al. 2005). 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.

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 (Tompa 2005, Uversky 2002). Intrinsic disorder has an important role in protein-protein interactions and

protein binding partner recognition (Tompa and Fuxreiter 2008) and is involved in posttranslational protein modifications including deacetylation and phosphorylation (Iakoucheva et al. 2004, Khan and Lewis 2005). It was reported that phosphorylation occurs predominantly within intrinsically unstructured regions of proteins (Iakoucheva et al. 2004) and the Hst2 enzyme, a member of the Sir2 family of NAD-dependent protein deacetylases, deacetylates the acetyl lysine residues situated in unstructured regions of proteins (Khan and Lewis 2005). Based on these observations we considered the intrinsic disorder as a possible factor influencing the substrate recognition of TG2 as well. The sequence analysis of TG2 substrate proteins revealed 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.

7. 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 was chosen for detailed analysis. Mass spectrometry-based studies of a representative part of the SWI1/SNF1-related chromatin remodeling protein p270 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, but 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.

8. 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 Andrejkovics, 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 colleagues 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.

9. REFERENCES

1. **Aeschlimann, D. & Thomazy, V.** (2000). Protein crosslinking in assembly and remodeling of extracellular matrices: The role of transglutaminases. Connect. Tissue Res. 41, 1–27.
2. **Aeschlimann, D., & Paulsson, M.** (1994). Transglutaminases: protein cross-linking enzymes in tissues and body fluids. Thromb. Hemost. 71, 402-415.
3. **Aeschlimann, D., Paulsson, M. & Mann, K.** (1992). Identification of Gln726 in Nidogen as the Amine Acceptor in Transglutaminase-catalyzed Cross-linking of Laminin-Nidogen Complexes. J. Biol. Chem. 267, 11316-11321.
4. **Ahmad, S., Gromiha, M., Fawareh, H. & Sarai, A.** (2004). ASAView, Solvent Accessibility Graphics for proteins. BMC Bioinformatic. 5, 51.
5. **Alford, S.C., Bazowski, J., Lorimer, H., Elowe, S., & Howard, P.L.** (2007). Transglutaminase 2 clusters soluble A-type ephrins into functionally active high molecular weight oligomers. Exp Cell Res. 313, 4170-4179.
6. **Altman, D.G.** (1992). Practical Statistics for Medical Research. pp. 351-388, Chapman&Hall.
7. **Ando, Y., Imamura, S., Owada, M.K. & Kannagi, R.** (1991). Calcium-induced intracellular cross-linking of lipocortin I by transglutaminase 2 in A431 cells. Augmentation by membrane phospholipids. J Biol Chem. 266, 1101-1108.
8. **Arentz-Hansen, H., McAdam, S.N., Molberg, O., Fleckenstein, B., Lundin, K.E.A., Jorgensen, T.J.D., Jung, G., Roepstorff, P. & Sollid, L.M.** (2002). Celiac lesion T cells recognize epitopes that cluster in regions of gliadin rich in proline residues. Gastroenterology 123, 803–809.

9. **Begg, G.E., Carrington, L., Stokes, P.H., Matthews, J.M., Wouters, M.A., Husain, A., Lorand, L., Iismaa, S.E., & Graham, R.M**. (2006). Mechanism of allosteric regulation of transglutaminase 2 by GTP. Proc Natl Acad Sci U S A. 103, 19683-19688.

10. **Cooper, A.J., Sheu, K.R., Burke, J.R., Onodera, O., Strittmatter, W.J., Roses, A.D. & Blass, J.P.** (1997). Transglutaminase-catalyzed inactivation of glyceraldehyde 3-phosphate dehydrogenase and alpha-ketoglutarate dehydrogenase complex by polyglutamine domains of pathological length. Proc Natl Acad Sci U S A. 94, 12604-12609.

11. **Cooper, A.J.L., Sheu, K-F.R., Burke, J.R., Onodera, O., Strittmatter, W.J., Roses, A.D. & Blass, J.** (1997). Polyglutamine Domains Are Substrates of Transglutaminase 2: Does Transglutaminase Play a Role in Expanded CAG/Poly-Q Neurodegenerative Diseases? J. Neurochem. 69, 431-434.

12. **Cordella-Miele, E., Miele, L. & Mukherjee, A.B.** (1990). A novel transglutaminase-mediated post-translational modification of phospholipase A2 dramatically increases its catalytic activity. J Biol Chem. 265, 17180–17188.

13. **Coussons, P.J., Price, N.C., Kelly, S.M., Smith, B. & Sawyer, L.** (1992). Factors that govern the specificity of transglutaminase-catalysed modification of proteins and peptides. Biochem. J. 282, 929-930.

14. **Curtis, C.G., Stenberg, P., Brown, K.L., Baron, A., Chen, K., Gray, A., Simpson, I., & Lorand, L.** (1974). Kinetics of transamidating enzymes. Production of thiol in the reactions of thiol esters with fibrinoligase. Biochemistry. 13, 3257-3262.

15. **Dallas, P.B., Pacchione, S., Wilsker, D., Bowrin, V., Kobayashi, R., & Moran, E.** (2000). The human SWI–SNF complex protein p270 is an ARID family member with non-sequence-specific DNA binding activity. Mol. Cell. Biol. 20, 3137–3146.

16. **Dosztányi, Z., Csizmok, V., Tompa, P. & Simon, I.** (2005). The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol. 347, 827-839.

17. **Esposito, C. & Caputo, I.** (2005). Mammalian transglutaminases: Identification of substrates as a key to physiological function and physiopathological relevance. FEBS J. 272, 615-631.

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

19. **Fésüs, L. & Szondy, Z.** (2005). Transglutaminase 2 in the balance of cell death and survival. FEBS Lett. 579, 3297-3302

20. **Fésüs, L., Thomazy, V., Autuori, F., Ceru, M.P., Tarcsa, E., & Piacentini, M.** (1989). Apoptotic hepatocytes become insoluble in detergents and chaotropic agents as a result of transglutaminase action. FEBS Lett. 245, 150-154.

21. **Folk, J.E. & Cole, P.W.** (1965). Structural requirements of specific substrates for guinea pig liver transglutaminase. J Biol Chem. 240, 2951–2960.

22. **Folk, J.E. & Finlayson, J.S.** (1977). The epsilon(gamma-glutamyl)lysine crosslink and the catalytic role of transglutaminases. Adv. Protein Chem. 31, 1-133.

23. **Folk, J.E.** (1983). Mechanism and basis for specificity of transglutaminase-catalyzed ε-(γ-glutamyl)lysine bond formation. Adv. Enzymol. 54, 1-56.

24. **Folk, J.E., & Cole, P.W.** (1965). Structural requirements of specific substrates for guinea pig liver transglutaminase. J Biol Chem. 240, 2951-2960.

25. **Gorman, J.J. & Folk, J.E.** (1981). Structural features of glutamine substrates for transglutaminases. Specificities of human plasma factor XIIIa and the guinea pig liver enzyme toward synthetic peptides. J Biol Chem. 256, 2712-2715.

26. **Gorman, J.J. & Folk, J.E.** (1984). Structural features of glutamine substrates for transglutaminases. Role of extended interactions in the specificity of human plasma factor XIIIa and of the guinea pig liver enzyme. J Biol Chem. 259, 9007-9010.

27. **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, 33066–33078.

28. **Griffin, M., Casadio, R. & Bergamini, C.M.** (2002). Transglutaminases: nature's biological glues. Biochem J. 368, 377-396.

29. **Groenen, P.J., Smulders, R.H., Peters, R.F., Grootjans, J.J., van den Ijssel, P.R., Bloemendal, H. & de Jong, W.W.** (1994). The amine-donor substrate specificity of tissue-type transglutaminase. Influence of amino acid residues flanking the amine-donor lysine residue. Eur J Biochem. 220, 795-799.

30. **Grootjans, J.J., Groenen, P.J.T.A. & de Jong, W.W.** (1995). Substrate Requirements for Transglutaminases: Influence of the amino acid preceding the amine donor lysine in a native protein. J. Biol. Chem. 270, 22855–22858.

31. **Han, J.A. & Park, S.C.** (2000). Transglutaminase-dependent modulation of transcription factor Sp1 activity. Mol Cells. 10, 612-618.

32. **Hanson, E.S. & Dallman, M.F.** (1995). Neuropeptide Y (NPY) may integrate responses of hypothalamic feeding systems and the hypothalamo-pituitary-adrenal axis. J Neurorndocrinol. 7, 273–279.

33. **Hasegawa, G., Suwa, M., Ichikawa, Y., Ohtsuka, T., Kumagai, S., Kikuchi, M., Sato, Y., & Saito, Y.** (2003). A novel function of tissue-type transglutaminase: protein disulphide isomerase. Biochem J. 373, 793-803.

34. **Hebert, S.S., Daviau, A., Grondin, G., Latreille, M., Aubin, R.A. & Blouin, R.** (2000). The mixed lineage kinase DLK is oligomerized by transglutaminase 2 during apoptosis. J Biol Chem. 275, 32482–32490.

35. **Heine, R.J., Van Gaal, L.F., Johns, D., Mihm, M.J., Widel, M.H., Brodows, R.G., GWAA Study Group.** (2005). Exenatide versus insulin glargine in patients with suboptimally controlled type 2 diabetes: a randomized trial. Ann Intern Med. 143, 559-569.

36. **Hevessy, Z., Patthy, A., Kárpáti, L. & Muszbek, L.** (2000). Alpha(2)-plasmin inhibitor is a substrate for transglutaminase 2: an in vitro study. Thromb Res. 99, 399-406.

37. **Humphrey, W., Dalke, A. & Schulten, K.** (1996). VMD – Visual Molecular Dynamics. J Molec Graphics. 14, 33-38.

38. **Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z. & Dunker, A.K.** (2004). The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. 32, 1037-1049.

39. **Ichikawa, A., Ishizaki, J., Morita, M., Tanaka, K., & Ikura, K.** (2008) Identification of new amine acceptor protein substrate candidates of transglutaminase in rat liver extract: use of 5-(biotinamido) pentylamine as a probe. Biosci Biotechnol Biochem. 72, 1056-1062.

40. **Ichikawa, A., Ohashi, Y., Terada, S., Natsuka, S., & Ikura, K.** (2004). In vitro modification of betaine-homocysteine S-methyltransferase by tissue-type transglutaminase. Int J Biochem Cell Biol. 36. 1981-1992.

41. **Iismaa, S.E.** (2002). Structure and function of transglutaminase 2. Minerva Biothec. 14, 113-119.

42. **Iismaa, S.E., Holman, S., Wouters, M.A., Lorand, L., Graham, R.M., & Husain, A.** (2003). Evolutionary specialization of a tryptophan indole group for transition-state stabilization by eukaryotic transglutaminases. Proc Natl Acad Sci U S A. 100, 12636-12641.

43. **Junn, E., Ronchetti, R.D., Quezado, M.M., Kim, S.Y., & Mouradian, M.M.** (2003). Transglutaminase 2-induced aggregation of alpha-synuclein: Implications for Lewy body formation in Parkinson's disease and dementia with Lewy bodies. Proc Natl Acad Sci U S A. 100, 2047-2052.

44. **Kaartinen, M.T., El-Maadawy, S., Räsänen, N.H. & McKee, M.D.** (2002). Transglutaminase 2 and its substrates in bone. J Bone Miner Res. 17, 2161-2173.

45. **Kaartinen, M.T., Sun, W., Kaipatur, N. & McKee, M.D.** (2005). Transglutaminase crosslinking of SIBLING proteins in teeth. J Dent Res. 84, 607-612.

46. **Khan, A.N. & Lewis, P.N.** (2005). Unstructured conformations are a substrate requirement for the Sir2 family of NAD-dependent protein deacetylases. J Biol Chem. 280, 36073-36078.

47. **Kim, J.H., Nam, K.H., Kwon, O-S., Kim, I.G., Bustin, M., Choy, H.E. & Park, S.C.** (2002). Histone cross-linking by transglutaminase. Biochem. Biophys. Res. Commun. 293, 1453-1457.

48. **Kim, S.Y., Marekov, L., Bubber, P., Browne, S.E., Stavrovskaya, I., Lee, J., Steinert, P.M., Blass, J.P., Beal, M.F., Gibson, G.E., & Cooper, A.J.** (2005). Mitochondrial aconitase is a transglutaminase 2 substrate: transglutamination is a probable mechanism contributing to high-molecular-weight aggregates of aconitase and loss of aconitase activity in Huntington disease brain. Neurochem Res. 30, 1245-1255.

49. **Kitahara, A., Mikawa, H., Ohtsuki, H., Yamagata, Y., Murachi, T. & Kannagi, R.** (1987). Specific localization of tissue-type transglutaminase in adrenocorticotropin-producing cells of the human pituitary gland as demonstrated by immunohistochemistry. J Clin Endocrinol Metab. 65, 885-890.
50. **Kojima, S., Inui, T., Muramatsu, H., Suzuki, Y., Kadomatsu, K., Yoshizawa, M., Hirose, S., Kimura, T., Sakakibara, S. & Muramatsu, T.** (1997). Dimerization of midkine by transglutaminase 2 and its functional implication. J Biol Chem. 272, 9410–9416.
51. **Lai, T.S., Bielawska, A., Peoples, K.A., Hannun, Y.A., & Greenberg, C.S.** (1997). Sphingosylphosphocholine reduces the calcium ion requirement for activating transglutaminase 2. J Biol Chem. 272, 16295-16300.
52. **Lai, T.S., Hausladen, A., Slaughter, T.F., Eu, J.P., Stamler, J.S., & Greenberg, C.S.** (2001). Calcium regulates S-nitrosylation, denitrosylation, and activity of transglutaminase 2. Biochemistry. 40, 4904-4910.
53. **Lee, K.N., Arnold, S.A., Birckbichler, P.J., Patterson, M.K. Jr, Fraij, B.M., Takeuchi, Y., & Carter, H.A.** (1993). Site-directed mutagenesis of human transglutaminase 2: Cys-277 is essential for transglutaminase activity but not for GTPase activity. Biochim Biophys Acta. 1202. 1-6.
54. **Lesort, M., Attanavanich, K., Zhang, J. & Johnson, G.V.** (1998). Distinct nuclear localization and activity of transglutaminase 2. J. Biol. Chem. 273, 11991–11994.
55. **Liu, S., Cerione, R.A., & Clardy, J.** (2002). Structural basis for the guanine nucleotide-binding activity of tissue transglutaminase and its regulation of transamidation activity. Proc Natl Acad Sci U S A. 99, 2743-2747.
56. **Lorand, L. & Graham, R.M.** (2003). Transglutaminases: crosslinking enzymes with pleiotropic functions. Nat. Rev. Mol. Cell Biol. 4, 140-156.

57. **Maccioni, R.B., & Arechaga, J.** (1986). Transglutaminase (TG) involvement in early embryogenesis. Exp Cell Res. 167, 266-270.

58. **Mandrusiak, L.M., Beitel, L.K., Wang, X., Scanlon, T.C., Chevalier-Larsen, E., Merry, D.E., & Trifiro, M.A.** (2003). Transglutaminase potentiates ligand-dependent proteasome dysfunction induced by polyglutamine-expanded androgen receptor. Hum Mol Genet. 12, 1497-1506.

59. **McDonough, J.L., Arrell, D.K., & Van Eyk, J.E.** (1999). Troponin I degradation and covalent complex formation accompanies myocardial ischemia/reperfusion injury. Circ Res. 84, 9-20.

60. **Mehta, K.** (2005). Mammalian transglutaminases: a family portrait. Prog Exp Tumor Res. 38, 1-18.

61. **Mishra, S., & Murphy, L.J.** (2004). Transglutaminase 2 has intrinsic kinase activity: identification of transglutaminase 2 as an insulin-like growth factor-binding protein-3 kinase. J Biol Chem. 279, 23863-23868.

62. **Murthy, S.N., Iismaa, S., Begg, G., Freymann, D.M., Graham, R.M., & Lorand, L.** (2002). Conserved tryptophan in the core domain of transglutaminase is essential for catalytic activity.Proc Natl Acad Sci U S A. 99, 2738-2742.

63. **Murthy, S.N., Wilson, J.H., Lukas, T.J., Kuret, J., & Lorand, L.** (1998). Cross-linking sites of the human tau protein, probed by reactions with human transglutaminase. J Neurochem. 71, 2607-2614.

64. **Nemes, Z., Devreese, B., Steinert, P.M., Van Beeumen, J. & Fésüs, L.** (2004). Cross-linking of ubiquitin, HSP27, parkin and a-synuclein by g-glutamyl-e-lysine bonds in Alzheimer's neurofibrillary tangles. FASEB J. 18, 1135-1137.

65. **Nemes, Z., Marekov, L.N., Fésüs, L., & Steinert, P.M.** (1999). A novel function for transglutaminase 1: attachment of long-chain omega-hydroxyceramides to involucrin by ester bond formation. Proc Natl Acad Sci U S A. 96, 8402-8407.

66. **Oliverio, S., Amendola, A., Di Sano, F., Farrace, M.G., Fésüs, L., Nemes, Z., Piredda, L., Spinedi, A. & Piacentini, M.** (1997). Transglutaminase 2-dependent posttranslational modification of the retinoblastoma gene product in promonocytic cells undergoing apoptosis. Mol Cell Biol. 10, 6040-6048.

67. **Overbye, A., Fengsrud, M., & Seglen, P. O.** (2007). Proteomic analysis of membrane-associated proteins from rat liver autophagosomes. Autophagy. 3, 300–322.

68. **Parameswaran, K.N., Cheng, X.F., Chen, E.C., Velasco, P.T., Wilson, J.H., & Lorand, L.** (1997). Hydrolysis of gamma:epsilon isopeptides by cytosolic transglutaminases and by coagulation factor XIIIa. J Biol Chem. 272, 10311-10317.

69. **Park, S.S., Kim, J.M., Kim, D.S., Kim, I.H. & Kim, S.Y.** (2006). Transglutaminase 2 mediates polymer formation of I-kappaBalpha through C-terminal glutamine cluster. J Biol Chem. 281, 34965-34972.

70. **Pastor, M.T., Diez, A., Perez-Paya, E. & Abad, C.** (1999). Addressing substrate glutamine requirements for transglutaminase 2 using substance P analogues. FEBS Lett. 451, 231-234.

71. **Peng, K., Radivojac, P., Vucetic, S., Dunker, A.K. & Obradovic, Z.** (2006). Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics. 7, 208.

72. **Pinkas, D.M., Strop, P., Brunger, A.T., & Khosla, C.** (2007). Transglutaminase 2 undergoes a large conformational change upon activation. PLoS Biol. 5, e327.

73. **Piredda, L., Amendola, A., Colizzi, V., Davies, P.J.A., Farrace, M.G., Fraziano, M., Gentile, V., Uray, I., Piacentini, M., & Fésüs, L.** (1997). Lack of transglutaminase 2 protein cross-linking leads to leakage of macromolecules from

dying cells, relationship to development of autoimmunity in MRLlpr/lpr mice. Cell Death Diff. 4, 463-472.

74. **Piredda, L., Farrace, M.G., Lo Bello, M., Malorni, W., Melino, G., Petruzzelli, R., & Piacentini, M.** (1999). Identification of 'tissue' transglutaminase binding proteins in neural cells committed to apoptosis. FASEB J. 13, 355-364.
75. **Rasmussen, L.K., Esben, S.S., Torben, E.P., Gliemann, J. & Jensen, P.H.** (1994). Identification of glutamine and lysine residues in Alzheimer amyloid b-A4 peptide responsible for transglutaminase-catalyzed homopolymerization and cross-linking to aMreceptor. FEBS Lett. 338, 161–166.
76. **Rodolfo, C., Mormone, E., Matarrese, P., Ciccosanti, F., Farrace, M.G., Garofano, E., Piredda, L., Fimia, G.M., Malorni, W., & Piacentini, M.** (2004). Transglutaminase 2 is a multifunctional BH3-only protein. J Biol Chem. 279, 54783-54792.
77. **Ruan, Q. & Johnson, G.V.** (2007).Transglutaminase 2 in neurodegenerative disorders. Front Biosci. 12, 891-904.
78. **Sakurai, T., Amemiya, A., Ishii, M., Matsuzaki, I., Chemelli, R.M., Tanaka, H., Williams, S.C., Richardson, J.A., Kozlowski, G.P., Wilson, S., Arch, J.R., Buckingham, R.E., Haynes, A.C., Carr, S.A., Annan, R.S., McNulty, D.E., Liu, W.S., Terrett, J.A., Elshourbagy, N.A., Bergsma, D.J. & Yanagisawa, M.** (1998). Orexins and orexin receptors: a family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. Cell. 92, 573-585.
79. **Siegel, J.M.** (1999). Narcolepsy: a key role for hypocretins (orexins). Cell. 98, 409-412.

80. **Singh, U.S., Kunar, M.T., Kao, Y-L. & Baker, K.M.** (2001). Role of transglutaminase II in retinoic acid-induced activation of RhoA-associated kinase-2. EMBO J. 20, 2413–2423.

81. **Sugimura, Y., Hosono, M., Wada, F., Yoshimura, T., Maki, M. & Hitomi, K.** (2006). Screening for the preferred substrate sequence of transglutaminase using a phage-displayed peptide library: identification of peptide substrates for TGASE 2 and Factor XIIIA. J Biol Chem. 281, 17699-17706.

82. **Tompa, P. & Fuxreiter, M.** (2008). Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. Trends Biochem Sci. 33, 2-8.

83. **Tompa, P.** (2005). The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett. 579, 3346-3354.

84. **Umar, S., Malavasi, F. & Mehta, K.** (1996). Post-translational modification of CD38 protein into a high molecular weight form alters its catalytic properties. J Biol Chem. 271, 15922–15927.

85. **Uversky, V.N.** (2002). Natively unfolded proteins: a point where biology waits for physics. Protein Sci. 11, 739-756.

86. **Vader, L.W., de Ru, A., van der Wal, Y., Kooy, Y.M.C., Benckhuijsen, W., Meqarin, M.L., Drijfhout, J.W., van Veelen, P., & Koning, F.** (2002). Specificity of transglutaminase 2 explains cereal toxicity in Celiac disease. J. Exp. Med. 195, 643–649.

87. **Williams-Ashman, H.G.** (1984). Transglutaminases and the clotting of mammalian seminal fluids. Mol Cell Biochem. 58, 51–61.