NBCZone: Universal three-dimensional construction of eleven amino acids near the catalytic nucleophile and base in the superfamily of (chymo)trypsin-like serine fold proteases

Denesyuk, Alexander I.; Johnson, Mark S.; Salo-Ahen, Outi M.H.; Uversky, Vladimir N.; Denessiouk, Konstantin

Published in:
International Journal of Biological Macromolecules

DOI:
10.1016/j.ijbiomac.2020.03.025

Published: 15/06/2020

Document Version
Accepted author manuscript

Document License
CC BY-NC-ND

This is the post-review accepted manuscript. The final version has been published in:

International Journal of Biological Macromolecules

Volume 153, 15 June 2020, Pages 399-411; https://doi.org/10.1016/j.ijbiomac.2020.03.025

NBCZone: Universal three-dimensional construction of eleven amino acids near the catalytic nucleophile and base in the superfamily of (chymo)trypsin-like serine fold proteases

Alexander I. Denesyuk,[1,2,*] Mark S. Johnson,[2] Outi M.H. Salo-Ahen,[2,3]

Vladimir N. Uversky,[1,4,*] and Konstantin Denessiouk[2,3]

[1]Institute for Biological Instrumentation of the Russian Academy of Sciences, Federal Research Center "Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences", 142290 Pushchino, Russia.

[2]Structural Bioinformatics Laboratory, Biochemistry, Faculty of Science and Engineering, Åbo Akademi University, 20520 Turku, Finland.

[3]Pharmaceutical Sciences Laboratory, Pharmacy, Faculty of Science and Engineering, Åbo Akademi University, 20520 Turku, Finland.

[4]Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA.

* Authors to whom correspondence should be addressed: AID – Structural Bioinformatics Laboratory, Biochemistry, Faculty of Science and Engineering, Åbo Akademi University, Tykistökatu 6, BioCity3A, 20520 Turku, Finland; E-mail: adenesyu@abo.fi; VNU - Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, 12901 Bruce B. Downs Blvd., MDC 07, Tampa, FL 33612, USA; E-mail: vuversky@healh.usf.edu

Abstract

(Chymo)trypsin-like serine fold proteases belong to the serine/cysteine proteases found in eukaryotes, prokaryotes, and viruses. Their catalytic activity is carried out using a triad of amino acids consisting of a nucleophile, a base, and an acid. For this superfamily of proteases, we propose the existence of a universal 3D structure comprising 11 amino acids near the catalytic nucleophile and base – Nucleophile-Base Catalytic Zone (NBCZone). The comparison of NBCZones among 169 eukaryotic, prokaryotic, and viral (chymo)trypsin-like proteases suggested the existence of 15 distinct groups determined by the combination of amino acids located at two "key" structure-functional positions 54_T and 55_T near the catalytic base $His57_T$. Most eukaryotic and prokaryotic proteases fell into two major groups, [ST]A and TN. Usually, proteases of [ST]A group contain a disulfide bond between cysteines $Cys42_T$ and $Cys58_T$ of the NBCZone. In contrast, viral proteases were distributed among seven groups, and lack this disulfide bond. Furthermore, only the [ST]A group of eukaryotic proteases contains glycine at position 43_T, which is instrumental for activation of these enzymes. In contrast, due to the side chains of residues at position 43_T prokaryotic and viral proteases do not have the ability to carry out the structural transition of the eukaryotic zymogen-zyme type.

Keywords: (Chymo)trypsin-like proteases; Catalytic triad; Structural motif; Structural framework

Introduction

Previously, when analyzing the spatial structures of a superfamily of proteins with an α/β-hydrolases fold (SCOP ID: [53473]), characterized by the 3-layer α/β/α architecture with a mixed β-sheet of eight β-strands placed in 12435678 order, and with β-strand 2 being antiparallel to the rest of the β-strands [1], the existence of a small internal position with variable contents was discovered. This position is filled with a water molecule or with an oxygen atom from the side-chain group of the catalytic acid residue. The set of amino acids surrounding this position was termed the catalytic acid zone [2].

In addition to the acid residue, the active site of proteins with an α/β hydrolase fold includes a nucleophile, base, and two residues of the oxyanion hole that stabilize the tetrahedral intermediate during catalysis [3]. Together, these residues form the catalytic machinery necessary for performing the hydrolase function. Similarly to the catalytic acid zone, the nucleophile and oxyanion zones, which co-ordinate the catalytic nucleophile and the residues of the oxyanion hole, were also described [4]; and it was speculated that the catalytic triad zones together form a conserved structural motif [2].

There is another superfamily of hydrolases, (chymo)trypsin-like serine fold proteases (SCOP ID: [50493]) [5]. These are the all-β proteins comprised of two six-stranded Greek key β-barrels lying perpendicular to one another with the active site cleft located between them [1]. The hydrolases of this type also have the same catalytic triad [5-7], but the three amino acids are arranged in a different sequential order (**Fig. 1**). The nucleophile of α/β hydrolases is located at the turn, known as the "nucleophile elbow", and is identified by the sequence motif Sm–X–Nuc–X–Sm (Sm, small residue; X, any residue; Nuc, nucleophile) [3]. The corresponding pentapeptide of (chymo)trypsin-like serine fold proteases has the consensus pattern G–[DE]–S–G–[GS] (https://prosite.expasy.org/; PROSITE documentation PDOC00124; TRYPSIN_SER, PS00135 [8]). In addition to this pattern, there is also a histidine active site pattern [LIVM]-[ST]-A-[STAG]-H-C (TRYPSIN_HIS, PS00134). In both families, the oxyanion hole is situated adjacent to the nucleophile and is mainly shaped by the main-chain nitrogen atoms of two residues, but again the sequential order is different.

Knowing the importance of the existence of a catalytic acid, nucleophile, and oxyanion zones for the function of the α/β hydrolases fold enzymes and with the possible presence of similar structural formations in other types of hydrolases, we carried out a detailed analysis of the spatial structures of (chymo)trypsin-like serine fold proteases near the catalytic triad residues.

As a result, fifteen variants of a unique structural Nucleophile-Base Catalytic Zone (NBCZone) were found that affect the conformations of the catalytic triad residue loops of all (chymo)trypsin-like serine fold proteases (eukaryotic, prokaryotic, and viral).

Results and Discussion

Eukaryotic serine (chymo)trypsin-like fold proteases

Nucleophile-Base Catalytic Zone (NBCZone) of trypsin

The presentation of our results will begin with an analysis of the tertiary structure of bovine trypsin active site (Protein Data Bank (PDB: [9]), PDB ID: 4I8H, chain A, **Fig. 1**) [10]. **Fig. 1** shows the catalytic triad that includes residues His57, Asp102 and Ser195. In addition to the base His57 and the nucleophile Ser195, the localization of Ser54, Ala55, and Cys58 of the PROSITE TRYPSIN_HIS pattern is shown, as well as amino acids of the PROSITE TRYPSIN_SER pattern: Gly193-Asp194-Ser195-Gly196-Gly197 and two main-chain nitrogens: N/Gly193 and N/Ser195, which are the canonical oxyanions OxyI and OxyII [11]. In trypsin, there are two amino acids that interact with both the PROSITE TRYPSIN_HIS pattern and the PROSITE TRYPSIN_SER pattern. The amino acids Gly43 and Val213 are in contact with the tripeptide Ser195-Gly196-Gly197 (**Fig. 2A**, **Table 1**). Main-chain atoms of Gly43 form a hydrogen bond with Ser195: N/Gly43-O/Ser195, and a weak hydrogen bond with Gly196: O/Gly43-CA/Gly196. Similar to Gly43, main-chain atoms of Val213 form two hydrogen bonds with the dipeptide Gly196-Gly197 (N/Val213-O/Gly197 and O/Val213-N/Gly196); the tetrapeptide Asp194-Gly197 forms a β-turn (data not shown). Main-chain atoms of Gly43 and Val213 are also in contact with the dipeptide Ser54-Ala55 from the PROSITE TRYPSIN_HIS pattern: contacts O/Gly43-OG/Ser54 and O/Val213-CB/Ala55 are shown (**Fig. 2A**). The group of connected amino acids presented in **Fig. 2A** we have named the 43/213 Nucleophile-Base Catalytic Zone (43/213 NBCZone) of trypsin.

Unlike the 43/213 NBCZone, which incorporates only two amino acids of the PROSITE TRYPSIN_HIS pattern, Ser54 and Ala55, the entire PROSITE TRYPSIN_HIS pattern is included into the neighboring 42/43 Base Catalytic Zone, which contains the pentapeptide Ser54-Ala55-Ala56-His57-Cys58 and the dipeptide Cys42-Gly43 (**Fig. 2B**). Cys58 and Cys42 of this zone are linked by a disulfide bond, which maintains the conformation of the polypeptide chain near the active site. In addition, these two cysteines have contacts with the

main-chain oxygen of the catalytic nucleophile (**Table 1**). In the remaining part of this article, the 43/213 Nucleophile-Base Catalytic Zone together with the 42/43 Base Catalytic Zone will be joined together and called the Nucleophile-Base Catalytic Zone (NBCZone) of trypsin. It is important to note that the side chain of the catalytic nucleophile Ser195, the side chain of the catalytic base His57 and the entire catalytic acid Asp102 are not part the NBCZone of trypsin.

The structure of trypsin may be either in active form [10] or inactive (zymogen) form [12] prior to proteolytic activation. Another name for the enzyme – trypsinogen – corresponds to the inactive form. However, in other proteases, the inactive form of the protein tertiary structure can be either similar or different from the zymogenic form of trypsin. Thus, in order to simplify the writing of the text of the article, we will use the structural term "zymogen" to describe only the zymogenic inactive form of trypsin and other proteases, while we will use the term "zyme" to collectively describe the remaining structural forms of proteases, which will include both non-zymogenic inactive forms ($zyme_{inact}$) and active forms ($zyme_{act}$) of the enzymes. In trypsin (PDB ID: 4I8H), the non-zymogenic form of the tertiary structure of the protein corresponds to $zyme_{act}$.

Structural comparison shows that the NBCZones of bovine trypsin ($zyme_{act}$; PDB ID: 4I8H) and trypsinogen (zymogen; PDB ID: 1TGT) are the same (**Table 1**). Indeed, the main structural differences between trypsin and trypsinogen lie in the activation domain [13], while the eleven amino acids of trypsin NBCZone lie outside of it.

[ST]A group

In the SCOP database, (chymo)trypsin-like fold serine proteins are divided into 4 families: eukaryotic, prokaryotic, viral serine, and viral cysteine proteases [1]. Looking at the tertiary structures among the eukaryotic (chymo)trypsin-like serine fold proteases in the PDB, SCOP, CATH [14], and MEROPS [15] databases we found, besides trypsin, another 14 proteins that have both zyme (either $zyme_{inact}$ or $zyme_{act}$) and zymogen forms of the three-dimensional structures, 49 proteases without the zymogen form and three with only the zymogen form, totaling to 67 tertiary structures (**Tables 1**, **S1** and **S2**). In this work, for each protease, we will use both the original numbering of the amino acid sequence and the canonical numbering of the amino acid sequence of trypsin which is shown by the index "T". The 67 eukaryotic

structures have either serine or threonine at position 54_T, and alanine at position 55_T. Therefore, the set of these 67 proteins was named the "[ST]A group" (**Table S1**).

The importance of alanine at position 55_T for the catalytic activity of chymotrypsin-like serine proteases has been established [16, 17]. Using the human plasmin model as an example, it was shown that replacing the alanine residue with a threonine leads to the formation of an unusual hydrogen bond between this threonine and the catalytic histidine [16]. The peculiarity of this interaction is that the catalytic histidine now adopts an inactive conformation. In contrast, with the bovine protein C as an example, replacing this alanine with a hydrophobic valine does not cause major changes in the conformation of the catalytic histidine [17].

The conservation of amino acids observed by us at the positions 54_T and 55_T fully complies with the definition of the PROSITE TRYPSIN_HIS pattern. Glycines at position 196_T and 197_T are also invariant. The same can be said about 42_T-58_T disulfide bond with the exception of four proteins: C1r, C1s, MASP-2 and MASP-3 (**Table S1**). C1r/C1s and MASP-1/-2/-3 form a family of mosaic serine proteases with identical domain organization [18], functioning as supramolecular complexes [19]. In each of the five mosaic serine proteases, two loops structurally corresponding to trypsin loop A (34_T-41_T) and loop B (56_T-64_T) [20] contact each other during the formation of such complexes [21-23]. In trypsin, cysteine at position 42_T is located at the carboxy-terminal end of loop A, and cysteine at position 58_T is placed at the amino-terminal end of loop B. Perhaps because of the need to form the supramolecular complexes, disulfide bonds are missing in all but MASP-1. As the analysis of the MASP-2 structure (PDB ID: 3TVJ, [22]) shows, despite the absence of a disulfide bond, contacts between amino acids at positions 42_T, 58_T and the catalytic nucleophile are conserved (**Table 1**). This indicates the existence of the NBCZones in these four proteins. In MASP-1 (PDB ID: 3GOV, [24]), unlike the other family members, there is a Cys475(42_T)-Cys491(58_T) disulfide bond. In addition, this protease has a very long loop B (Ala488-Asp513), compared with the corresponding loops in other proteases of this family [24]. It is possible that the structural and functional features of loop B, together with loop A and the amino acids adjacent to them, determine the presence of the Cys475(42_T)-Cys491(58_T) disulfide bond in MASP-1.

TN group

Although the probability is very high for alanine to be at position 55_T of eukaryotic (chymo)trypsin-like serine fold proteases, the residue is not absolutely conserved. Asparagine is observed at position 55_T in *Homo sapiens* HtrA and *Arabidopsis thaliana* Do-like serine proteases (8 proteases, **Tables 1** and **S1**), joined together into HtrA family [25, 26]. In these enzymes the 54_T position is occupied by only threonine, and thus, this group of HtrA family proteases was named the "TN group". Because in the TN group the 55_T position is occupied by an asparagine and not by an alanine like in the [ST]A group, the 55_T-213_T interaction (**Fig. 2A**) is modified such that instead of the CB atom of alanine, the main chain oxygen of amino acid at position 213_T interacts with the ND2 atom of asparagine (e.g. ND2/Asn196–O/Asn321 in human mitochondrial serine protease HtrA2 (PDB ID: 5M3N, [27]; **Table 1**, **Fig. 3A**).

The catalytic triad of HtrA2 is found in a catalytically incompetent conformation [27]. The distance between the ND1/His198 (base) and OG/Ser306 (nucleophile) atoms is 6.2 Å. There are also no hydrogen bonds between His198 (catalytic base) and Asp228 (catalytic acid). Asn196 is, however, directly involved in the mutual separation of the base and the nucleophile from each other, forming two hydrogen bonds: ND2/Asn196-OG/Ser306=2.8 Å and OD1/Asn196-N/His198=3.1 Å (**Fig. 3A**). In particular, the tripeptide Asn196-Ala197-His198 forms an Asx-turn [28].

The more complicated networks of interactions within the NBCZones are observed in the case of human HtrA1 (PDB ID: 3TJN, [29]). The PDB file 3TJN, 3 Å resolution, contains coordinates for the A, B, and D chains; chain D is relatively poorly ordered overall. Chain A has an incompetent conformation of active site: ND1/His220-OG/Ser328=7.6 Å, that is essentially the same as seen for the HtrA2 structure. The active site residues of chain B: Ser328 (nucleophile), His220 (base) and Asp250 (acid) are properly positioned for catalytic activity (**Fig. 3B**, **Table 1**). As in HtrA2, the contacts O/Asn343-ND2/Asn218 are present in chains A and B of HtrA1; however, atom OD1/Asn218, instead of forming the Asx-turn OD1/Asn218-N/His220 as in HtrA2, is now involved in the contacts with a catalytic acid Asp250.

There is no essential difference between the NBCZones of HtrA3 and HtrA2 that belong to set I (**Table S1**, column 4). Further analysis of prokaryotic and viral proteases will show that all these proteases of set I have an incompetent conformation of catalytic histidine, and as the result, the tripeptide $Asn55_T$-$Xaa56_T$-$His57_T$ forms an Asx-turn. Thus, the analysis of the structures of HtrA1, HtrA2, and HtrA3 proteases demonstrates that $Asn55_T$ is characterized

by large conformational differences between the incompetent and competent conformations for substrate binding active site regions.

Although the HtrA and Do-like proteases are within the same TN group, there are some structural differences in their NBCZones. For example, in the proteases HtrA1 and Do-like 5 (set II, **Table S1**) atom ND2/Asn55$_T$ forms the hydrogen bond in the NBCZone and atom OD1/Asn55$_T$ interacts with the catalytic acid Asp102$_T$ (**Fig. 3B**). However, in protease Do-like 1, atom OD1/Asn171 forms the hydrogen bond in the NBCZone and atom ND2/Asn171 plays a key role in the interactions with the catalytic acid Asp102$_T$ (PDB ID: 3QO6 [30], **Fig. 3C**, **Table 1**). Similar NBCZones occur in the proteases Do-like 2, Do-like 8 and Do-like 9: set III (**Table S1**). Further analysis of prokaryotic and viral proteases also show that all proteases of set II have the catalytically competent or incompetent conformation of catalytic histidine, but all proteases of set III have only the catalytically competent conformation of the catalytic histidine (**Table S1**, column 4). The structural diversity of the side chain of Asn55$_T$ agrees well with the conformational changes in the active sites of the HtrA family proteases [25-27, 29].

The HtrA family proteases are multidomain proteins, which besides a proteolytic domain also contain at least one C-terminal PDZ domain [25, 26]. The functional unit of the HtrA family proteases ranges from a trimer to a dodecamer. Loops A and B play important structural and regulatory roles in the HtrA multimer complexes [25, 31]. It is possible that for the implementation of these functions, loops A and B require a certain mobility. The observed presence of asparagine at position 55$_T$, the lack of the disulfide bond Cys42$_T$-Cys58$_T$ and the substitution of the cysteine at position 42$_T$ for the small amino acid glycine (**Table S1**) contribute to this requirement.

Prokaryotic serine (chymo)trypsin-like fold proteases

TA and TN groups

As in the case of eukaryotic (chymo)trypsin-like serine fold proteases, most prokaryotic proteases fall into the TA and TN groups (**Tables 1** and **S1**). Only now the TA and TN groups are almost equal in terms of the number of the available tertiary structures that have fallen into them. The presence (TA group) or absence (TN group) of a disulfide bond Cys42$_T$-Cys58$_T$ is also similar in prokaryotic and eukaryotic proteases. Furthermore, when the

disulfide bond is missing in the TN group, then there is a small amino acid – either glycine or alanine – at position 42_T of the prokaryotic proteases, as is the case with the eukaryotic proteases.

Most prokaryotic proteases of the TN group fall into set III. In **Table S1**, there are also examples of tertiary structures included in set I and set II, and even one structure: *Staphylococcus aureus* serine protease SplE, PDB ID: 5MM8 [32], in which Asn37 has two conformations corresponding to sets I and III.

Prokaryotic proteases of the TN group, without the $Cys42_T$-$Cys58_T$ disulfide bond, belong mostly to three subgroups: the HtrA family, Spl proteases, and exfoliative toxins (**Table S1**). **Fig. 3D** shows structural details of the NBCZone (TN group) for serine protease Spl from *Staphylococcus aureus*, PDB ID: 2AS9 [33]. The Spl protease is not in an active form due to a rotation of the side chain of catalytic His40, and its NBCZone is not similar to other NBCZones formed with the participation of asparagine at position 55_T. The tripeptide Asn38-Lys39-His40 forms a modified Asx-turn. Therefore, the structure of this protease belongs to a separate set IV.

The structural significance of the amino acid at position 55_T for the stabilization of the catalytic triad of prokaryotic (chymo)trypsin-like serine fold proteases is analyzed in detail in several publications [34, 35]. Using V8 protease and glutamyl-endopeptidase as examples, it was assumed that accommodation of an asparagine instead of alanine in position 55_T is impossible without some rearrangement of interactions between the catalytic histidine and the catalytic acid. In particular, weakening of the interactions of the catalytic acid $D102_T$ with the amides of residues 56_T and 57_T were observed. In addition, it was predicted that in glutamate-specific endopeptidase from *Bacillus subtilis*, the replacement of the conserved $Gly193_T$ (**Fig. 1**) with a cysteine could lead to the formation of a new disulfide bond that stabilizes the conformation of the oxyanion hole.

43&[STG][AV] group

Another noticeable difference between the NBCZones of eukaryotic and prokaryotic proteases is the appearance of the 43&[STG]V group in prokaryotic proteins (**Tables 1** and **S1**). These proteases belong to the SPATE family [36, 37]. The presence of number 43 in the group name is due to the fact that the six proteins of this group have the amino acid main

chain conformation at position 43_T different compared to the amino acid main chain conformation at position 43_T of all proteins analyzed so far. As an example, in the *Haemophilus influenzae* immunoglobulin A1 protease (**Fig. 4**, PDB ID: 3H09, [38]), the contact with O/Ser288 is formed not by the N/Ile86 atom, but by CG2/Ile86. A change in the course of the polypeptide chain at position 43_T leads to a complete impossibility of forming the $Cys42_T$-$Cys58_T$ disulfide bond. Nevertheless, the contact of atom CD1/Ile86(43_T) with atom CG1/Val101(58_T) indicates the presence of the NBCZone (**Table 1**).

It is possible that the formation of a truncated loop A (34_T-41_T) is directly related to the specificity of the catalytic activity of SPATE family proteases. Instead, a conservative tyrosine (Tyr239 in immunoglobulin A1 protease) of the unique functional loop D (143_T-149_T) is located in place of the loop A bend of the polypeptide chain of these proteases [38].

Another structural feature observed in the 43&[STG]V group is the presence of valine at position 55_T (Val98 in immunoglobulin A1 protease). Interesting amino acid variability is also observed at position 54_T, with serine or threonine found in five proteins and glycine in one protein. In the immunoglobulin A1 protease that has glycine at position 54_T, instead of the OG atom of the side chain of $Ser54_T$, there is a water molecule HOH1135, which completely replaces the OG atom in the construction of the NBCZone (**Fig. 4**). The possible structural and functional role of the existence of a water molecule near the amino acid at position 54_T is discussed below using the viral cysteine proteases as examples.

Viral serine (chymo)trypsin-like fold proteases

TA, [ST]Ψ and [KR]P groups

SCOP divides viral (chymo)trypsin-like fold proteases into two superfamilies: serine and cysteine proteases [1]. None of the viral proteases in **Table S1** have a $Cys42_T$-$Cys58_T$ disulfide bond. In turn, viral serine (chymo)trypsin-like fold proteases are divided into three groups: TA, [ST]Ψ and [KR]P (Ψ – amino acids with large aliphatic side chains: V, I, L; [39]) (**Tables 1** and **S1**). In all proteases that fall into these three groups, the amino acids at position 55_T do not show new structural features in the formation of the NBCZone compared to Val98 of the immunoglobulin A1 protease (**Table 1** and **Fig. 4**). Instead of cysteine, glycine is located at position 42_T, with the exception of phenylalanine in the HCV NS3 protease (**Table S1**). Since the [KR]P group has lysine or arginine instead of threonine or serine at position

54_T, as pointed out above, only the corresponding representative structure of this group (Sindbis virus capsid protein, PDB ID: 1SVP, [40]) is presented in **Table 1** and **Fig. 5**. The inclusion of lysine and proline at positions 54_T and 55_T, respectively, does not cause any major steric problems in the construction of the NBCZone.

Substitutions of Thr54 and Val55 in the HCV NS3 protease/helicase ([ST]Ψ group) affect the level of drug resistance of this virus [41-44]. The Thr54Ala mutation changes the type of the hydrogen bond with Leu44(43_T), as a result of which the conformations of amino acids Leu44(43_T) and Phe43(42_T) may change, and thus the protease/helicase binding to the inhibitor is weakened. The Thr54/Ser mutation was associated with medium level drug resistance. The Val55Ala, Arg155(214_T)Lys/Thr and Ala156(215_T)Thr resistant variants have been also identified.

Viral cysteine (chymo)trypsin-like fold proteases

Viral cysteine (chymo)trypsin-like fold proteases are divided into four groups: [TA]N, T[TSA], [ΨC][PQ], and 43&[VR]N (**Tables 1** and **S1**, [45-48]). Perhaps, to accommodate the cysteine nucleophile, the active site of the viral cysteine (chymo)trypsin-like fold proteases is larger than that of the serine proteases [49]. This structural result is consistent with the observation that the contact between the amino acid at position 58_T and the catalytic cysteine has disappeared (**Table 1**).

The NBCZone of the nuclear inclusion protein A from *Tobacco vein mottling virus* ([TA]N group) demonstrates a previously found feature in immunoglobulin A1 protease (**Table S1**, 43&[STG][AV] group): the amino acid Ala43 at position 54_T for contact with amino acid Phe33 at position 43_T uses a water molecule HOH246 as an intermediary. Perhaps this presence of a water molecule is caused by the existence of large hydrophobic Leu32 and Leu47 instead of cysteines at the positions 42_T and 58_T (**Table S1**). The largest number of these viral protease structures belong to the [ΨC][PQ] group.

There are two more noticeable differences between the NBCZones of serine and cysteine proteases. The first difference is the presence of glutamate instead of aspartate (catalytic acid) at position 102_T of the 3C and 3C-like proteases (T[TSA] and [ΨC][PQ] groups). The second difference is that half of the proteases from the [ΨC][PQ] group do not have a catalytic acid at position 102_T at all; i.e., they have a catalytic dyad in the active site instead of the catalytic

triad. However, all these dyad proteases, instead of the missing catalytic acid, have a water molecule (**Table S1**, column 10), which forms several hydrogen bonds with the residues surrounding it, including the catalytic histidine [50].

With one exception, the group of proteases with a catalytic dyad have cysteine and proline at positions 54_T and 55_T, respectively (**Table S1**). The alignment of the primary structures of such proteases, given in the work of Kanitz *et al.* (2019), shows that the number of dyad proteases with the dipeptide Ile-Gln at positions 54_T and 55_T is approximately equal to the number of dyad proteases with the dipeptide Cys-Pro [47]. **Table 1** lists the structural parameters of the 3Cl protease from *Alphamesonivirus 1* with the participation of the dipeptide Ile-Gln in the active site (PDB ID: 5LAC, [47]). Gln46, located in position 55_T, is the largest amino acid of all structurally similar amino acids listed in **Table S1**. Therefore, despite the large size of Ile45 located at position 54_T, Ile45 required a water molecule HOH537 to contact with Arg35 (position 43_T) (**Table 1**). The existence of water HOH537 correlates with the presence of two leucine residues at positions 42_T and 58_T, as it was shown earlier for the nuclear inclusion protein A. It is possible that the presence of a water molecule (**Table S1**, column 7) in the 2A proteinase from the 43&[VR]N group is also associated with the existence of leucine at position 58_T and structural specificity of 43&[VR]N group. The connection between the structural specificity of a group and the presence of a water molecule is also supported by the existence of a similar water molecule in immunoglobulin A1 protease (43&[STG][AV] group), which was described earlier. However, the structural reasons for the presence of the water near the residues at positions 54_T and 43_T in these six proteins are not entirely clear. It can only be assumed that the replacement of a direct contact of amino acids at positions 54_T and 43_T with water-mediated contacts demonstrates a weakening of the interaction of the nucleophilic loop with a β-sheet containing the catalytic base and the catalytic acid. This weakening is somehow related to the functional characteristics of these six proteins. In conclusion, we note that the presence of water molecules in a similar place in (chymo)trypsin-like serine fold proteases has not been previously established.

In the [TA]N and 43&[VR]N groups of cysteine proteases, rotation of the side chain of asparagine at position 55_T is observed, as has already been noted for the representative of the TN group of eukaryotic and prokaryotic serine proteases.

Unlike the HCV NS3 protease/helicase, the Thr27(54_T)Ala/Ser/Val mutations of 3C-like norovirus protease (T[TSA] group) do not affect the catalytic activity of this protease. Rather,

it was suggested that Thr27 is involved in stabilizing the conformation [51]. Mutation Leu19(58_T)Ser in the HRV2 2A proteinase (43&[VR]N group) leads to a similar result [52]. However, the Asn16(55_T)Ala mutation inhibits proteolytic activity completely. Mutations of residues around the nucleophilic cysteine: Pro103(192_T)Gly and Asp105(194_T)Thr/Asn, also impair the proteinase activity.

Inactive (chymo)trypsin-like fold proteases

In addition to 161 protease structures, 8 proteins with the (chymo)trypsin-like fold were found that are not proteases (**Table S1**). These 8 proteins do not have a catalytic nucleophile, and five of them do not have a catalytic base either. Seven structures belong to the eukaryotic proteins (TA and T[TG] groups) and one structure (TT group) is a prokaryotic protein. Six out of the seven eukaryotic proteins have a $Cys42_T$-$Cys58_T$ disulfide bond.

The active site of *Holotrichia diomphalia* prophenoloxidase activating factor-II demonstrates the zymogenic conformation (PDB ID: 2B9L, [53]). In addition, due to the lack of a side chain on glycine 55_T (Gly198), the contact O/Val374-CA/Gly198=5.4 (4.5) Å is weak **(Table 1**). However, in this protein, the catalytic serine 195_T is also replaced by glycine (Gly353). Perhaps for this reason, the CE1 atom of the catalytic histidine 57_T (His200) forms the contact O/Val374-CE1/His200=3.5 (2.4) Å. According to the canonical rule of the Derewenda *et al*. [54], the CE1 atom of a catalytic histidine should form a weak hydrogen bond with the main-chain oxygen of Ala375 (O/Ala375-CE1/His200, 3.1 (3.0) Å), the residue following Val374 in the amino acid sequence. Therefore, for the main-chain oxygen of the amino acid at position 213_T (Val374), a new structure-catalytic role is discovered as a fixator of a catalytic histidine. Despite the loss of protease activity, all 8 proteins have a characteristic NBCZone.

NBCZones based conclusions

Summarizing the results on a structural comparison of the NBCZones for 169 (chymo)trypsin-like fold proteases, we can conclude:

1) For the majority of eukaryotic and prokaryotic proteins, the presence of a $Cys42_T$-$Cys58_T$ disulfide bond and the location of alanine at position 55_T are interrelated. In those cases where

the analyzed proteins lack the $Cys42_T$-$Cys58_T$ disulfide bond, position 42_T is predominantly occupied by glycine, position 55_T by asparagine, and position 58_T by valine.

2) Viral proteases do not have the $Cys42_T$-$Cys58_T$ disulfide bond. In these proteases, position 42_T is predominantly glycine, phenylalanine, or leucine (cysteine proteases), position 55_T can be occupied by 9 different amino acids (predominantly hydrophobic residues or proline in serine proteases and hydrophilic uncharged residues or proline in cysteine proteases), and the amino acid at position 58_T is either valine or leucine (cysteine proteases).

It was shown that the presence or absence of the $Cys42_T$-$Cys58_T$ disulfide bond affects the overall thermal stability of trypsin [55]. Moreover, mutations Cys42Ala, Cys58Ala/Val and Ser195Thr convert the serine protease trypsin to a functional threonine protease.

Eukaryotic serine (chymo)trypsin-like fold proteases

Extension of trypsin NBCZone

As aforementioned, NBCZones of trypsin and trypsinogen do not differ from each other. An additional visual inspection of the trypsin tertiary structure showed that near the nucleophile-oxyanion loop Gly193-Gly197 there are two water molecules HOH1015 (position X) and HOH1003 (position Y) that form two hydrogen bonds with the main-chain oxygen atoms of Gly193 and Asp194, and two weak hydrogen bonds with the Cα-atoms of Asp194 and Gly197 (**Fig. 6A**, **Table 2**). Water molecules HOH1015 and HOH1003 are located at a distance comparable to the distance of a hydrogen bond. This conformation of the nucleophile-oxyanion loop corresponds to the zyme type trypsin conformation. The contacts HOH1015-O/Gly193 and HOH1015-CA/Asp194 dispose the N/Asp194 atom to the position required for the formation of an important β-turn contact (O/Cys191-N/Asp194; not shown). The contacts HOH1003-O/Asp194 and HOH1003-CA/Gly197 are important for the correct orientation of the nucleophile Ser195 and its N(OxyII) oxyanion atom. Consequently, HOH1015 and HOH1003 do not affect the position of the N(OxyI) oxyanion atom. In trypsin, the nitrogen atom N/Gly193 is located at the position (type II β-turn) that ensures the activity of this protease [10]. However, in other proteases with the same zyme pattern contacts, as shown in **Fig. 6A**, nitrogen N(OxyI) may not be appropriate for the activity of the protein (type I β-turn; see below). Therefore, the zyme structural organization is necessary, but it is

not sufficient to conclude whether the particular tertiary structure corresponds to the active or inactive state of the protease.

For Asp194, the amino acid of the neighboring nucleophile, significant zyme-zymogen conformational changes are observed [13]. The result of these changes is that the side-chain oxygen OD1 of Asp194 of trypsinogen occupies the position of water molecule HOH1015 (position X) that is found in the structure of trypsin (**Fig. 6B**, **Table 2**). Atom OD1/Asp194 has no contact with the main chain oxygen of Gly193. Tetrapeptide Cys191-Asp194 no longer forms a β-turn conformation in the trypsinogen structure (not shown). Position Y of the trypsinogen structure is still occupied by a water molecule HOH701. Contacts of HOH701 with O/Asp194 and CA/Gly197 atoms are conserved, but there is also an additional contact with CB/Asp194 atom.

Due to the catalytic importance of the described structural differences between trypsin and trypsinogen, it is necessary to extend the NBCZone of proteases by including the atomic contents of positions X and Y.

[ST]A group

In 58 eukaryotic (chymo)trypsin-like serine fold proteases of the [ST]A group there are two water molecules that are structurally similar to the waters HOH1015 (position X) and HOH1003 (position Y) in trypsin (**Table S2**). Once again, we emphasize here that the presence in these 58 proteins of the identical zyme pattern contacts does not automatically mean that their full catalytic center is in the active configuration, as in trypsin. The structural zymogen subgroup consists of 18 proteases (**Tables 2** and **S2)**.

Glycine at position 193_T and aspartic acid at position 194_T are absolutely conserved in these eukaryotic (chymo)trypsin-like serine fold proteases, and glycine is the most commonly found amino acid at positions 43_T and 197_T. However, of eleven proteases (**Tables 2** and **S2)**, five have alanine, four have serine, one has arginine, and one methionine instead of glycine at position 43_T [56, 57]. In addition, we found one example of glycine replaced at position 197_T by another amino acid, namely serine, in granzyme A (PDB ID: 1ORF, [58]), and alanine is found at position 43_T instead of the conserved glycine. However, this does not affect the presence of a water molecule at position X. Water is also present at position X in the remaining four proteases, all of which have alanine at position 43_T. In the case of serine at

position 43_T, a water molecule may be present or absent at position X. In the structure of granzyme A, there is no water molecule at position Y. Therefore, in 6 out of 11 proteases, the amino acid substitutions at position 43_T do not change the zyme or zymogen (2 proteases) contact diagrams, which are shown in **Figs. 6A** and **6B**, respectively.

At position 43_T of eukaryotic proteases, one can find not only small amino acids, such as glycine, alanine, or serine, but also large residues, e.g., arginine or methionine. These large amino acids at position 43_T are seen in complement factors C2 and B (PDB IDs: 2ODP and 1RRK, respectively [57, 59]); in these cases a pair of carbon atoms of the side chains at position 43_T are located at position X (**Tables 2** and **S2**).

TN group

As we showed above, although the probability is very high for the existence of a glycine residue at position 43_T of eukaryotic (chymo)trypsin-like serine fold proteases, this position is not absolutely conserved. The HtrA and Do-like families of *Homo sapiens* and *Arabidopsis thaliana* serine proteases also have a serine at position 43_T (**Tables 2** and **S2**) that leads to structural changes whereby the OG atom of $Ser43_T$ is locates to position X instead of a water molecule. The example of the chloroplastic protease Do-like 2 (PDB ID: 5ILB, [60]) shows the zyme contacts near the nucleophile-oxyanion loop (**Fig. 7A**).

While analyzing the tertiary structures of the eukaryotic (chymo)trypsin-like serine fold proteases, we never encountered an enzyme in which, as a result of activation, the side chain atom of the residue at position 194_T would be located in position X instead of the side chain atom of the residue at position 43_T. In particular, at position 194_T of TN group proteases there is asparagine instead of aspartate found at this position in the [ST]A group. Given the conservative changes in amino acids at positions 43_T, 55_T, and 194_T and the absence of the disulfide bond 42_T-58_T, it seems that proteases of the TN group do not have the ability to undergo a structural transition of the zymogen-zyme type. It is possible that this rule applies also to the complement factors C2 and B. This assumption is fully supported by the literature [29, 57, 59, 61].

Prokaryotic serine (chymo)trypsin-like fold proteases

TA and TN groups

Only two examples of the NBCZone extension in prokaryotic TA group proteases with the zyme (chymo)trypsin-like form and two examples of the NBCZone extension with the zymogen form were found in the PDB: trypsins from *Saccharopolyspora erythraea* (PDB ID: 5KWM, [62]) and *Streptomyces griseus* (PDB ID: 1SGT, [63]), VESB and VesC proteases from *Vibrio cholerae* (PDB ID: 4LK4, [64] and PDB ID: 6BQM, [65]) (**Tables 2** and **S2**). In terms of the organization of (chymo)trypsin-like zyme and zymogen forms, eukaryotic and prokaryotic NBCZone extensions of these four proteases are not different. Therefore, in the eukaryotic TA group proteases, the (chymo)trypsin-like form of NBCZone extension is dominant but in the prokaryotic proteases of this group it is auxiliary.

Most prokaryotic TA group proteases have the amino acid serine or threonine at position 43_T, and possess glycine at position 197_T (**Table S1**). In the corresponding TN group, all proteases have serine or threonine at position 43_T, and glycine or serine at position 197_T. As a result, in the cases where glycine is *not* located at position 43_T and 197_T, there are no water molecules at positions X and Y, but atoms of the side chains of serine or threonine are located there instead.

Among the prokaryotic TN group proteases, there is one exception from the viewpoint of building a zyme NBCZone extension pattern. With the serine protease SplE (**Table S1**, PDB ID: 5MM8, [32]) the dipeptide 193_T-194_T has a different conformation compared to the conformation of the corresponding dipeptide in the remaining proteases. As a result, the contact of the side-chain atom of the amino acid at position 43_T with the main-chain oxygen of amino acid at position 193_T is absent (Thr25 and Gly153, respectively, in the serine protease SplE). Earlier, we stated that, apparently, only members of eukaryotic and prokaryotic [ST]A groups can demonstrate two alternative conformations of the NBCZone extension pattern: zyme and zymogen. The NBCZone extension pattern of SplE shows that, sometimes, neither zyme nor zymogen form is possible. Instead a third form is formed, which is an atypical variant. Therefore, the letter A (Atypical) is added to the number of the structures of this protease and to the numbers of all similar structures in **Table S1**. This structural feature of the prokaryotic SplE protein is discussed in detail in the work [32].

43&[STG][AV] group

Six (chymo)trypsin-like fold structures belonging to the 43&[STG][AV] group have five different amino acids, other than glycine, at position 43_T (**Tables 2** and **S2**). However, all of these proteases have a water molecule at position X. This is due to the removal of the side chain of amino acid 43_T from position X as a result of a break in the course of the polypeptide chain in this place. Currently there are no data on the existence of the activation structure changes near the catalytic site in SPATE family proteases. Their N-terminal amino acid often forms a hydrogen bond with aspartate, an amino acid preceding the catalytic nucleophile [38].

Viral serine/cysteine (chymo)trypsin-like fold proteases

As aforementioned, a significant number of prokaryotic serine (chymo)trypsin-like fold proteases have amino acids, other than glycine, at positions 43_T and 197_T (**Table S1**, 22 structures), and only one such protease was observed among eukaryotic proteases (granzyme A). Analysis of the tertiary structures of viral proteases possessing such a fold showed that there are 37 such structures. With the exception of two proteases from the 43&[VR]N group, all viral proteases lack a glycine at position 43_T (**Tables 1**, **S1** and **S2**). As a result, there is no water molecule in position X, except among the four viral cysteine 3C proteases from the [ΨC][PQ] group (**Tables 2** and **S2**) [66].

Besides a greater number of such viral structures, there is another difference between prokaryotic and viral proteases. The prokaryotic proteases show conservation of the amino acids occupying positions 43_T (Thr) and 197_T (Ser), while the viral proteases demonstrate a sequence variability at these positions, since 14 and 8 different amino acids are located at the positions 43_T and 197_T, respectively (**Table S1**). Mutations Asn28(43_T) and Ser147(197_T) to alanine modulate the dimerization (active form of enzyme) and completely inactivate the main viral 3C-like proteinase from human SARS coronavirus [67, 68]. Interestingly, mutation Ser139(189_T)/Ala had a slight effect on the activity and dimer stability of the proteinase. Mutation Ser144(194_T)/Ala showed a two-fold decrease in catalytic efficiency compared to that of the wild type, but maintained a similar dimeric state. Ser139, Ser144 and Ser147 form a cluster of conserved serine residues near the catalytic nucleophile Cys145 of 3C-like proteinase.

Viral proteases with an atypical NBCZone

As with the members of the prokaryotic TN group, viral serine/cysteine (chymo)trypsin-like fold proteases also have several members with an atypical NBCZone (**Tables 2** and **S2**). They are: a putative serine protease (PDB ID: 2W5E, [69]), whose active site residues are not in the typical (chymo)trypsin-like conformation, and 4 additional proteins, non-structural protein NSP4 (PDB IDs: 5Y4L and 3FAN, [70, 71]), 3C-like protease (PDB IDs: 5E0G and 5E0J, [72]), 3C-like proteinase (PDB IDs: 5C5O and 2QCY, [50, 73]), and infectious bronchitis virus (IBV) main protease (PDB IDs: 2Q6F and 2Q6D, [74]), all of which demonstrate both a typical and atypical conformation of the dipeptide, residues 193_T-194_T.

Inactive proteases

As aforementioned, in eukaryotic serine (chymo)trypsin-like fold proteases, only glycine (with one exception) is found at the 197_T position of the nucleophile-oxyanion loop 193_T-197_T. We found five inactive proteins, in which this requirement is not the case (**Tables S1** and **S2**). In all these five proteins, instead of glycine in position 197_T there is serine, threonine, or aspartate amino acid. The presence of any residue other than glycine at position 197_T leads to a structural change in which a water molecule is replaced by the atom(s) of the side chain of the amino acid of residue 197_T. **Figure 7B** shows an example of a zyme contact pattern for such a protein: Heparin binding protein (PDB ID: 1A7S, [75]) has Thr177 at a position equivalent to position 197_T in trypsin (**Table 2**). The CB and CG2 atoms of the side chain of Thr177 have the same structural role in this protein as the water molecule HOH1003 in the structure of trypsin (see **Fig. 6A**). In position 43_T there is Gly27. Therefore, a water molecule HOH451 is located at the X position. The other four eukaryotic proteins also have water molecule at position X because they have glycine at position 43_T (**Table S2**).

Extension of NBCZone based conclusions

Summarizing the results of a structural comparison of the extension of the NBCZone for 169 (chymo)trypsin-like fold proteases, we can conclude:

1) The vast majority of eukaryotic zyme type proteases have glycine residues at positions 43_T and 197_T. As a result, one water molecule is located at position X and another at Y.

2) Eukaryotic zymogen type proteases place the side-chain oxygen of $Asp194_T$ at position X and water molecule at position Y.

3) In almost all prokaryotic and viral proteases, the amino acid at position 43_T is *not* glycine. This leads to the displacement of a water molecule at position X by the atom(s) of the side chain of the amino acid at position 43_T.

4) Due to the presence of a side chain in the amino acid at position 43_T, prokaryotic and viral proteases do not have the ability to undergo a structural transition from the zymogen to zyme type.

Conflict of interest

The authors declare no conflict of interest.

Author contributions

OMHS-A conceived and AID and KD designed the study. KD performed the structural bioinformatics study. AID, MSJ, and VNU analyzed the data. AID, KD, MSJ, and VNU wrote the manuscript. OMHS-A critically reviewed the manuscript.

Acknowledgements

This work is supported by a grant from the Sigrid Juselius Foundation and Joe, Pentti, and Tor Borg Memorial Fund. We thank the Biocenter Finland Bioinformatics Network (Dr. Jukka Lehtonen) and CSC IT Center for Science for computational support for the project. The Structural Bioinformatics Laboratory is part of the Drug Development and Diagnostics Platform of Åbo Akademi University.

Appendix A. Supplementary data

References

[1] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, J Mol Biol 247(4) (1995) 536-540.

[2] P.S. Dimitriou, A. Denesyuk, S. Takahashi, S. Yamashita, M.S. Johnson, T. Nakayama, K. Denessiouk, Alpha/beta-hydrolases: A unique structural motif coordinates catalytic acid residue in 40 protein fold families, Proteins 85(10) (2017) 1845-1855.

[3] M. Nardini, B.W. Dijkstra, Alpha/beta hydrolase fold enzymes: the family keeps growing, Curr Opin Struct Biol 9(6) (1999) 732-737.

[4] P.S. Dimitriou, A.I. Denesyuk, T. Nakayama, M.S. Johnson, K. Denessiouk, Distinctive structural motifs co-ordinate the catalytic nucleophile and the residues of the oxyanion hole in the alpha/beta-hydrolase fold enzymes, Protein Sci 28(2) (2019) 344-364.

[5] E. Di Cera, Serine proteases, IUBMB Life 61(5) (2009) 510-515.

[6] G. Dodson, A. Wlodawer, Catalytic triads and their relatives, Trends Biochem Sci 23(9) (1998) 347-352.

[7] L. Hedstrom, Serine protease mechanism and specificity, Chem Rev 102(12) (2002) 4501-4524.

[8] C.J. Sigrist, E. de Castro, L. Cerutti, B.A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, I. Xenarios, New and continuing developments at PROSITE, Nucleic Acids Res 41(Database issue) (2013) D344-D347.

[9] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res 28(1) (2000) 235-242.

[10] D. Liebschner, M. Dauter, A. Brzuszkiewicz, Z. Dauter, On the reproducibility of protein crystal structures: five atomic resolution structures of trypsin, Acta Crystallogr D Biol Crystallogr 69(Pt 8) (2013) 1447-1462.

[11] R. Menard, A.C. Storer, Oxyanion hole interactions in serine and cysteine proteases, Biol Chem Hoppe Seyler 373(7) (1992) 393-400.

[12] J. Walter, W. Steigemann, T.P. Singh, H. Bartunik, W. Bode, R. Huber, On the disordered activation domain in trypsinogen. Chemical labelling and low-temperature crystallography, Acta Crystallogr B38 (1982) 1462-1472.

[13] R. Huber, W. Bode, Structural Basis of the Activation, Action and Inhibition of Trypsin, H-S Z Physiol Chem 360(4) (1979) 489-489.

[14] N.L. Dawson, T.E. Lewis, S. Das, J.G. Lees, D. Lee, P. Ashford, C.A. Orengo, I. Sillitoe, CATH: an expanded resource to predict protein function through structure and sequence, Nucleic Acids Research 45(D1) (2017) D289-D295.

[15] N.D. Rawlings, J. Alan, P.D. Thomas, X.D. Huang, A. Bateman, R.D. Finn, The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database, Nucleic Acids Research 46(D1) (2018) D624-D632.

[16] M. Takeda-Shitaka, H. Umeyama, Elucidation of the cause for reduced activity of abnormal human plasmin containing an Ala55-Thr mutation: importance of highly conserved Ala55 in serine proteases, FEBS Lett 425(3) (1998) 448-452.

[17] M. Takeda-Shitaka, H. Umeyama, Effect of exceptional valine replacement for highly conserved alanine-55 on the catalytic site structure of chymotrypsin-like serine protease, Chem Pharm Bull (Tokyo) 46(9) (1998) 1343-1348.

[18] P. Gal, L. Barna, A. Kocsis, P. Zavodszky, Serine proteases of the classical and lectin pathways: similarities and differences, Immunobiology 212(4-5) (2007) 267-277.

[19] P. Gal, J. Dobo, P. Zavodszky, R.B. Sim, Early complement proteases: C1r, C1s and MASPs. A structural insight into activation and functions, Mol Immunol 46(14) (2009) 2745-2752.

[20] J.J. Perona, C.S. Craik, Evolutionary divergence of substrate specificity within the chymotrypsin-like serine protease fold, J Biol Chem 272(48) (1997) 29987-29990.

[21] J. Kardos, V. Harmat, A. Pallo, O. Barabas, K. Szilagyi, L. Graf, G.N. Szabo, Y. Goto, P. Zavodszky, P. Gal, Revisiting the mechanism of the autoactivation of the complement protease C1r in the C1 complex: Structure of the active catalytic region of C1r, Molecular Immunology 45(6) (2008) 1752-1760.

[22] D. Heja, V. Harmat, K. Fodor, M. Wilmanns, J. Dobo, K.A. Kekesi, P. Zavodszky, P. Gal, G. Pal, Monospecific Inhibitors Show That Both Mannan-binding Lectin-associated Serine Protease-1 (MASP-1) and-2 Are Essential for Lectin Pathway Activation and Reveal Structural Plasticity of MASP-2, J Biol Chem 287(24) (2012) 20290-20300.

[23] C. Gaboriaud, R.K. Gupta, L. Martin, M. Lacroix, L. Serre, F. Teillet, G.J. Arlaud, V. Rossi, N.M. Thielens, The Serine Protease Domain of MASP-3: Enzymatic Properties and Crystal Structure in Complex with Ecotin, Plos One 8(7) (2013).

[24] J. Dobo, V. Harmat, L. Beinrohr, E. Sebestyen, P. Zavodszky, P. Gal, MASP-1, a Promiscuous Complement Protease: Structure of Its Catalytic Region Reveals the Basis of Its Broad Specificity, J Immunol 183(2) (2009) 1207-1214.

[25] T. Clausen, C. Southan, M. Ehrmann, The HtrA family of proteases: Implications for protein composition and cell fate, Mol Cell 10(3) (2002) 443-455.

[26] D. Zurawa-Janicka, T. Wenta, M. Jarzab, J. Skorko-Glonek, P. Glaza, A. Gieldon, J. Ciarkowski, B. Lipinska, Structural insights into the activation mechanisms of human HtrA serine proteases, Arch Biochem Biophys 621 (2017) 6-23.

[27] M. Merski, C. Moreira, R.M.V. Abreu, M.J. Ramos, P.A. Fernandes, L.M. Martins, P.J.B. Pereira, S. Macedo-Ribeiro, Molecular motion regulates the activity of the Mitochondrial Serine Protease HtrA2, Cell Death Dis 8 (2017).

[28] W.Y. Wan, E.J. Milner-White, A natural grouping of motifs with an aspartate or asparagine residue forming two hydrogen bonds to residues ahead in sequence: Their occurrence at alpha-helical N termini and in other situations, Journal of Molecular Biology 286(5) (1999) 1633-1649.

[29] C. Eigenbrot, M. Ultsch, M.T. Lipari, P. Moran, S.J. Lin, R. Ganesan, C. Quan, J. Tom, W. Sandoval, M.V. Campagne, D. Kirchhofer, Structural and Functional Analysis of HtrA1 and Its Subdomains, Structure 20(6) (2012) 1040-1050.

[30] J. Kley, B. Schmidt, B. Boyanov, P.C. Stolt-Bergner, R. Kirk, M. Ehrmann, R.R. Knopf, L. Naveh, Z. Adam, T. Clausen, Structural adaptation of the plant protease Deg1 to repair photosystem II during light exposure, Nat Struct Mol Biol 18(6) (2011) 728-731.

[31] T. Wenta, P. Glaza, M. Jarzab, U. Zarzecka, D. Zurawa-Janicka, A. Lesner, J. Skorko-Glonek, B. Lipinska, The role of the LB structural loop and its interactions with the PDZ domain of the human HtrA3 protease, Bba-Proteins Proteom 1865(9) (2017) 1141-1151.

[32] N. Stach, M. Kalinska, M. Zdzalik, R. Kitel, A. Karim, K. Serwin, W. Rut, K. Larsen, A. Jabaiah, M. Firlej, B. Wladyka, P. Daugherty, H. Stennicke, M. Drag, J. Potempa, G. Dubin, Unique Substrate Specificity of SplE Serine Protease from Staphylococcus aureus, Structure 26(4) (2018) 572-579.

[33] G.M. Popowicz, G. Dubin, J. Stec-Niemczyk, A. Czarny, A. Dubin, J. Potempa, T.A. Holak, Functional and structural characterization of Spl proteases from Staphylococcus aureus, Journal of Molecular Biology 358(1) (2006) 270-279.

[34] J.A.R.G. Barbosa, J.W. Saldanha, R.C. Garratt, Novel features of serine protease active sites and specificity pockets: Sequence analysis and modelling studies of glutamate-specific endopeptidases and epidermolytic toxins, Protein Eng 9(7) (1996) 591-601.

[35] R. Meijers, E.V. Blagova, V.M. Levdikov, G.N. Rudenskaya, G.G. Chestukhina, S.V. Kostrov, V.S. Lamzin, I.P. Kuranova, The crystal structure of glutamyl endopeptidase from Bacillus intermedius reveals a structural link between zymogen activation and charge compensation, Biochemistry-Us 43(10) (2004) 2784-2791.

[36] K. Nishimura, N. Tajima, Y.H. Yoon, S.Y. Park, J.R.H. Tame, Autotransporter passenger proteins: virulence factors with common structural themes, J Mol Med 88(5) (2010) 451-458.

[37] F. Ruiz-Perez, J.P. Nataro, Bacterial serine proteases secreted by the autotransporter pathway: classification, specificity, and role in virulence, Cell Mol Life Sci 71(5) (2014) 745-770.

[38] T.A. Johnson, J.Z. Qiu, A.G. Plaut, T. Holyoak, Active-Site Gating Regulates Substrate Selectivity in a Chymotrypsin-Like Serine Protease: The Structure of Haemophilus influenzae Immunoglobulin A1 Protease, Journal of Molecular Biology 389(3) (2009) 559-574.

[39] R. Aasland, C. Abrams, C. Ampe, L.J. Ball, M.T. Bedford, G. Cesareni, M. Gimona, J.H. Hurley, T. Jarchau, V.P. Lehto, M.A. Lemmon, R. Linding, B.J. Mayer, M. Nagai, M. Sudol, U. Walter, S.J. Winder, Normalization of nomenclature for peptide motifs as ligands of modular protein domains, Febs Letters 513(1) (2002) 141-144.

[40] S. Lee, K.E. Owen, H.K. Choi, H. Lee, G.G. Lu, G. Wengler, D.T. Brown, M.G. Rossmann, R.J. Kuhn, Identification of a protein binding site on the surface of the alphavirus nucleocapsid and its implication in virus assembly, Structure 4(5) (1996) 531-541.

[41] Y. Zhou, D.J. Bartels, B.L. Hanzelka, U. Muh, Y. Wei, H.M. Chu, A.M. Tigges, D.L. Brennan, B.G. Rao, L. Swenson, A.D. Kwong, C. Lin, Phenotypic characterization of resistant Val36 variants of hepatitis C virus NS3-4A serine protease, Antimicrob Agents Chemother 52(1) (2008) 110-120.

[42] C. Welsch, F.S. Domingues, S. Susser, I. Antes, C. Hartmann, G. Mayr, A. Schlicker, C. Sarrazin, M. Albrecht, S. Zeuzem, T. Lengauer, Molecular basis of telaprevir resistance

due to V36 and T54 mutations in the NS3-4A protease of the hepatitis C virus, Genome Biol 9(1) (2008) R16.

[43] A.J. Thompson, J.G. McHutchison, Antiviral resistance and specifically targeted therapy for HCV (STAT-C), J Viral Hepat 16(6) (2009) 377-387.

[44] L.B. Zeminian, J.L. Padovani, S.M. Corvino, G.F. Silva, M.I. Pardini, R.M. Grotto, Variability and resistance mutations in the hepatitis C virus NS3 protease in patients not treated with protease inhibitors, Mem Inst Oswaldo Cruz 108(1) (2013) 13-17.

[45] J. Phan, A. Zdanov, A.G. Evdokimov, J.E. Tropea, H.K. Peters, 3rd, R.B. Kapust, M. Li, A. Wlodawer, D.S. Waugh, Structural basis for the substrate specificity of tobacco etch virus protease, J Biol Chem 277(52) (2002) 50564-50572.

[46] J. Yin, M.M. Cherney, E.M. Bergmann, J. Zhang, C. Huitema, H. Pettersson, L.D. Eltis, J.C. Vederas, M.N. James, An episulfide cation (thiiranium ring) trapped in the active site of HAV 3C proteinase inactivated by peptide-based ketone inhibitors, J Mol Biol 361(4) (2006) 673-686.

[47] M. Kanitz, S. Blanck, A. Heine, A.A. Gulyaeva, A.E. Gorbalenya, J. Ziebuhr, W.E. Diederich, Structural basis for catalysis and substrate specificity of a 3C-like cysteine protease from a mosquito mesonivirus, Virology 533 (2019) 21-33.

[48] Y. Sun, X. Wang, S. Yuan, M. Dang, X. Li, X.C. Zhang, Z. Rao, An open conformation determined by a structural switch for 2A protease from coxsackievirus A16, Protein Cell 4(10) (2013) 782-792.

[49] E.M. Bergmann, M.N.G. James, The 3C proteinases of picornaviruses and other positive-sense, single-stranded RNA viruses. , in: K. Von der Helm, B. Korant, J.C. Cheronis (Eds.), Proteases as Target for Therapy. , Springer-Verlag Berlin Heidelberg, 2000, pp. 117-143.

[50] J. Shi, J. Sivaraman, J. Song, Mechanism for controlling the dimer-monomer switch and coupling dimerization to catalysis of the severe acute respiratory syndrome coronavirus 3C-like protease, J Virol 82(9) (2008) 4620-4629.

[51] Y. Someya, N. Takeda, Functional consequences of mutational analysis of norovirus protease, FEBS Lett 585(2) (2011) 369-374.

[52] W. Sommergruber, J. Seipelt, F. Fessl, T. Skern, H.D. Liebig, G. Casari, Mutational analyses support a model for the HRV2 2A proteinase, Virology 234(2) (1997) 203-214.

[53] S. Piao, Y.L. Song, J.H. Kim, S.Y. Park, J.W. Park, B.L. Lee, B.H. Oh, N.C. Ha, Crystal structure of a clip-domain serine protease and functional roles of the clip domains, EMBO J 24(24) (2005) 4404-4414.

[54] Z.S. Derewenda, U. Derewenda, P.M. Kobos, (His)C epsilon-H...O=C < hydrogen bond in the active sites of serine hydrolases, J Mol Biol 241(1) (1994) 83-93.

[55] T.T. Baird, Jr., W.D. Wright, C.S. Craik, Conversion of trypsin to a functional threonine protease, Protein Sci 15(6) (2006) 1229-1238.

[56] B.T. Riley, O. Ilyichova, M.G. Costa, B.T. Porebski, S.J. de Veer, J.E. Swedberg, I. Kass, J.M. Harris, D.E. Hoke, A.M. Buckle, Direct and indirect mechanisms of KLK4 inhibition revealed by structure and dynamics, Sci Rep 6 (2016) 35385.

[57] V. Krishnan, Y. Xu, K. Macon, J.E. Volanakis, S.V. Narayana, The crystal structure of C2a, the catalytic fragment of classical pathway C3 and C5 convertase of human complement, J Mol Biol 367(1) (2007) 224-233.

[58] J.K. Bell, D.H. Goetz, S. Mahrus, J.L. Harris, R.J. Fletterick, C.S. Craik, The oligomeric structure of human granzyme A is a determinant of its extended substrate specificity, Nat Struct Biol 10(7) (2003) 527-534.

[59] K. Ponnuraj, Y. Xu, K. Macon, D. Moore, J.E. Volanakis, S.V. Narayana, Structural analysis of engineered Bb fragment of complement factor B: insights into the activation mechanism of the alternative pathway C3-convertase, Mol Cell 14(1) (2004) 17-28.

[60] M. Ouyang, X. Li, S. Zhao, H. Pu, J. Shen, Z. Adam, T. Clausen, L. Zhang, The crystal structure of Deg9 reveals a novel octameric-type HtrA protease, Nat Plants 3(12) (2017) 973-982.

[61] T. Clausen, M. Kaiser, R. Huber, M. Ehrmann, HTRA proteases: regulated proteolysis in protein quality control, Nat Rev Mol Cell Biol 12(3) (2011) 152-162.

[62] E. Blankenship, D.T.S. Lodowski, S. erythraea trypsin long construct, 2017.

[63] R.J. Read, M.N. James, Refined crystal structure of Streptomyces griseus trypsin at 1.7 A resolution, J Mol Biol 200(3) (1988) 523-551.

[64] S. Gadwal, K.V. Korotkov, J.R. Delarosa, W.G. Hol, M. Sandkvist, Functional and structural characterization of Vibrio cholerae extracellular serine protease B, VesB, J Biol Chem 289(12) (2014) 8288-8298.

[65] C.S. Rule, Y.J. Park, K.V. Korotkov, J.R. Delarosa, S. Turley, F. DiMaio, W.G.J. Hol, M. Sandkvist, Secondary mutations in Type II secretion mutants of Vibrio cholerae: inactivation of VesC, 2018.

[66] J. Wang, T. Fan, X. Yao, Z. Wu, L. Guo, X. Lei, J. Wang, M. Wang, Q. Jin, S. Cui, Crystal structures of enterovirus 71 3C protease complexed with rupintrivir reveal the roles of catalytically important residues, J Virol 85(19) (2011) 10021-10030.

[67] J. Barrila, S.B. Gabelli, U. Bacha, L.M. Amzel, E. Freire, Mutation of Asn28 disrupts the dimerization and enzymatic activity of SARS 3CL(pro), Biochemistry-Us 49(20) (2010) 4308-4317.

[68] J. Barrila, U. Bacha, E. Freire, Long-range cooperative interactions modulate dimerization in SARS 3CLpro, Biochemistry-Us 45(50) (2006) 14908-14916.

[69] S. Speroni, J. Rohayem, S. Nenci, D. Bonivento, I. Robel, J. Barthel, V.B. Luzhkov, B. Coutard, B. Canard, A. Mattevi, Structural and biochemical analysis of human pathogenic astrovirus serine protease at 2.0 A resolution, J Mol Biol 387(5) (2009) 1137-1152.

[70] Y. Shi, Y. Lei, G. Ye, L. Sun, L. Fang, S. Xiao, Z.F. Fu, P. Yin, Y. Song, G. Peng, Identification of two antiviral inhibitors targeting 3C-like serine/3C-like protease of porcine reproductive and respiratory syndrome virus and porcine epidemic diarrhea virus, Vet Microbiol 213 (2018) 114-122.

[71] X. Tian, G. Lu, F. Gao, H. Peng, Y. Feng, G. Ma, M. Bartlam, K. Tian, J. Yan, R. Hilgenfeld, G.F. Gao, Structure and cleavage specificity of the chymotrypsin-like serine protease (3CLSP/nsp4) of Porcine Reproductive and Respiratory Syndrome Virus (PRRSV), J Mol Biol 392(4) (2009) 977-993.

[72] P.M. Weerawarna, Y. Kim, A.C. Galasiti Kankanamalage, V.C. Damalanka, G.H. Lushington, K.R. Alliston, N. Mehzabeen, K.P. Battaile, S. Lovell, K.O. Chang, W.C. Groutas, Structure-based design and synthesis of triazole-based macrocyclic inhibitors of norovirus protease: Structural, biochemical, spectroscopic, and antiviral studies, Eur J Med Chem 119 (2016) 300-318.

[73] Y. Shimamoto, Y. Hattori, K. Kobayashi, K. Teruya, A. Sanjoh, A. Nakagawa, E. Yamashita, K. Akaji, Fused-ring structure of decahydroisoquinolin as a novel scaffold for SARS 3CL protease inhibitors, Bioorg Med Chem 23(4) (2015) 876-890.

[74] X. Xue, H. Yu, H. Yang, F. Xue, Z. Wu, W. Shen, J. Li, Z. Zhou, Y. Ding, Q. Zhao, X.C. Zhang, M. Liao, M. Bartlam, Z. Rao, Structures of two coronavirus main proteases: implications for substrate binding and antiviral drug design, J Virol 82(5) (2008) 2515-2527.

[75] S. Karlsen, L.F. Iversen, I.K. Larsen, H.J. Flodgaard, J.S. Kastrup, Atomic resolution structure of human HBP/CAP37/azurocidin, Acta Crystallogr D Biol Crystallogr 54(Pt 4) (1998) 598-609.

[76] Dassault Systèmes BIOVIA, Discovery Studio Modeling Environment, Dassault Systèmes, San Diego, 2016.

[77] J.V. Lehtonen, D.J. Still, V.V. Rantanen, J. Ekholm, D. Bjorklund, Z. Iftikhar, M. Huhtala, S. Repo, A. Jussila, J. Jaakkola, O. Pentikainen, T. Nyronen, T. Salminen, M. Gyllenberg, M.S. Johnson, BODIL: a molecular modeling environment for structure-function analysis and drug design, J Comput Aided Mol Des 18(6) (2004) 401-419.

[78] P.J. Kraulis, MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures., J Appl Cryst 24 (1991) 946–950.

[79] E.A. Merritt, D.J. Bacon, Raster3D: photorealistic molecular graphics, Methods Enzymol 277 (1997) 505-524.

Figure legends

Fig. 1. TRYPSIN_SER and TRYPSIN_HIS patterns, catalytic triad: Asp102, His57 and Ser195, and two oxyanion nitrogen atoms in the active site of trypsin (PDB ID: 4I8H). Positions of amino acid residues 43 and 213, disulfide bond C42-C58, and two water molecules in positions X and Y are also shown. Structural data were visualized and analyzed using Discovery Studio [76] and Bodil [77]. Figures were drawn with MolScript [78] and Raster3D [79].

Fig. 2. 43/213 NBCZone (A) and 42/43 Base Catalytic Zone (B) of trypsin. 43/213 NBCZone consists of four fragments of the amino acid sequence forming a ring-shaped structure due to the presence of six hydrogen bonds between them (shown by a dotted line). 42/43 Base Catalytic Zone consists of two fragments of the amino acid sequence.

Fig. 3. Two types of structural organization of the 43/213 NBCZone as a result of changes in the conformation of $Asn55_T$ side chain in the TN group of trypsin-like serine fold proteases. A) and B) $Asn55_T$ ND2 atom takes part in the formation of the 43/213 NBCZone. In this case, OD1 atom either forms an Asx-turn with the catalytic histidine (A) or interacts with the main chain oxygen of the catalytic acid (B). C) and D) $Asn55_T$ OD1 atom takes part in the formation of the 43/213 NBCZone. In this case, ND2 atom is in contact with the catalytic acid (C) or base (D).

Fig. 4. In the 43&[STG]V group of proteases (Ile86 in immunoglobulin A1 protease) the canonical hydrogen bond NH...O is replaced with a weak hydrogen bond CH...O due to a change in the conformation of the amino acid backbone at position 43_T.

Fig. 5. Structural organization of the 43/213 NBCZone in [KR]P group of proteases.

Fig. 6. Expansion of the NBCZone due to A) the inclusion of two water molecules in the positions X and Y of trypsin, and B) an OD1 atom of Asp194 and one water molecule in the corresponding positions of trypsinogen.

Fig. 7. Substitution of water molecules at the X or Y positions with the atom(s) of the side chains of amino acids at positions 43_T (A) and 197_T (B), respectively.

Figures

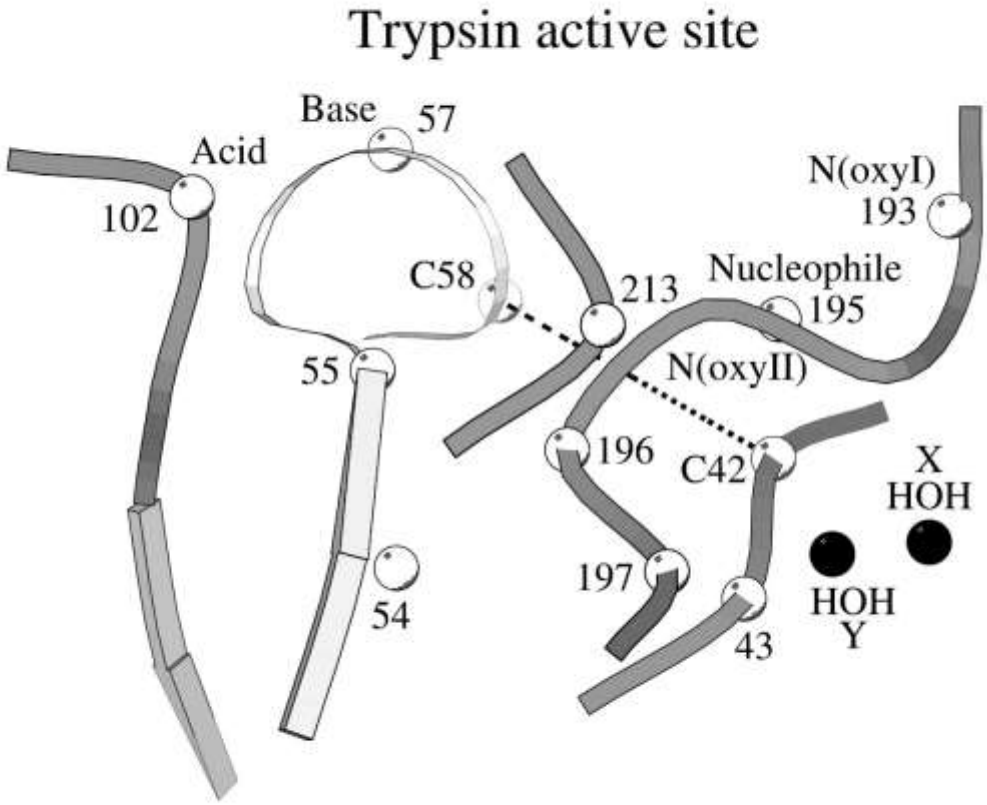

Figure1.

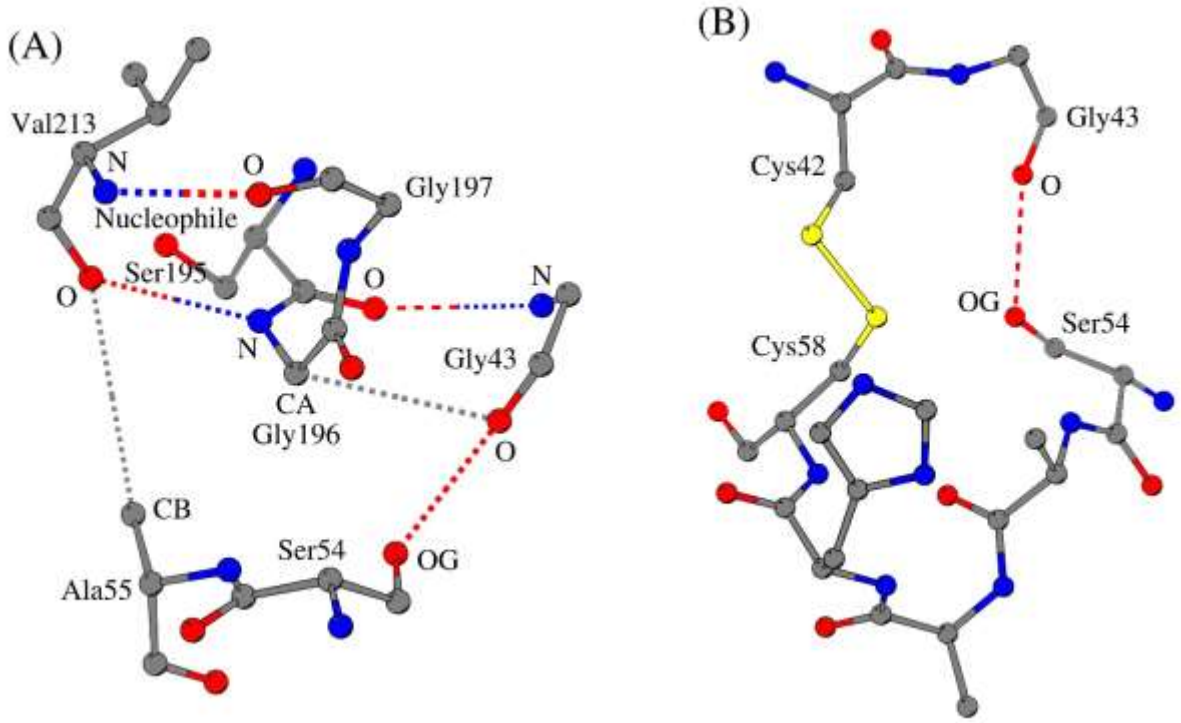
(A)
Val213
N
O
Gly197
Nucleophile
Ser195
O
O
N
N
Gly43
CA
Gly196
O
CB
Ser54
OG
Ala55
(B)
Gly43
Cys42
O
OG
Ser54
Cys58

Figure 2.

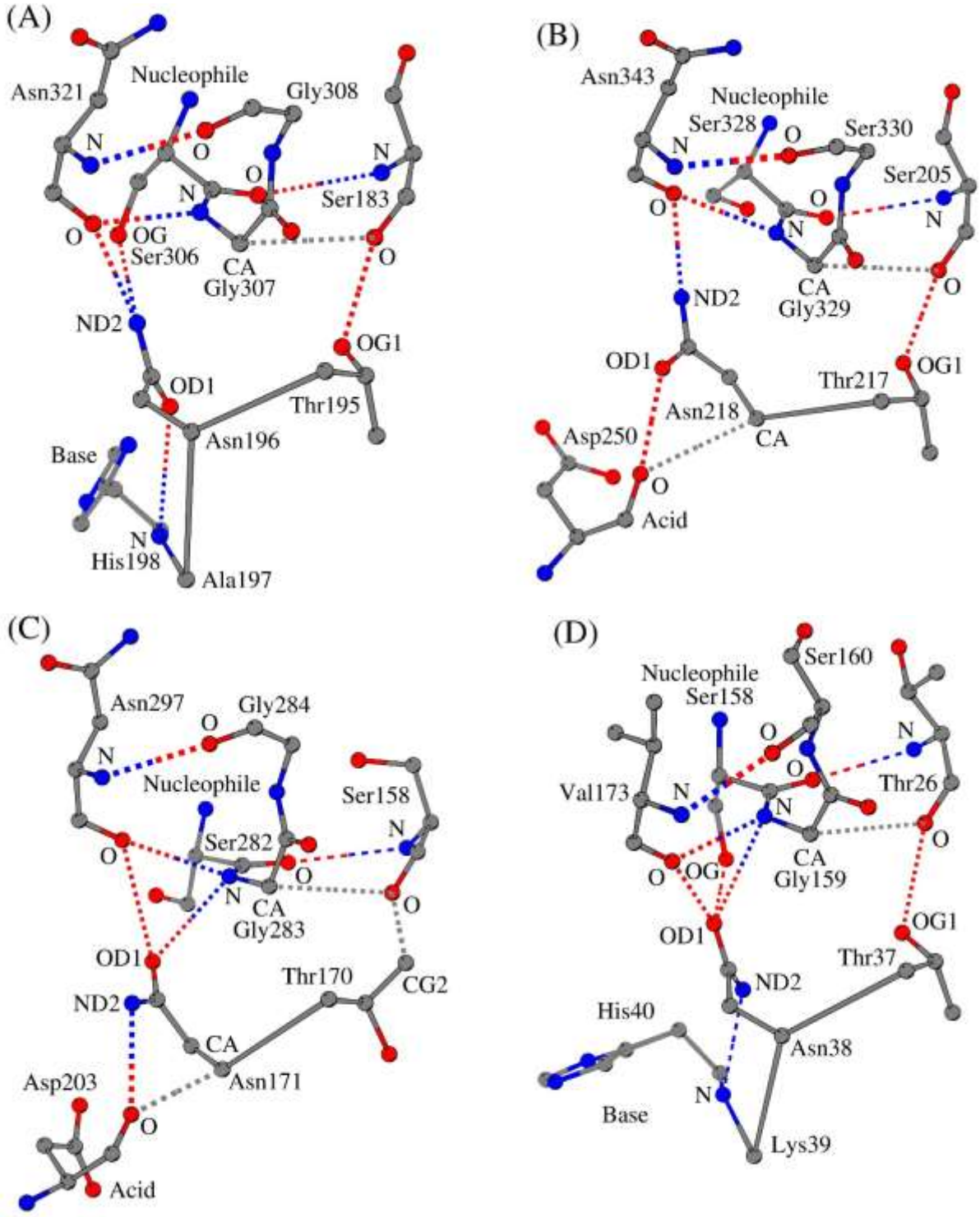
(A)
Asn321
Nucleophile
Gly308
N
O
O
N
N
Ser183
O
OG
Ser306
CA
Gly307
O
ND2
OG1
OD1
Thr195
Asn196
Base
N
His198
Ala197
(B)
Asn343
Nucleophile
Ser328
O
Ser330
N
Ser205
O
O
N
N
CA
Gly329
O
ND2
OD1
Thr217
OG1
Asn218
CA
Asp250
O
Acid
(C)
Asn297
Gly284
O
N
Nucleophile
Ser158
Ser282
N
O
O
N
CA
Gly283
O
OD1
Thr170
CG2
ND2
CA
Asp203
Asn171
O
Acid
(D)
Ser160
Nucleophile
Ser158
O
N
O
Val173
N
Thr26
N
O
CA
Gly159
O
OG
OD1
OG1
Thr37
ND2
His40
Asn38
N
Base
Lys39

Figure 3.

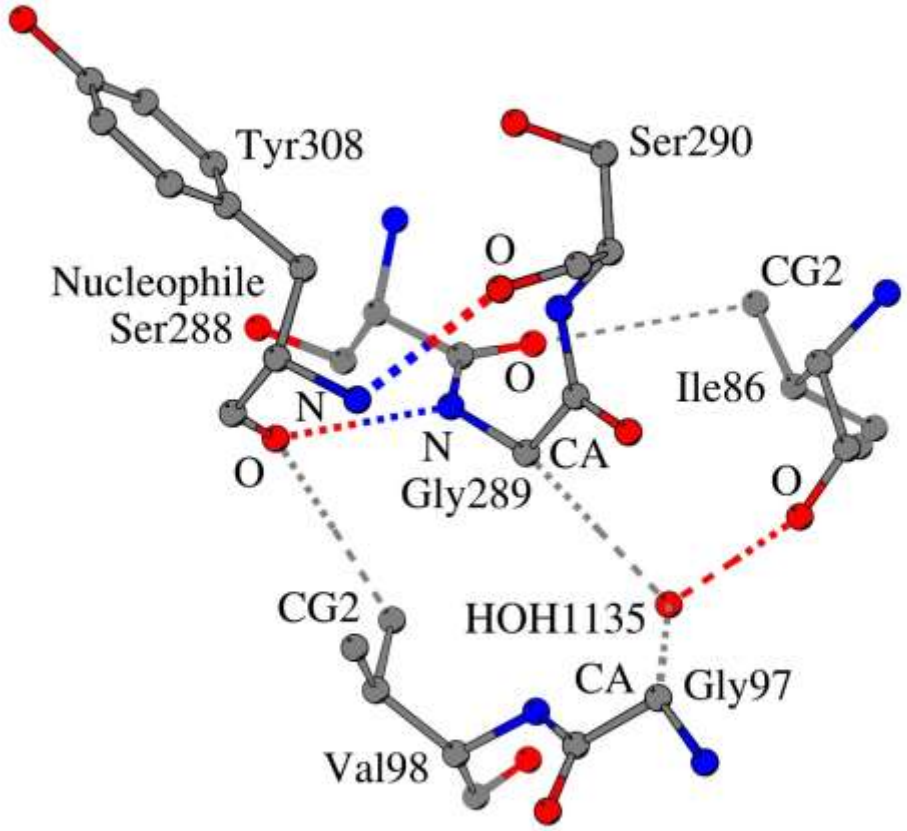
Tyr308
Ser290
Nucleophile
Ser288
O
CG2
O
Ile86
N
N
CA
O
Gly289
O
CG2
HOH1135
CA
Gly97
Val98

Figure 4.

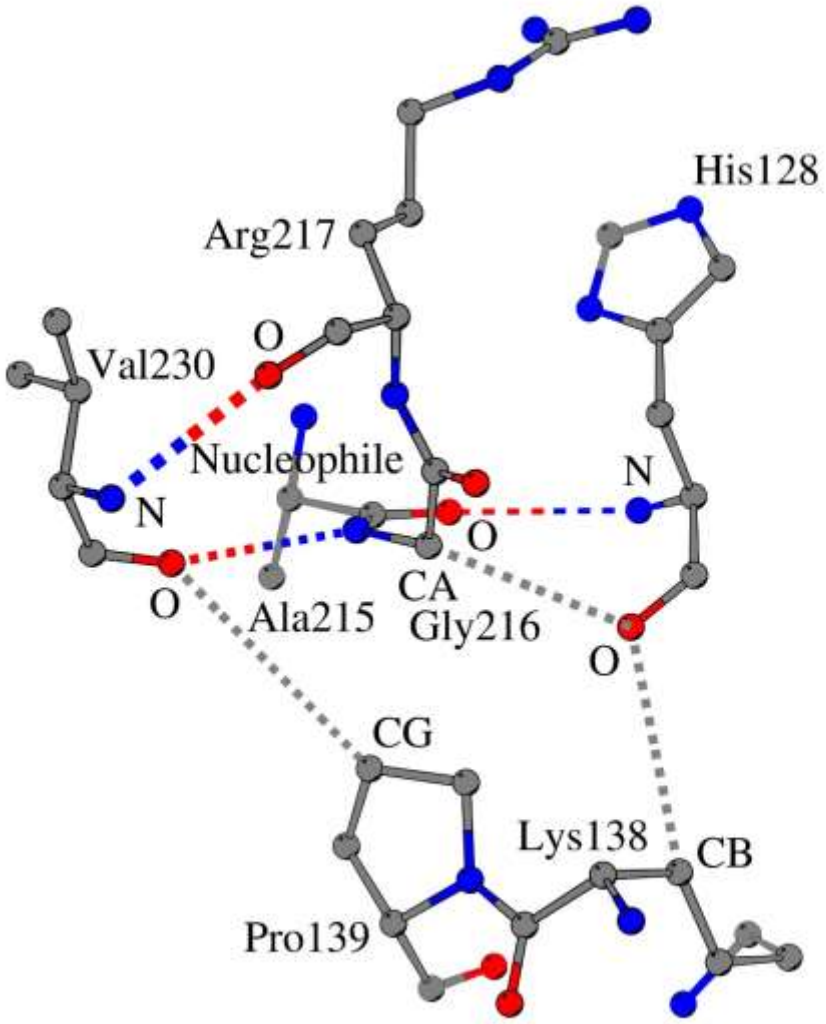
His128
Arg217
O
Val230
Nucleophile
N
N
O
O
CA
Ala215
Gly216
O
CG
Lys138
CB
Pro139

Figure 5.

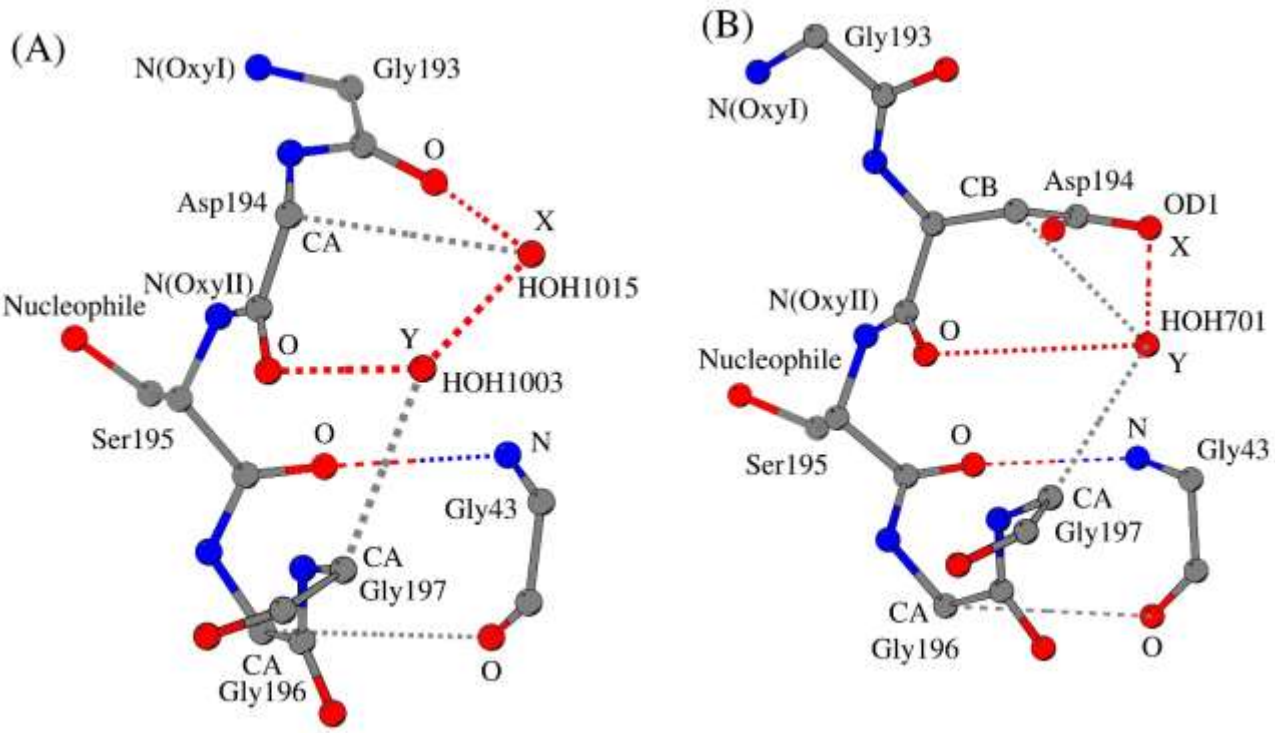
(A)
N(OxyI)
Gly193
O
Asp194
CA
X
N(OxyII)
HOH1015
Nucleophile
O
Y
HOH1003
Ser195
O
N
Gly43
CA
Gly197
CA
Gly196
O
(B)
Gly193
N(OxyI)
CB
Asp194
OD1
X
N(OxyII)
HOH701
O
Nucleophile
Y
Ser195
O
N
Gly43
CA
Gly197
CA
Gly196
O

Figure 6.

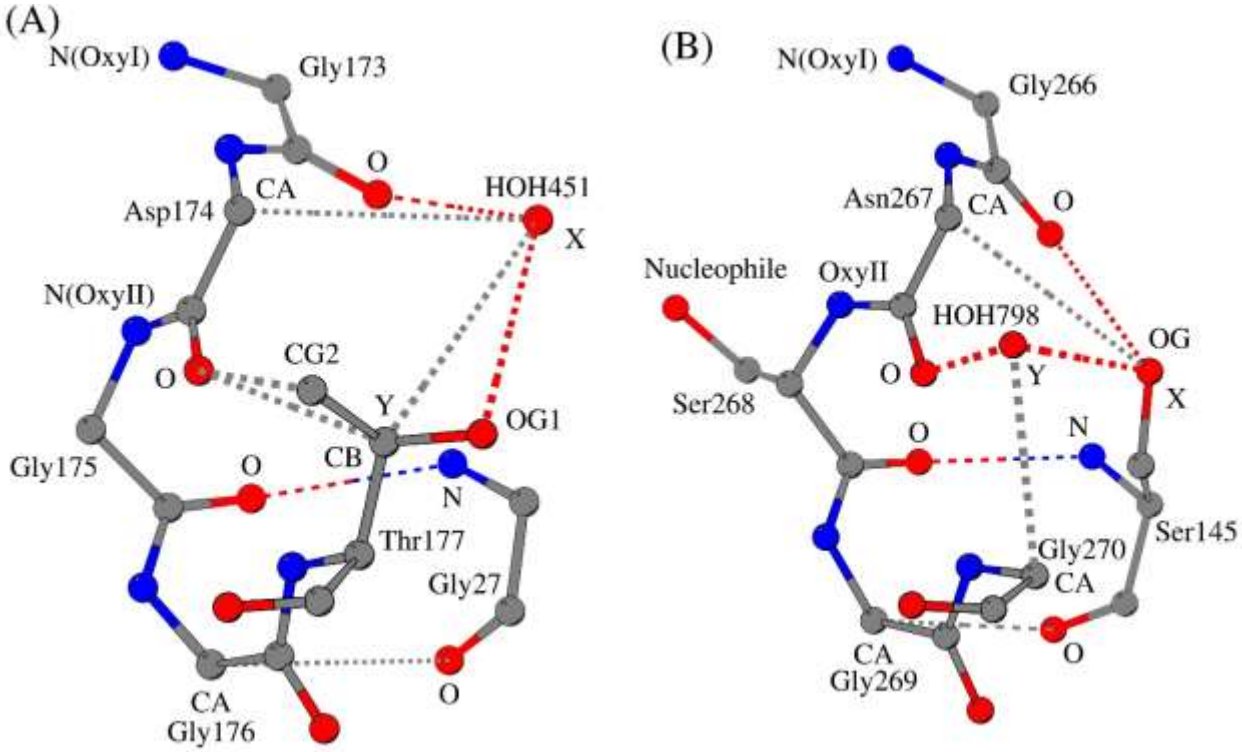
(A)
N(OxyI)
Gly173
O
CA
HOH451
Asp174
X
N(OxyII)
CG2
O
Y
OG1
Gly175
O
CB
N
Thr177
Gly27
CA
Gly176
O
(B)
N(OxyI)
Gly266
Asn267
CA
O
Nucleophile
OxyII
HOH798
OG
O
Y
X
Ser268
O
N
Ser145
Gly270
CA
CA
Gly269
O

Figure 7.

Tables

Table 1. Geometrical parameters of interactions within the amino acid sets forming NBCZones in representative structures of nine (chymo)trypsin-like serine fold proteases groups

Protein	Organism	PDB ID Resolution	Hydrogen bonds of amino acid at position 43_T	Hydrogen bonds of amino acid at position 213_T	Interactions of amino acids at positions 42_T&58_T	Ref.
Eukaryotic proteases						
[ST]A group						
Trypsin	*Bos taurus*	4I8H_A R=0.75 Å	N/G43-O/S195, 2.8 O/G43-CA/G196, 3.4 (2.7) O/G43-OG/S54, 2.8	N/V213-O/G197, 2.9 O/V213-N/G196, 2.9 O/V213-CB/A55, 4.0 (3.0)	CB/C42-O/S195, 3.3 (2.7) SG/C58-O/S195, 3.6 C42-C58	10
Trypsino-gen	*Bos taurus*	1TGT_A R=1.70 Å	N/G43-O/S195, 2.7 O/G43-CA/G196, 3.6 (2.8) O/G43-OG/S54, 2.8	N/V213-O/G197, 2.7 O/V213-N/G196, 2.9 O/V213-CB/A55, 4.1 (3.2)	CA/C42-O/S195, 3.4 (2.8) SG/C58-O/S195, 3.6 C42-C58	12
Mannan-binding lectin serine protease 2	*Homo sapiens*	3TVJ_B R=1.28 Å	N/A469-O/S633, 2.9 O/A469-CA/G634, 3.2 (2.3) O/A469-OG1/T480, 2.8	N/V653-O/G635, 2.9 O/V653-N/G634, 3.0 O/V653-CB/A481, 4.0 (3.0)	CB/A468-O/S633, 3.7 (2.8) CB/A484-O/S633, 5.0 (4.3) CB/A468-CB/A484, 4.3	22
TN group						
Serine protease HTRA2, mitochon-drial	*Homo sapiens*	5M3N_A R=1.65 Å Set I	N/S183-O/S306, 3.0 O/S183-CA/G307, 3.2 (2.5) O/S183-OG1/T195, 2.7	N/N321-O/G308, 2.9 O/N321-N/G307, 2.8 O/N321-ND2/N196, 3.2	CA/G182-O/S306, 3.6 (2.9) CG2/V199-O/S306, 4.0 (3.2) CA/G182-CG2/V199, 4.4	27
Serine protease HTRA1	*Homo sapiens*	3TJN_B R=3.00 Å Set II	N/S205-O/S328, 3.1 O/S205-CA/G329, 3.3 (2.4) O/S205-OG1/T217, 2.5	N/N343-O/G330, 3.0 O/N343-N/G329, 2.9 O/N343-ND2/N218, 3.0	CA/G204-O/S328, 4.0 (3.6) CG2/V221-O/S328, 3.5 (2.5) CA/G204-CG1/V221, 4.0	29
Protease Do-like 1, chloro-plastic	*Arabi-dopsis thaliana*	3QO6_A R=2.50 Å Set III	N/S158-O/S282, 3.1 O/S158-CA/G283, 3.3 (2.3) O/S158-CG2/T170, 2.6 (2.0)	N/N297-O/G284, 2.9 O/N297-N/G283, 2.9 O/N297-OD1/N171, 3.3 N/G283-OD1/N171, 2.9	CA/G157-O/S282, 4.1 (3.6) CG2/V174-O/S282, 4.0 (3.2) CA/G157-CG1/V174, 4.2	30
Prokaryotic proteases						
TN group						
Serine protease Spl	*Staphy-lococcus aureus*	2AS9_A R=1.70 Å Set IV	N/T26-O/S158, 2.8 O/T26-CA/G159, 3.2 (2.3) O/T26-OG1/T37, 2.9	N/V173-O/S160, 2.8 O/V173-N/G159, 3.0 O/V173-OD1/N38, 2.9 N/G159-OD1/N38, 3.1	CB/A25-O/S158, 3.6 (2.7) CG2/V41-O/S158, 4.3 (3.5) CB/A25-CG2/V41, 3.5	33
43&[STG]V group						
Immuno-globulin A1 protease	*Haemo-philus influ-enzae*	3H09_A R=1.75 Å	CG2/I86-O/S288, 3.4 (2.6) O/I86-HOH1135, 2.9 HOH1135-CA/G289, 3.4 (2.5) HOH1135-CA/G97, 3.4 (2.5)	N/Y308-O/S290, 2.8 O/Y308-N/G289, 2.8 O/Y308-CG2/V98, 3.7 (2.9)	CG2/V101-O/S288, 3.6 (2.6) CD1/I86-CG1/V101, 3.5	38
Viral serine proteases						
[KR]P group						
Sindbis virus capsid protein	*Sindbis virus*	1SVP_A R=2.00 Å	N/H128-O/S215A, 2.9 O/H128-CA/G216, 3.3 (2.4) O/H128-CB/K138, 3.8 (3.3)	N/V230-O/R217, 3.0 O/V230-N/G216, 2.9 O/V230-CG/P139, 4.4 (3.8)	CA/G127-O/S215A, 3.3 (2.7) CG2/V142-O/S215A, 4.4 (3.4) CA/G127-CG1/V142, 3.7	40
Viral cysteine proteases						

[TA]N group						
Nuclear inclusion protein A	*Tobacco etch virus*	1LVM_A R=1.80 Å Set III	N/Y33-O/C151, 2.8 O/Y33-CA/G152, 3.3 (2.3) O/Y33-OG1/T43, 2.7	N/H167-O/S153, 2.9 O/H167-N/G152, 2.9 O/H167-ND2/N44, 3.2 N/G152-OD1/N44, 2.9	CB/L32-O/C151, 3.5 (2.5) CD2/L32-CD1/L47, 3.7	45
[ΨC][PQ] group						
Hepatitis A protease 3C	*Human hepatitis A virus*	2HAL_A R=1.35 Å	N/N30-O/C172, 3.0 O/N30-CA/G173, 3.4 (2.3) O/N30-CG2/V41, 4.1 (3.0)	N/H191-O/G174, 2.9 O/H191-N/G173, 2.9 O/H191-CG/P42, 4.1 (3.5)	CB/M29-O/C172, 3.5 (2.4) CE/M29-CB/A45, 3.7	46
3Cl protease	*Alphamesonivirus 1*	5LAC_B R=1.94 Å	N/R35-O/C153, 3.0 O/R35-CA/G154, 3.4 (2.3) O/R35-HOH537, 2.7 HOH537-CA/I45, 3.3 (2.4)	N/H168-O/G155, 2.7 O/H168-N/G154, 2.9 O/H168-NE2/Q46, 2.9	CB/L34-O/C153, 3.6 (2.5) CG/L34-CD1/L49, 4.0	47
43&[VR]N group						
2A proteinase	*Coxsackievirus A16*	4MG3_A R=1.80 Å Set III	N/A O/G8-CA/G111, 3.6 (3.1) O/G8-HOH303, 2.6 HOH303-CG2/V18, 2.8 (2.2)	N/V124-O/G112, 2.8 O/V124-N/G111, 3.0 O/V124-ND2/N19, 2.8 HOH303-OD1/N19, 2.7 N/G111-OD1/N19, 3.0	CD1/L22-O/C110, 3.9 (3.6) CD1/L22-HOH303, 3.8 (2.9)	48
Inactive proteases						
Eukaryotic proteases						
T[TG] group						
Prophenoloxidase activating factor-II	*Holotrichia diomphalia*	2B9L_A R=2.00 Å	N/G186-O/G353, 2.6 O/G186-CA/G354, 3.7 (3.0) O/G186-OG1/T197, 2.9	N/V374-O/S355, 2.9 O/V374-N/G354, 2.8 O/V374-CA/G198, 5.4 (4.5) O/V374-CE1/H200, 3.5 (2.4)	CB/C185-O/G353, 3.1 (2.5) SG/C201-O/G353, 3.8 C185-C201	53

Sets “I-IV” refer to four subgroups of TN groups proteases with different orientation of $Asn55_T$. The values within the parentheses indicate distances to hydrogen atoms.

Table 2. Geometrical parameters of interactions within the positions X and Y of active sites in representative structures of (chymo)trypsin-like serine fold proteases

Protein	Organism	PDB ID Resolution	Nuc195 Xaa43 Xaa197	Hydrogen bonds of water molecule or amino acid at position X	Hydrogen bonds of water molecule or amino acid at position Y	Ref.
Eukaryotic proteases						
[ST]A group						
Trypsin	*Bos taurus*	4I8H_A R=0.75 Å	Ser195 Gly43 Gly197	HOH1015-O/G193, 2.8 HOH1015-CA/D194, 3.7 (2.7) HOH1015-HOH1003, 2.8	HOH1003-O/D194, 2.9 HOH1003-CA/G197, 3.3 (2.2)	10
Trypsinogen	*Bos taurus*	1TGT_A R=1.70 Å	Ser195 Gly43 Gly197	OD1/D194-HOH701, 2.9	HOH701-CB/D194, 3.4 (2.5) HOH701-O/D194, 3.7 HOH701-CA/G197, 3.1 (2.4)	12
Mannan-binding lectin serine protease 2	*Homo sapiens*	3TVJ_B R=1.28 Å	Ser633 Ala469 G635	HOH6-O/G631, 2.8 HOH6-CA/D632, 3.8 (2.8) HOH6-CB/A469, 3.5 (2.6) HOH6-HOH10, 2.9	HOH10-O/D632, 2.9 HOH10-CA/G635, 3.6 (2.7) HOH10-CB/A469, 3.2 (2.7)	22
Kallikrein-4	*Homo sapiens*	4K8Y_A R=1.00 Å	S195 S43 G197	HOH549-O/G193, 2.8 HOH549-CA/D194, 3.6 (2.6) HOH549-OG/S43, 2.7 HOH549-HOH301, 2.8	HOH301-O/D194, 2.8 HOH301-CA/197, 3.4 (2.4) HOH301-CB/S43, 3.4 (2.6)	56
Complement factor C2	*Homo sapiens*	2ODP_A R=1.90 Å	Ser659 Arg473 Gly661	NH1/R473-O/G657, 3.1 CG/R473-O/G657, 3.6 (2.5) CG/R473-CA/E658, 4.0 CB/R473-HOH960, 3.7 (3.1) CG/R473-HOH960, 3.7 (3.0)	HOH960-O/E658, 2.8 HOH960-CA/G661, 3.4 (2.4)	57
TN group						
Protease Do-like 2, chloroplastic	*Arabidopsis thaliana*	5ILB_A R=1.85 Å	Ser268 Ser145 Gly270	OG/S145-O/G266, 2.7 OG/S145-CA/N267, 4.0 (3.3) OG/S145-HOH798, 4.6	HOH798-O/N267, 2.8 HOH798-CA/G270, 3.7 (2.8)	60
Prokaryotic proteases						
TA group						
Trypsin	*Saccharopolyspora erythraea*	5KWM_A R=0.78 Å	Ser179 Gly28 Gly181	HOH490-O/G177, 2.9 HOH490-CA/D178, 3.7 (2.8) HOH490-HOH480, 2.8	HOH480-O/D194, 2.9 HOH480-CA/G197, 3.3 (2.4)	62
VESB protease	*Vibrio cholerae*	4LK4_A R=2.40 Å	S221A Gly64 Gly223	OD1/D194-HOH455, 3.0	HOH455-O/D220, 3.3 HOH455-CA/G223, 4.6 (3.6)	64
43&[STG][AV] group						
Immunoglobulin A1 protease	*Haemophilus influenzae*	3H09_A R=1.75 Å	S288 I86 S290	HOH1038-O/G286, 2.9 HOH1038-CA/D287, 3.6 (2.6) HOH1038-CG2/I86, 3.6 (3.1) HOH1038-CB/S290, 3.7 (2.7)	OG/S290-O/D287, 2.6	38
Viral serine proteases						
TA group						
Putative serine protease	*Human astrovirus-1*	2W5E_A R=2.00 Å	S551 T448 A553	HOH2022-O/G549, 3.2 HOH2022-CA/M550, 3.6 (2.9) HOH2022-OG1/T448, 2.7 HOH2022-CB/A553, 5.5 (4.5)	CB/A553-O/M550, 3.1 (2.7) CB/A553-CB/T448, 4.5	69

Viral cysteine proteases						
[ΨC][PQ] group						
EV71 3C protease	*Enterovirus A71*	3R0F_A R=1.31 Å	C147 T26 G149	HOH186-O/G145, 2.8 HOH186-CA/Q146, 3.3 (2.7)	No HOH	66
Inactive proteases						
Eukaryotic proteases						
TA group						
Heparin binding protein	*Homo sapiens*	1A7S_A R=1.12 Å	G175 Gly27 Thr177	HOH451-O/G173, 2.9 HOH451-CA/D174, 4.0 (3.0) HOH451-OG1/T177, 4.5 HOH451-CB/T177, 4.6 (3.8)	CB/T177-O/D174, 3.3 (2.7) CG2/T177-O/D174, 3.2 (2.5)	75