Biol. Chem., Vol. 393, pp. 843–851, September 2012 • Copyright © by Walter de Gruyter • Berlin • Boston. DOI 10.1515/hsz-2012-0167

Review

Current and prospective applications of non-proteinogenic amino acids in profiling of proteases substrate specificity

Paulina Kasperkiewicz[a], Anna D. Gajda[a] and Marcin Drąg*

Division of Bioorganic Chemistry, Faculty of Chemistry, Wroclaw University of Technology, 50-370 Wroclaw, Poland

* Corresponding author
e-mail: marcin.drag@pwr.wroc.pl

Abstract

Proteases recognize their endogenous substrates based largely on a sequence of proteinogenic amino acids that surrounds the cleavage site. Currently, several methods are available to determine protease substrate specificity based on approaches employing proteinogenic amino acids. The knowledge about the specificity of proteases can be significantly extended by application of structurally diverse families of non-proteinogenic amino acids. From a chemical point of view, this information may be used to design specific substrates, inhibitors, or activity-based probes, while biological functions of proteases, such as posttranslational modifications can also be investigated. In this review, we discuss current and prospective technologies for application of non-proteinogenic amino acids in protease substrate specificity profiling.

Keywords: combinatorial library; endopeptidase; exopeptidase.

Introduction

Proteases are hydrolases that irreversibly cleave peptide bonds in protein substrates. These enzymes participate in several regulation stages of cellular control, such as apoptosis, selective proteasome-based protein degradation, coagulation, or fibrinolysis (Drag and Salvesen, 2010). Proteolytic enzymes are also major players in the development of disorders, such as cancer, diabetes, malaria, primary hypertension, or HIV (Turk, 2006).

An individual feature of each protease is its ability to recognize specific protein substrates (Poreba and Drag, 2010). According to the Schechter and Berger (1967) nomenclature, this can be achieved through binding of side chains of amino acids of a peptide sequence (called P1′, P2′, P1, P2, and so on) into pockets (S1′, S2′, S1, S2, and so on) on the enzyme surface localized near the active site (Figure 1). Due to the ability to hydrolyze peptide bonds internal to peptide or protein sequences, endopeptidases usually have more specific pockets, while in the case of exopeptidases (aminopeptidases, carboxypeptidases, or dipeptidyl peptidases), this is limited usually to no more than three or four positions.

To date, several methods are available for determination of protease substrate specificity: proteomics, combinatorial chemistry peptide synthesis, or phage display (Schilling and Overall, 2007; Schneider and Craik, 2009; Poreba and Drag, 2010). Information obtained using these approaches has been used to design several useful substrates, inhibitors, and activity-based probes (ABPs), and even for identification of endogenous substrates (Rano et al., 1997; Backes et al., 2000; Choe et al., 2006; Drag et al., 2008). One of the biggest limitations of these approaches is the limitation of utilizing only proteinogenic amino acids. While this is sufficient for the investigation of biological functions of proteases, this is not always enough in the case of the design of specific substrates or inhibitors for individual enzymes. For example, despite several substrate specificity profiling approaches available to date, it has not been possible to design very specific low molecular weight substrates or inhibitors based only on proteinogenic amino acid scaffold for certain protease families, including caspases, cathepsins, or matrix metalloproteases (MMPs) (McStay et al., 2008; Drag and Salvesen, 2010).

A solution to this problem can be the application of non-proteinogenic amino acids in libraries used for substrate specificity profiling. Currently, there are hundreds of commercially available and structurally different non-proteinogenic derivatives of amino acids suitable for immediate application in peptide synthesis. While in the case of proteomics or the phage display approach, this could be difficult to achieve due to methodological problems, properly designed combinatorial libraries of synthetic fluorogenic or chromogenic substrates could yield very useful information about protease substrate specificity.

Non-proteinogenic amino acids

Natural or chemically designed and synthesized amino acids, which are non-genetically coded, are considered as non-proteinogenic amino acids. In a more rigorous classification for non-proteinogenic amino acids, we consider all amino acids

[a]These authors contributed equally to the work.

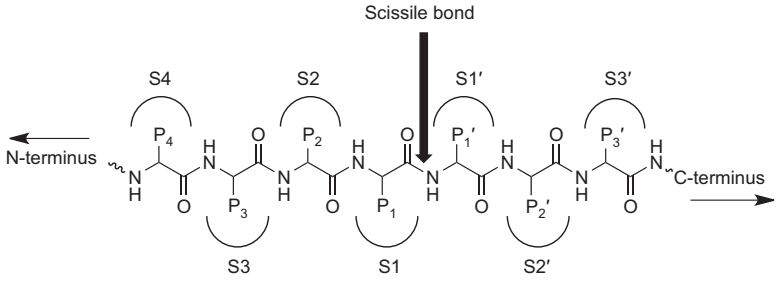

Figure 1 General architecture of a protease active site demonstrating binding pockets (S_n) and their assigned amino acid side chains (P_n) from the peptidic substrate.

different in structure from the 20 proteinogenic natural amino acids encoded by the universal genetic code.

Non-proteinogenic amino acids play very important roles in peptide and drug discovery research. Good examples are amino acids used in the treatment of Parkinson's disease, e.g., L-DOPA (L-3,4-dihydroxyphenylalanine), or drugs used to alleviate high blood pressure symptoms, e.g., enalapril, which contains hPhe (L-homophenylalanine) (Figure 2).

Due to the need of proteinogenic amino acids to mimic substrate design, α-amino derivatives play the most important role. Application of non-proteinogenic amino acids in the combinatorial approach significantly extends the amount of information about hotspot binding preferences on the enzyme surface, and thus facilitates further structure-activity studies. Due to the quite high price and the lack of general methodologies in the design of substrate peptide libraries with non-proteinogenic amino acids, these compounds have not been used on a large scale in proteolytic enzyme studies. More examples of non-proteinogenic amino acid applications can be found in inhibitors or ABPs designed for proteases (Powers et al., 2002; Berger et al., 2006; Skinner-Adams et al., 2007; Mucha et al., 2010).

Non-proteinogenic amino acids in protease substrate specificity profiling

Aminopeptidases

Aminopeptidases are exopeptidases that release one amino acid at a time from the N-terminal end of a peptidic substrate. In humans, these enzymes play important roles in most cellular and extracellular events, and their upregulation or downregulation is observed in pathogenic disorders, such as tumor growth, cancer, parasite infections, and inflammation (Hooper and Lendeckel, 2004). Aminopeptidases are present both in prokaryotic and eukaryotic organisms and can have either very high substrate specificity toward proteinogenic amino acids (e.g., methionine aminopeptidases, which recognize only methionine) or can tolerate quite diverse proteinogenic amino acid side chains in the S1 pocket [aminopeptidase N (APN), leucine aminopeptidase (LAP)] (Rawlings et al., 2011).

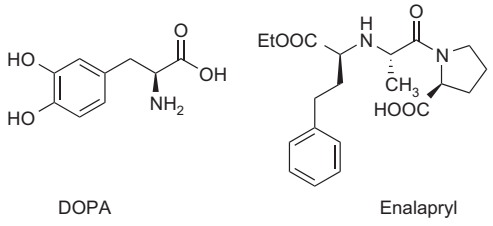

Figure 2 Structures of drugs containing non-proteinogenic amino acids in the scaffold.

The application of non-proteinogenic amino acids on a large scale in substrate specificity profiling of aminopeptidases was reported for the first time by Drag et al. (2010) for human, rat, and pig orthologs of APN (CD13). In this approach, a library of 61 individual fluorogenic substrates containing the reporter group 7-amino-4-carbamoylmethylcoumarin (ACC) was obtained by an Fmoc-based solid-phase peptide synthesis method (Maly et al., 2002). Besides the 19 proteinogenic amino acids (cysteine was omitted owing to problems with oxidation), 42 non-proteinogenic amino acids were chosen in consideration of their diverse functional groups, covering a broad range of interactions in the S1 pocket of aminopeptidases.

The primary requirement of metallo-aminopeptidases for free a N-terminal amino group at the end of the peptide sequence is often forced by a GAMEN motif (Gly-Ala-Met-Glu-Asn, where Gly – L-glycine, Ala – L-alanine, Met – L-methionine, Glu – L-glutamic acid, and Asn – L-asparagine) (Luciani et al., 1998; Ito et al., 2006). The Glu residue plays a particularly important role in this sequence, interacting with the free α-N-terminal amino group of the substrate, and therefore the majority of compounds in the library contain free alpha amino group. However, substrates with other functional groups like β-Ala (β-alanine), 6-Ahx (6-aminohexanoic acid; amine group in a position other than the α-position), Tic (tetrahydroisoquinoline-1-carboxylic acid; secondary amine), or Apns (3-amino-2-hydroxy-4-phenyl-butanoic acid; α-amino group replaced by a hydroxy group) were also used to determine the influence of an α-amino group on substrate binding.

Substrate specificity studies revealed that all three mammalian orthologs of APN have very similar substrate specificity patterns and tolerate several proteinogenic and non-proteinogenic amino acids in the S1 pocket. An interesting observation

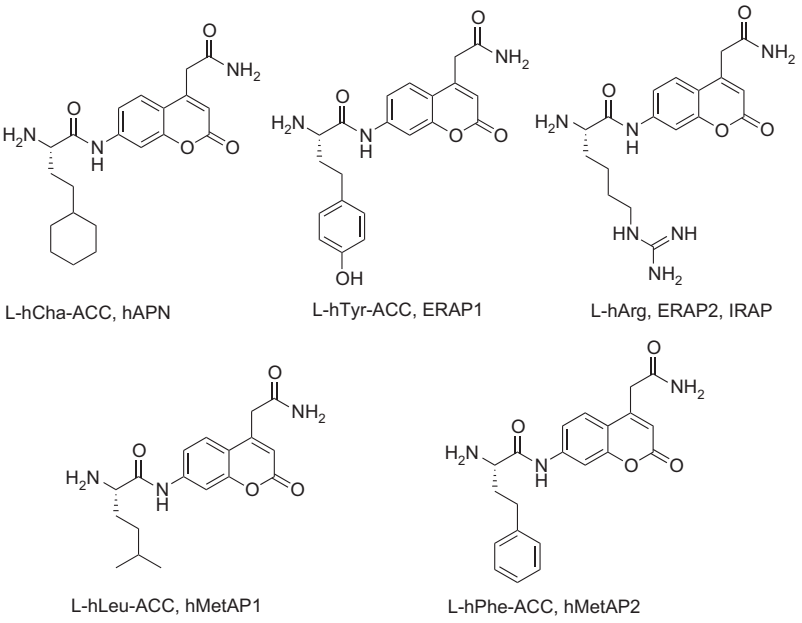

Figure 3 Non-proteinogenic amino acid-based fluorogenic substrates that are preferentially recognized by human aminopeptidases.

revealed that methionine was a better substrate than alanine (an alternate name for APN is alanine aminopeptidase). Several non-proteinogenic amino acids were almost equal to the best proteinogenic amino acid – methionine. The most preferred non-proteinogenic amino acids were those with rather large hydrophobic side chains like hCha (L-homocyclohexylalanine), styryl-Ala (L-styrylalanine), hArg (L-homoarginine), Nle (L-norleucine) and hPhe. Substrates without α-amines were not recognized by the tested aminopeptidases.

This approach has been applied to substrate specificity profiling of other human (ERAP1, ERAP2, IRAP, MetAP-1, MetAP-2), bacterial (*Escherichia coli* MetAP), malaria (PfM1AAP and PfM17LAP), and plant (oat seeds) aminopeptidases, and even in the detection of total aminopeptidase activity in malaria cell lysates (3D7 strain of *Plasmodium falciparum*) (Figure 3) (Gajda et al., 2012; Poreba et al., 2012a,b). In the case of ERAP1, ERAP2, and IRAP, fluorogenic substrate libraries were additionally extended by D-enantiomer derivatives of almost all proteinogenic amino acids (Zervoudi et al., 2011). The absence of observable cleavage of these compounds (minimal D-arginine processing by ERAP2 was observed) confirmed the high demand of these enzymes for L-enantiomers. Analysis of the described data reveals the utility of single non-proteinogenic amino acids connected to the ACC flurophore for the determination of important enzyme features, such as function, shape of the S1 pocket, and distinctions from other closely related enzymes.

Dipeptidyl peptidases

Dipeptidyl peptidases sequentially remove N-terminal dipeptides from peptides and proteins. Non-proteinogenic amino acids used for substrate specificity profiling of this group of enzymes focused initially on cathepsin C (DPP-IV), which is a cysteine protease involved in the activation of hematopoietic serine proteases, such as granzyme A and B, chymase, cathepsin G (CatG), or neutrophil elastase (Turk et al., 2001). Importantly, deficiency of DPP-IV leads to Haim-Munk syndrome or Papillon-Lefevre syndrome (Hart et al., 2000).

Li et al. (2009) synthesized dipeptide derivatives attached to a rhodamine fluorophore and monitored cathepsin C proteolytic activity in live cells by flow cytometry assay (FACS). In this approach, a collection of 16 different individual dipeptidic substrates containing proteinogenic and non-proteinogenic amino acids were synthesized. The best substrate identified was the diamide derivative $(Abu\text{-}hPhe)_2$-Rd containing hPhe in the P1 position and Abu (1-aminobutyric acid) in P2 (Figure 4).

Endopeptidases

Endopeptidases, which hydrolyze peptide bonds within the protein substrate, have also been a subject of substrate

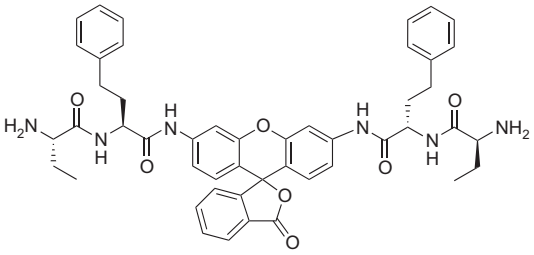

Figure 4 Potent rhodamine-based substrate $(Abu\text{-}hPhe)_2$-Rd for cathepsin C based on non-proteinogenic amino acids.

specificity studies using non-proteinogenic amino acids. The majority of these approaches describe the synthesis of various individual substrates containing a mixture of proteinogenic and non-proteinogenic amino acids. Initially, two types of substrates are used. One type uses a fluorogenic group attached to the C-terminal to the P1 position and therefore only covers the non-primed side of the substrate-binding cleft. The other type (fluorescently quenched peptides) may allow interaction with both the primed and non-primed sites. To date, there are no detailed descriptions of combinatorial approaches for substrate library design for exopeptidases that employ a large array of non-proteinogenic amino acids. In the classic positional scanning-synthetic combinatorial library (PS-SCL) approach described originally for caspases by Thornberry et al. (1997), the only non-proteinogenic amino acids used were D-Ala (D-alanine) and Nle, the latter applied instead of oxidation-prone methionine.

Matrix metallopeptidases are endopeptidases with a zinc ion responsible for the catalytic mechanism of action. They are involved in tissue remodeling processes and in the control of important cell functions, such as proliferation, differentiation, or apoptosis. MMPs are also involved in the development of various pathological disorders, such as cancer or arthritis (Overall and Lopez-Otin, 2002).

To explore the shape of the S1′ pocket of MMPs, substrates with non-proteinogenic amino acids in the P1′ position were synthesized by Mucha et al. (1998). The choice of the P1′ position was motivated by the importance of this pocket in primary substrate recognition and the need for the examination of the side chain length effect influencing MMP14 and ST1 activity. Heptapetides with P3-P4′ positions chosen from the canonical sequence characteristic for MMP formed the basis of this analysis. The N-terminal end was substituted with quenching Dns (dansyl), and as a fluorophore Trp (L-tryptophan) at the P2′ position was used. The substrate sequence design incorporated Dns-Pro-Leu-Ala-Î-Xaa-Trp-Ala-Arg-NH_2, where Î is the scissile peptide bond; Arg, L-arginine; Leu, L-leucine; Pro, L-proline; and Xaa is either Phe (L-phenylalanine), hPhe, pPhe (2-amino-5-phenylpentanoic acid), Ser(Bzl) (*O*-benzyl-L-serine), Cys(Bzl) (*S*-benzyl-L-cysteine), Cys(OMeBzl) (*S-para*-methoxybenzyl L-cysteine), Leu, Nle, Met, nPent (2-amine-pentanoic acid), or nHex (2-amine-hexanoic acid).

Analysis of substrate cleavage efficiency clearly indicated the dominant role of the P1′ amino acid side chain length. Application of longer phenylalanine derivatives like hPhe and pPhe significantly improved k_{cat}/K_M values. In general, peptides containing non-proteinogenic amino acids with very long side chains were much better substrates compared with substrates with proteinogenic amino acids. One of the best substrates of MT1-MMP and ST3 proteases found in this work was a peptide with a Cys(OMeBzl) residue in the P1′ position.

In another approach, the detailed substrate specificity of the S2 to S2′ pockets of gelatinase MMP-9 and collagenase MMP-1 was explored by McGeehan et al. (1994), employing four mixtures of substrates in which proteinogenic and non-proteinogenic amino acids were incorporated into the sequence. This resulted in 88 unique substituents at every position, yielding a total of 352 individual substrate sequences. Hydrolysis of substrates, investigated by HPLC and mass spectrometry, revealed several non-proteinogenic amino acids that were more efficiently processed by MMP-1 and MMP-9 compared with substrates with only proteinogenic amino acids in the peptide sequence. Interestingly, none of the D-amino acids were recognized by these MMPs.

The serine proteases proteinase 3 (PR3), human neutrophil elastase (HNE), and CatG are present in an active form in neutrophils. PR3 participates in immune defense by processing cytokines and degrading the extracellular matrix, and also has antimicrobial actions. PR3 can lead to autoimmune disorders, such as Wegener's granulomatosis – characterized by the presence of anti-PR3 antibodies, anti-neutrophil cytoplasmic antibodies (ANCA), in the patient's serum. High levels of ANCA increase neutrophil activation and the progression of disease symptoms.

To find the difference between PR3 and HNE at the P1′ position, Wysocka et al. (2008) synthesized a library of FRET-based substrates [aminobenzoic acid (ABZ) and 5-amino-2-nitrobenzoic acid (ANB-NH_2) were used as donor and acceptor, respectively]. As a result of this library analysis, a substrate containing the non-proteinogenic amino acid Abu in the peptide structure ABZ-Tyr-Tyr-Abu-ANB-NH_2 (where Tyr – L-tyrosine) was identified as the most optimal in terms of kinetic parameters for PR3 (k_{cat}/K_M=189×10^3 M^{-1} s^{-1}). To further optimize the substrate sequence, Wysocka et al. (2010) developed another substrate library, which was used to determine the optimal specificity pattern in prime sites. Using previously optimized substrates containing the non-proteinogenic amino acid Abu – ABZ-Tyr-Tyr-Abu-ANB-X-NH_2 (X=proteinogenic amino acid), it was found that glutamine is preferentially recognized in the investigated prime position (k_{cat}/K_M=275×10^3 M^{-1} s^{-1}) (Figure 5) (Wysocka et al., 2010).

CatG preferentially recognizes the bulky and hydrophobic amino acids Phe and Tyr at P1. The primary substrate specificity of this enzyme in the P1-P4 positions was determined using proteinogenic amino acids in chromophore-based library analyses (Wysocka et al., 2007). On the basis of these data, the leading sequence, Phe-Val-Thr-Tyr-Anb^{5,2}-NH_2 (where Thr – L-threonine and Val – L-valine), was found. To investigate in more detail the substrate specificity at P1, a second generation of the library with the general structure Ac-Phe-Val-Thr-X-Anb^{5,2}-NH_2 was synthesized using proteinogenic and non-proteinogenic amino acids [in the X (P1) position: Pal (L-pyridyl-alanine), Nif (*p*-nitro-L-phenylalanine), Phe(*p*-NH_2) (*p*-amino-L-phenylalanine), Cbf (*p*-carboxy-L-

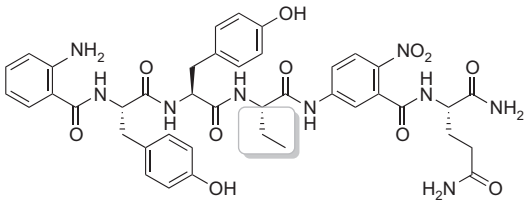

Figure 5 Optimal substrate ABZ-Tyr-Tyr-Abu-Anb^{5,2}-Gln for PR3 with 2-aminobutyric acid in the P1 position.

phenylalanine), Phe(*p*-guan) (*p*-guanidine-L-phenylalanine), Mcf (*p*-methyloxycarbonyl-L-phenylalanine), Phe(*p*-CN) (*p*-cyano-L-phenylalanine), Phe, Tyr, Arg, and Lys (L-lysine)] (Wysocka et al., 2007). All substrates were recognized by CatG except those containing Mcf, Pal, and Cbf at the P1 position, and data revealed that the specificity rate (k_{cat}/K_M) for a substrate with the non-proteinogenic amino acid Phe(*p*-guan) at P1 was about 12 times higher ($k_{cat}/K_M=95.3\times10^3$ M^{-1} s^{-1}) than the specificity rate for the best proteinogenic amino acid – Phe ($k_{cat}/K_M=7.9\times10^3$ M^{-1} s^{-1}). Furthermore, the substrate specificity constant for the most widely used CatG substrate, Suc-Phe-Phe-Pro-Ala-*p*NA ($k_{cat}/K_M=1.1\times10^3$ M^{-1} s^{-1}) (where Suc – succinyl residue, *p*-NA – *p*-nitroaniline), was more than seven times lower compared with the new substrate containing only proteinogenic amino acids, Ac-Phe-Val-Thr-Phe-Anb^{5,2}-NH_2 ($k_{cat}/K_M=7.9\times10^3$ M^{-1} s^{-1}), and about 90 times lower compared with the substrate with non-proteinogenic amino acids, Ac-Phe-Val-Thr-[Phe(p-guan)]-Anb^{5,2}-NH_2 ($k_{cat}/K_M=95.5\times10^3$ M^{-1} s^{-1}) (Figure 6). Finally, Ac-Phe-Val-Thr-Gnf-Anb^{5,2}-NH_2 was found to be recognized by CatG, while PR3, HNE, and chymotrypsin did not recognize this substrate.

Trypsin is one of the most well-studied serine proteases with very high specificity for basic amino acids (Arg, Lys) in the S1 pocket. To further optimize S1 pocket specificity, Lesner et al. (2001) developed a library of chromogenic substrates. On the basis of previous research, Ac-Ala-Val-Abu-Pro-X-*p*NA was chosen as a leading structure, where at the P1 position (X) two types of amino acids were used. The first group consisted of positively charged amino acids like Orn (L-ornithine), Lys, Arg, hArg, and Arg(NO_2) (*N*′-nitro-L-arginine), while in a second group amino acids that potentially form hydrogen bonds in enzyme-substrate interactions, Cit (L-citrulline), Hci (L-homocitrulline), Phe(*p*-CN), and Phe(*p*-NH_2), were used. None of the tested non-proteinogenic amino acids were better than Arg and Lys, and the activity order based on the k_{cat}/K_M value was as follows (lowest to highest): Phe(*p*-CN), Arg(NO_2), Phe(*p*-NH_2), hArg, Lys, and Arg. Substrates with Cit, Hci, and Orn in the P1 position were completely inactive toward trypsin.

Dengue fever flavivirus produces the NS2BNS3 protease, which is one of the enzymes crucial for the biological development of the disease. To investigate the substrate specificity in the S2 pocket of this enzyme, Gouvea et al. (2007) designed and synthesized a collection of fluorogenic dipeptide substrates with the general formula Bz-X-Arg-MCA (X represents a basic amino acid with an amine or guanidine substituent attached to an aromatic or aliphatic group). The non-proteinogenic amino acids used to optimize the P2 pocket were Ama (*trans*-4-aminomethylcyclohexyl-alanine), Amf (4-aminomethyl-phenylalanine), Phe(*p*-guan), Iaf (4-amino-methyl-*N*-isopropyl-L-phenylalanine), Pya (3-pyridyl-L-alanine), Ppa (4-piperidinyl-alanine), and Aca (4-amino-methyl-cycloheksyl-alanine). Kinetic analyses revealed that the non-proteinogenic amino acids Amf, Ama, and Aca are much better substrates in terms of the k_{cat}/K_M value (0.50, 0.58, and 0.48 mM^{-1} s^{-1}, respectively) compared with the proteinogenic amino acid Arg ($k_{cat}/K_M=0.44$ mM^{-1} s^{-1}) (Figure 7). Substrates with Phe(p-guan), Iaf, Pya, and Ppa were poorly hydrolyzed ($k_{cat}/K_M=0.07$, 0.06, 0.05, and 0.07 mM^{-1} s^{-1}, respectively) by the enzyme.

Hepatitis C virus (HCV), which is responsible for posttransfusion and community-acquired hepatitis type C, produces the NS3 protease, a non-structural protein involved in virus growth, considered as a major target in anti-HCV therapy. Substrate specificity studies with proteinogenic amino acids revealed an essential influence of the P1 and P3 positions on the overall substrate specificity of this enzyme. To investigate in more detail the S1 pocket binding preferences, a set of proteinogenic and non-proteinogenic amino acids similar in structure to cysteine was chosen [hCys (L-homocysteine), Cys(Me) (*S*-methyl L-cysteine), Ala, Ecy (*S*-ethyl L-cysteine), Thr, Met, D-Cys (D-cysteine), Ser, and Pen (penicillamine)] (Zhang et al., 1997). These residues were introduced into the P1 (X) position of the modified NS5A/5B P8P8′K substrate with the parent sequence DTEDVVAX-SMSYTWTG-K-OH. HPLC analysis of substrate hydrolysis was used to determine kinetic parameters. Interestingly, the reported data revealed that none of the tested substrates was better compared with one with Cys at the P1 position, and that even small modifications lead to significant decrease of activity toward the enzyme. The lack of hydrolysis of D-Cys confirmed the high stereospecificity in the S1 pocket of NS3 protease.

Perspectives

Non-proteinogenic amino acids constitute a large group of compounds with great potential in substrate specificity

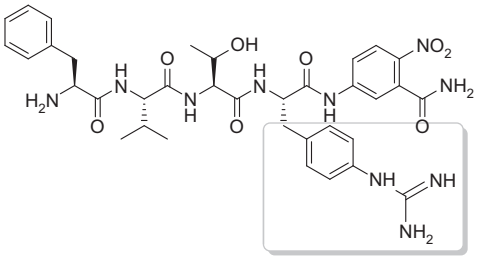

Figure 6 Optimal substrate Phe-Val-Thr-[Phe(p-guan)]-Anb^{5,2}-NH_2 for CatG with 4-guanidine-L-phenylalanine acid in the P1 position.

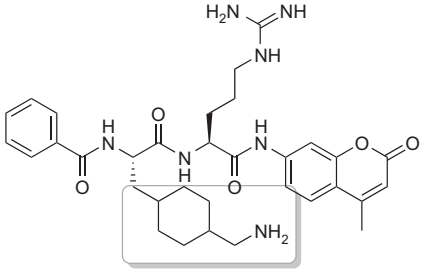

Figure 7 Optimal substrate Bz-Ama-Arg-MCA for NS2BNS3 protease with *trans*-4-aminomethylcyclohexyl-alanine acid in the P2 position.

profiling studies of all protease types. There is a great need for new types of substrates and substrate libraries that could help in differentiating individual members of the same families of proteolytic enzymes, such as caspases, deubiquitinating and deSUMOylating enzymes, MMPs, and cathepsins. To date, available technologies for application of non-proteinogenic amino acids in profiling of aminopeptidases or dipeptidyl peptidases are general enough to be applied to other members of amino-terminal exopeptidase families. At present, one of the biggest challenges is to use non-proteinogenic amino acids for investigations of endopeptidases. Currently, there are insufficiently developed technologies to allow facile incorporation of non-proteinogenic amino acids into combinatorial substrate libraries to enable synthesis of tetrapeptides or longer peptides for substrate specificity studies. The major problem in the synthesis of these libraries is the design of the isokinetic mixtures containing both proteinogenic and non-proteinogenic amino acids, which would be needed to guarantee equimolar substitution of all amino acids from a mixture (Ostresh et al., 1994). There are no described general solutions to this problem, and the only known solution is to apply a 'mix and split' strategy. Regrettably, this approach is quite tedious on a small laboratory scale (Furka et al., 1991). Another solution would be to apply only proteinogenic amino acids in the isokinetic mixture and apply only non-proteinogenic amino acids in fixed positions, as depicted in Figure 8. Such a library is designed to optimize the P2, P3, and P4 positions with non-proteinogenic amino acids. A good prognosis for this type of solution is the successful application of the non-proteinogenic amino acid norleucine instead of methionine in PS-SCL. We expect that incorporation of any other non-proteinogenic amino acid should be equally good.

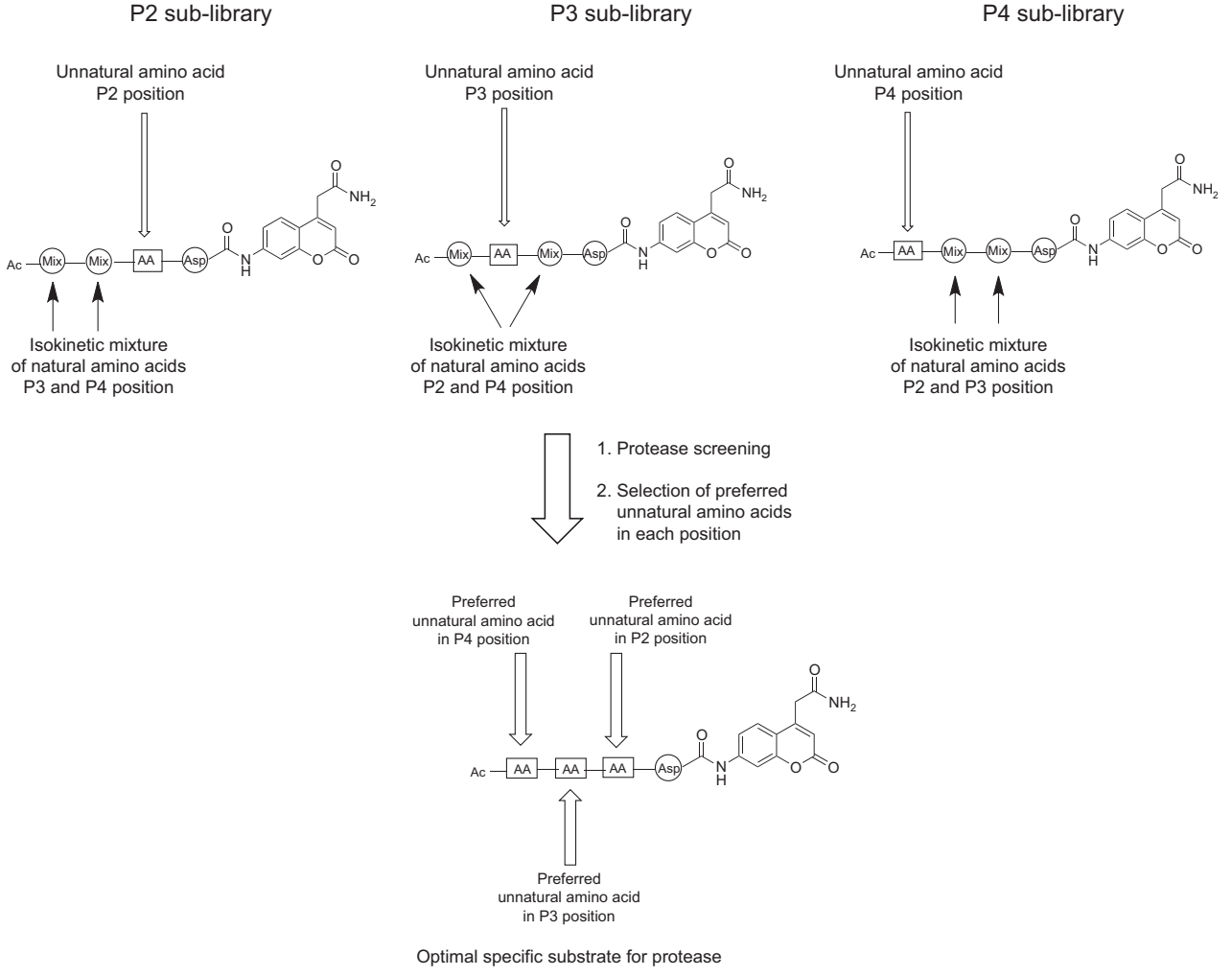

Figure 8 A concept of combinatorial substrate library design to screen for optimal non-proteinogenic amino acids in the P2, P3, and P4 positions for caspases.

Each P sublibrary could contain a certain number of individual fluorogenic substrate mixtures based on the proposed structure. Application of fixed aspartic acid in the P1 position will guarantee recognition of the substrate by caspase in regular screening. Subsequent combination of P1 Asp with optimal non-proteinogenic amino acids resulting from P2, P3, and P4 sublibrary screening may yield specific individual optimal substrates for the investigated caspase.

Another yet-to-be explored area of non-proteinogenic amino acids is their use for investigation of posttranslational modifications influencing protease substrate specificity. Keeping in mind that small structural differences influence the substrate specificity of proteolytic enzymes, one could imagine that occupation of Sn pockets by methylation of glutamic or aspartic acid side chains, hydroxylation of Pro, or oxidation of Met could result in large changes in binding preferences (Figure 9). This fascinating area of protease research has not been rationally investigated yet and opens a new field for application of non-proteinogenic amino acids.

Results from library screens containing non-proteinogenic amino acids could also be widely used in the design of new, specific substrates for individual proteases. This would greatly facilitate further research of certain protease families for which such substrates are currently unavailable. Additionally, inhibitors and ABPs based on substrate recognition patterns could also be designed and synthesized (Figure 10). Such approaches with proteinogenic amino acids have already been investigated and resulted in the selection of new low molecular weight molecules, which currently are invaluable tools in protease research. For example, Choe et al. (2006) used PS-SCL of fluorogenic tetrapeptides to determine the substrate specificity profile of proteinogenic amino acids in the S1-S4 pockets of the cysteine proteases cathepsins L, V, K, S, F, and B, as well as papain and bromelain. This resulted in identification of a high preference for proline in the S2 pocket by cathepsin K. Subsequent deconvolution of the library resulted in the peptide sequence Ac-HGPR-ACC, which was recognized only by cathepsin K, while other cathepsins used in the studies did not hydrolyze this substrate. Finally, an optimal and highly specific cathepsin K substrate sequence was used to design a specific inhibitor of this enzyme by a simple exchange of the ACC substrate fluorogenic group by the inhibitory acyloxymethyl ketone group (Choe et al., 2006). In another approach, Mahrus and Craik (2005) determined substrate specificity in the S1-S4 pockets of human serine proteases from the granzyme family (granzyme A, B, H, K, and M), also using tetrapeptidic fluorogenic PS-SCL. The optimal sequences from these studies were used for the design and synthesis of phosphonate ABPs for granzyme A and B. In the case of granzyme A, in the P1 position of the ABP, arginine was exchanged with the non-proteinogenic amino acid *p*-guanidine-α-phenylglycine. This approach resulted in the selection of ABPs for granzyme A and B that are very specific toward proteases investigated in cell-based experiments (Mahrus and Craik, 2005). The application of PS-SCL libraries aided in the discovery of differences between closely related proline-specific dipeptidases (DPP-II, -IV, and -VII), and this information was employed for drug design of dipeptidyl peptidase IV, involved diabetes type II (Leiting et al., 2003). Substrate specificity data obtained for both proteinogenic and non-proteinogenic amino acids were also used by Drag and colleagues to design optimal phosphonate inhibitors of human and pig APN (Drag et al., 2010; Grzywa et al., 2010). Due to the susceptibility of the peptide bond to the hydrolysis by proteases, it is unlikely that substrate specificity studies performed using either proteinogenic or non-

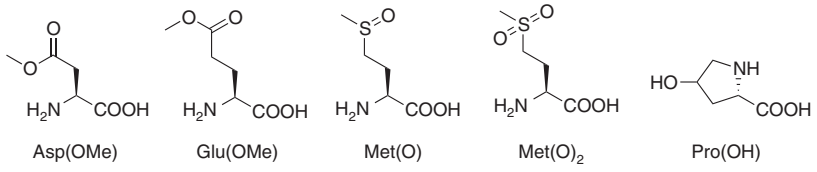

Figure 9 Examples of posttranslationally modified amino acids.

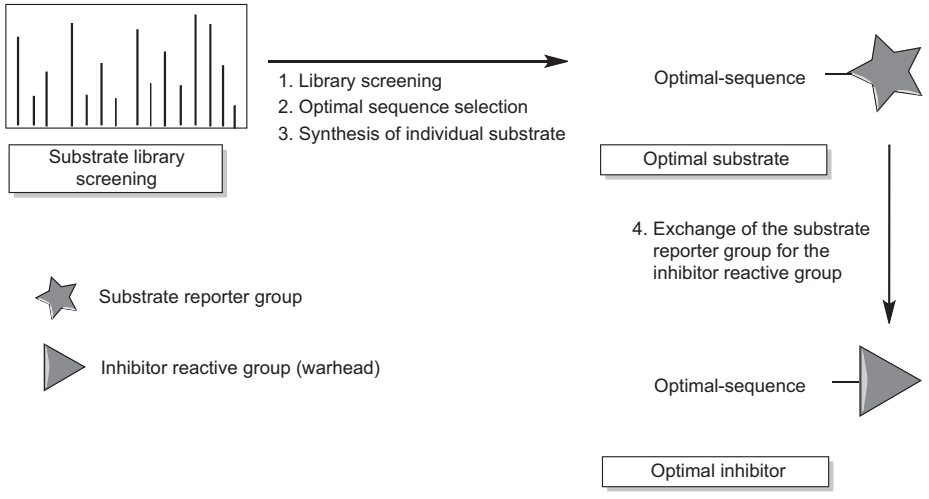

Figure 10 A general scheme for protease inhibitor design based on information from substrate specificity screening.

proteinogenic amino acid peptide-like molecules will yield ideal drug candidates after incorporating an electrophile, like an aldehyde or a phosphonate, at the position of the reporter group. However, this information can be invaluable for the design of molecules used by scientists to address questions related to the physiological function of proteases, as in the case of the above examples of inhibitor and ABP designs. Moreover, examples from the past demonstrate that, in some cases, substrate specificity may also help in drug design as in the case of DPP-IV protease or the dipeptidic proteasome inhibitor – Velcade (Adams and Kauffman, 2004).

Conclusions

For a long time, non-proteinogenic amino acids have attracted considerable interest from medicinal chemists as very attractive scaffolds in the design of molecules with potential biological activity. The variety of structural modifications is theoretically unlimited and the universe of new, interesting structures is growing rapidly. Unfortunately, their application in the design of synthetic substrates for proteases in the past was rather modest. Currently, we are in urgent need of new synthetic approaches that could be applied to all families of proteolytic enzymes. The examples presented here serve as good indicators for further development of methodologies for the synthesis of new types of combinatorial and focused substrate libraries to investigate proteases. This would significantly broaden our understanding of this group of enzymes and may help design more specific and active substrates, inhibitors, and ABPs.

Acknowledgments

The Drag laboratory is supported by the Foundation for Polish Science and the State for Scientific Research Grant N N401 042838 in Poland. Paulina Kasperkiewicz and Anna Gajda are supported by the European Union Human Capital National Cohesion Strategy. We would like to thank Guy S. Salvesen for critical reading of the manuscript.

References

Adams, J. and Kauffman, M. (2004). Development of the proteasome inhibitor Velcade (Bortezomib). Cancer Invest. *22*, 304–311.

Backes, B.J., Harris, J.L., Leonetti, F., Craik, C.S., and Ellman, J.A. (2000). Synthesis of positional-scanning libraries of fluorogenic peptide substrates to define the extended substrate specificity of plasmin and thrombin. Nat. Biotechnol. *18*, 187–193.

Berger, A.B., Witte, M.D., Denault, J.B., Sadaghiani, A.M., Sexton, K.M., Salvesen, G.S., and Bogyo, M. (2006). Identification of early intermediates of caspase activation using selective inhibitors and activity-based probes. Mol. Cell *23*, 509–521.

Choe, Y., Leonetti, F., Greenbaum, D.C., Lecaille, F., Bogyo, M., Bromme, D., Ellman, J.A., and Craik, C.S. (2006). Substrate profiling of cysteine proteases using a combinatorial peptide library identifies functionally unique specificities. J. Biol. Chem. *281*, 12824–12832.

Drag, M. and Salvesen, G.S. (2010). Emerging principles in protease-based drug discovery. Nat. Rev. Drug Discov. *9*, 690–701.

Drag, M., Mikolajczyk, J., Krishnakumar, I.M., Huang, Z., and Salvesen, G.S. (2008). Activity profiling of human deSUMOylating enzymes (SENPs) with synthetic substrates suggests an unexpected specificity of two newly characterized members of the family. Biochem. J. *409*, 461–469.

Drag, M., Bogyo, M., Ellman, J.A., and Salvesen, G.S. (2010). Aminopeptidase fingerprints, an integrated approach for identification of good substrates and optimal inhibitors. J. Biol. Chem. *285*, 3310–3318.

Furka, A., Sebestyen, F., Asgedom, M., and Dibo, G. (1991). General method for rapid synthesis of multicomponent peptide mixtures. Int. J. Pept. Protein Res. *37*, 487–493.

Gajda, A.D., Pawelczak, M., and Drag, M. (2012). Substrate specificity screening of oat (*Avena sativa*) seeds aminopeptidase demonstrate unusually broad tolerance in S1 pocket. Plant Physiol. Biochem. *54C*, 6–9.

Gouvea, I.E., Izidoro, M.A., Judice, W.A., Cezari, M.H., Caliendo, G., Santagada, V., dos Santos, C.N., Queiroz, M.H., Juliano, M.A., Young, P.R., et al. (2007). Substrate specificity of recombinant dengue 2 virus NS2B-NS3 protease: influence of natural and unnatural basic amino acids on hydrolysis of synthetic fluorescent substrates. Arch. Biochem. Biophys. *457*, 187–196.

Grzywa, R., Oleksyszyn, J., Salvesen, G.S., and Drag, M. (2010). Identification of very potent inhibitor of aminopeptidase N (CD13). Bioorg. Med. Chem. Lett. *20*, 2497–2499.

Hart, T.C., Hart, P.S., Michalec, M.D., Zhang, Y., Firatli, E., Van Dyke, T.E., Stabholz, A., Zlotogorski, A., Shapira, L., and Soskolne, W.A. (2000). Haim-Munk syndrome and Papillon-Lefevre syndrome are allelic mutations in cathepsin C. J. Med. Genet. *37*, 88–94.

Hooper, N.M. and Lendeckel, U., eds. (2004). Aminopeptidases in Biology and Disease, Series: Proteases in Biology and Disease, Vol. 2. (New York, USA: Kluwer Academic/Plenum Publishers).

Ito, K., Nakajima, Y., Onohara, Y., Takeo, M., Nakashima, K., Matsubara, F., Ito, T., and Yoshimoto, T. (2006). Crystal structure of aminopeptidase N (proteobacteria alanyl aminopeptidase) from *Escherichia coli* and conformational change of methionine 260 involved in substrate recognition. J. Biol. Chem. *281*, 33664–33676.

Leiting, B., Pryor, K.D., Wu, J.K., Marsilio, F., Patel, R.A., Craik, C.S., Ellman, J.A., Cummings, R.T., and Thornberry, N.A. (2003). Catalytic properties and inhibition of proline-specific dipeptidyl peptidases II, IV and VII. Biochem. J. *371*, 525–532.

Lesner, A., Kupryszewski, G., and Rolka, K. (2001). Chromogenic substrates of bovine β-trypsin: the influence of an amino acid residue in P1 position on their interaction with the enzyme. Biochem. Biophys. Res. Commun. *285*, 1350–1353.

Li, J., Petrassi, H.M., Tumanut, C., Masick, B.T., Trussell, C., and Harris, J.L. (2009). Substrate optimization for monitoring cathepsin C activity in live cells. Bioorg. Med. Chem. *17*, 1064–1070.

Luciani, N., Marie-Claire, C., Ruffet, E., Beaumont, A., Roques, B.P., and Fournie-Zaluski, M.C. (1998). Characterization of Glu350 as a critical residue involved in the N-terminal amine binding site of aminopeptidase N (EC 3.4.11.2): insights into its mechanism of action. Biochemistry *37*, 686–692.

Mahrus, S. and Craik, C.S. (2005). Selective chemical functional probes of granzymes A and B reveal granzyme B is a major effector of natural killer cell-mediated lysis of target cells. Chem. Biol. *12*, 567–577.

Maly, D.J., Leonetti, F., Backes, B.J., Dauber, D.S., Harris, J.L., Craik, C.S., and Ellman, J.A. (2002). Expedient solid-phase synthesis of fluorogenic protease substrates using the 7-amino-

4-carbamoylmethylcoumarin (ACC) fluorophore. J. Org. Chem. *67*, 910–915.

McGeehan, G.M., Bickett, D.M., Green, M., Kassel, D., Wiseman, J.S., and Berman, J. (1994). Characterization of the peptide substrate specificities of interstitial collagenase and 92-kDa gelatinase. Implications for substrate optimization. J. Biol. Chem. *269*, 32814–32820.

McStay, G.P., Salvesen, G.S., and Green, D.R. (2008). Overlapping cleavage motif selectivity of caspases: implications for analysis of apoptotic pathways. Cell Death Differ. *15*, 322–331.

Mucha, A., Cuniasse, P., Kannan, R., Beau, F., Yiotakis, A., Basset, P., and Dive, V. (1998). Membrane type-1 matrix metalloprotease and stromelysin-3 cleave more efficiently synthetic substrates containing unusual amino acids in their P1′ positions. J. Biol. Chem. *273*, 2763–2768.

Mucha, A., Drag, M., Dalton, J.P., and Kafarski, P. (2010). Metalloaminopeptidase inhibitors. Biochimie *92*, 1509–1529.

Ostresh, J.M., Winkle, J.H., Hamashin, V.T., and Houghten, R.A. (1994). Peptide libraries – determination of relative reaction rates of protected amino acids in competitive couplings. Biopolymers *34*, 1681–1689.

Overall, C.M. and Lopez-Otin, C. (2002). Strategies for MMP inhibition in cancer: innovations for the post-trial era. Nat. Rev. Cancer *2*, 657–672.

Poreba, M. and Drag, M. (2010). Current strategies for probing substrate specificity of proteases. Curr. Med. Chem. *17*, 3968–3995.

Poreba, M., Gajda, A., Picha, J., Jiracek, J., Marschner, A., Klein, C.D., Salvesen, G.S., and Drag, M. (2012a). S1 pocket fingerprints of human and bacterial methionine aminopeptidases determined using fluorogenic libraries of substrates and phosphorus based inhibitors. Biochimie *94*, 704–710.

Poreba, M., McGowan, S., Skinner-Adams, T.S., Trenholme, K.R., Gardiner, D.L., Whisstock, J.C., To, J., Salvesen, G.S., Dalton, J.P., and Drag, M. (2012b). Fingerprinting the substrate specificity of M1 and M17 aminopeptidases of human malaria, *Plasmodium falciparum*. PLoS One *7*, e31938.

Powers, J.C., Asgian, J.L., Ekici, O.D., and James, K.E. (2002). Irreversible inhibitors of serine, cysteine, and threonine proteases. Chem. Rev. *102*, 4639–4750.

Rano, T.A., Timkey, T., Peterson, E.P., Rotonda, J., Nicholson, D.W., Becker, J.W., Chapman, K.T., and Thornberry, N.A. (1997). A combinatorial approach for determining protease specificities: application to interleukin-1β converting enzyme (ICE). Chem. Biol. *4*, 149–155.

Rawlings, N.D., Barrett, A.J., and Bateman, A. (2011). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. *40*, D343–D350.

Schechter, I. and Berger, A. (1967). On the size of the active site in proteases. I. Papain. Biochem. Biophys. Res. Commun. *27*, 157–162.

Schilling, O. and Overall, C.M. (2007). Proteomic discovery of protease substrates. Curr. Opin. Chem. Biol. *11*, 36–45.

Schneider, E.L. and Craik, C.S. (2009). Positional scanning synthetic combinatorial libraries for substrate profiling. Methods Mol. Biol. *539*, 59–78.

Skinner-Adams, T.S., Lowther, J., Teuscher, F., Stack, C.M., Grembecka, J., Mucha, A., Kafarski, P., Trenholme, K.R., Dalton, J.P., and Gardiner, D.L. (2007). Identification of phosphinate dipeptide analog inhibitors directed against the *Plasmodium falciparum* M17 leucine aminopeptidase as lead antimalarial compounds. J. Med. Chem. *50*, 6024–6031.

Thornberry, N.A., Rano, T.A., Peterson, E.P., Rasper, D.M., Timkey, T., Garcia-Calvo, M., Houtzager, V.M., Nordstrom, P.A., Roy, S., Vaillancourt, J.P., et al. (1997). A combinatorial approach defines specificities of members of the caspase family and granzyme B. Functional relationships established for key mediators of apoptosis. J. Biol. Chem. *272*, 17907–17911.

Turk, B. (2006). Targeting proteases: successes, failures and future prospects. Nat. Rev. Drug. Discov. *5*, 785–799.

Turk, D., Janjic, V., Stern, I., Podobnik, M., Lamba, D., Dahl, S.W., Lauritzen, C., Pedersen, J., Turk, V., and Turk, B. (2001). Structure of human dipeptidyl peptidase I (cathepsin C): exclusion domain added to an endopeptidase framework creates the machine for activation of granular serine proteases. EMBO J. *20*, 6570–6582.

Wysocka, M., Legowska, A., Bulak, E., Jaskiewicz, A., Miecznikowska, H., Lesner, A., and Rolka, K. (2007). New chromogenic substrates of human neutrophil cathepsin G containing non-natural aromatic amino acid residues in position P(1) selected by combinatorial chemistry methods. Mol. Divers. *11*, 93–99.

Wysocka, M., Lesner, A., Guzow, K., Mackiewicz, L., Legowska, A., Wiczk, W., and Rolka, K. (2008). Design of selective substrates of proteinase 3 using combinatorial chemistry methods. Anal. Biochem. *378*, 208–215.

Wysocka, M., Lesner, A., Majkowska, G., Legowska, A., Guzow, K., Rolka, K., and Wiczk, W. (2010). The new fluorogenic substrates of neutrophil proteinase 3 optimized in prime site region. Anal. Biochem. *399*, 196–201.

Zervoudi, E., Papakyriakou, A., Georgiadou, D., Evnouchidou, I., Gajda, A., Poreba, M., Salvesen, G.S., Drag, M., Hattori, A., Swevers, L., et al. (2011). Probing the S1 specificity pocket of the aminopeptidases that generate antigenic peptides. Biochem. J. *435*, 411–420.

Zhang, R., Durkin, J., Windsor, W.T., McNemar, C., Ramanathan, L., and Le, H.V. (1997). Probing the substrate specificity of hepatitis C virus NS3 serine protease by using synthetic peptides. J. Virol. *71*, 6208–6213.

Received March 23, 2012; accepted April 26, 2012