Comprehensive protease specificity profiling

Bo Zhu,[a], [b], [d], # Michael D. Lane,[a], [b], # , Joshua A. Baller,[c] Pratik D. Jagtap,[a] [c] Timothy J. Griffin,[a] and Burckhard Seelig*[a] [b]

[a] Department of Biochemistry, Molecular Biology and Biophysics
University of Minnesota, Minneapolis, Minnesota, 55455, USA.
* E-mail: seelig@umn.edu

[b] BioTechnology Institute
University of Minnesota, St. Paul, Minnesota, 55108, USA.

[c] Minnesota Supercomputing Institute
University of Minnesota, Minneapolis, Minnesota, 55455, USA.

[d] Current address: Laboratory for Chemistry and Life Science, Institute of Integrated Research
Institute of Science Tokyo, Yokohama, Kanagawa, 226-8501, Japan.

# These authors contributed equally: B. Zhu and M. D. Lane

Abstract: Protease enzymes are of great importance in medicine, industry, and as research tools. Despite the crucial need for detailed knowledge of their proteolytic cleavage specificity, many proteases are poorly characterized. We present a method for fully characterizing the cleavage specificity of proteases through the comprehensive profiling of all possible permutations of octamer peptide substrates in a single experiment. The powerful combination of in vitro selection with high-throughput sequencing, mass spectrometry, and automated motif mining enabled the screening of mixtures of $>10^{12}$ unique peptides. We developed freely available software that easily integrates the massive amounts of cleavage data into user-friendly specificity information. We applied this method to three different proteases that had either narrow (factor Xa) or broad specificity (ADAM17 and streptopain). The resulting specificity maps revealed motifs that corroborate canonical known cleavage sites, yet step further into extended spectrum preferences and yield insights into the function of broad specificity proteases.

Introduction

Proteases perform a vast array of critical functions in the cell. These activities are dependent upon the highly-refined specificity of the protease, including evolved natural cleavage specificities and co-evolved cleavage sites[1]. Recent efforts have assessed broad-specificity proteases using human proteomic screens[2], designed peptide libraries[3], or photo-crosslinked native substrates to protease[4]. However, these methods have drawbacks that limit their ability to generate reliable specificity maps for broad specificity proteases, including low numbers of tested peptides (usually dozens, up to thousands), proteomic library bias, or the need for modified or unnatural amino acids in the cleaved substrates. To increase the throughput of protease cleavage site analyses, phage display was also used[5]. However, even a phage-displayed peptide library still can only test 4% of all possible octamer substrate sequences and the potential cleavage of the g3p protein that is crucial for displaying the substrates introduces bias. Furthermore, the reported method requires prior knowledge of the protease specificity to perform the cleavage site-guided motif generation[5c]. Here we present a method that overcomes these limitations by combining *in vitro* selection by mRNA display, high-throughput sequencing, and LC-MS/MS to increase the number of analyzed cleavage sites by several orders of magnitude, eliminate proteomic or modified amino acid biases, provide knowledge-free cleavage site-guided alignment, and enable screening of all possible substrate octamers in a single experiment without repeated selection rounds (Fig. 1).

Proteases typically recognize between 2-4 residues on either side of the cleaved bond and, therefore, their cleavage preferences can be challenging to elucidate[6]. To scan all possible permutations of eight residues (octamers), we used the mRNA display technology[7], which generated up to 10^{12} substrates and thereby oversampled all possible octamer substrates (20^8 = 26 billion) by > 30-fold. We applied this method to comprehensively assess the protease specificity of both C- and N- terminals to the cleaved bond. Subsequent high-throughput sequencing enabled us to detect up to millions

of cleavage sequences, which is a substantial increase over the best existing method. Finally, mass spectrometry (MS) identified the location of cleavage for a subset of preferred substrates and guided the alignment of all cleaved sequences to generate the final specificity map.

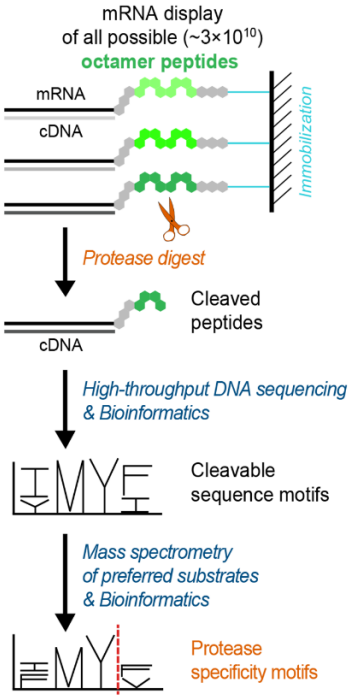

Figure 1. General scheme for comprehensive protease specificity profiling by mRNA display, high-throughput sequencing, LC-MS/MS, and automated motif mining. A library containing nearly all octamer peptide substrates is mRNA displayed, immobilized, and subsequently digested with a protease of interest. The cDNA-peptide fusions released from immobilization by protease cleavage of the respective peptide substrate are analyzed by high-throughput sequencing to identify the cleavable sequences. The cleavable sequence motifs are obtained through the motif mining computation pipeline. Highly enriched sequences from the cleavable sequence motifs are chemically synthesized and the cleavage position is identified by LC-MS/MS. These cleavable sequence data and the identified cleavage positions are merged to generate specificity motifs with cleavage positions. The orange dashed line indicates the cleavage position.

The key advantages of our method, in addition to the unique ability to sample all possible octamer substrates, are: (i) a substantial reduction of biases because our substrate library contained an even distribution of only natural amino acids and required no additional protease processing prior to selection; (ii) the simultaneous assessment of both prime and non-prime cleavage specificity; and (iii) the ability of our purely *in vitro* technology to freely assess protease function with respect to temperature, pH, ionic strength, or inhibitors[8]. Moreover, while lower-throughput methods might be sufficient to characterize narrow-specificity proteases, higher-throughput is needed to reveal the intricate details of broad-specificity proteases.

We applied this method to study a representative group of narrow- and broad-specificity, well- and poorly-characterized protease enzymes: factor Xa, ADAM17, and streptopain. The comprehensive high-resolution specificity maps generated with our method will lead to a better understanding of proteases in general, and broad-specificity proteases in particular.

Results and Discussion

Design of the comprehensive octamer peptide substrate library

We aimed to present all possible octamer substrates for cleavage. Through mRNA display, we generated the random octamer peptide library, where every individual peptide variant was linked at one end to its encoding mRNA, and at the other end to an immobilization tag (Fig. 2a). mRNA display also requires additional constant regions on each terminus of the peptide sequence that is to be displayed[9]. We incorporated an alkyne moiety at the conserved N-terminus to allow bioorthogonal immobilization via copper-catalyzed click chemistry[10]. This covalent immobilization to azide-agarose resin enabled stringent washes to remove nonspecific binders during the subsequent selection step. We used an orthogonal tRNA/aminoacyl-tRNA synthetase pair to incorporate the alkyne group specifically at an amber stop codon[11]. The entire translated peptide construct consisted of MG, alkyne moiety (Z), 6 × His (H) affinity tag, PQP frame around the random octamer substrate (X_8), and ended with MPTGY for mRNA display purposes (Fig. 2b). PQP was chosen as a frame because this motif is generally difficult for most proteases to cleave. We constructed the random octamer section of the library through chemical DNA synthesis using a mixture of nucleotide trimers encoding all 20 amino acids[12]. This strategy eliminated the risk of introducing stop codons and degeneracy bias, which can be an issue when using the more common NNS or NNK libraries (Supplementary Table 1). Through *in vitro* translation of this library, we were able to produce up to 10^{12} peptides as mRNA-displayed fusions, which oversamples the library of all possible octamers (2.6×10^{10}) by >30-fold. This comprehensive library of all octamer peptide substrates can be directly used for profiling the cleavage specificity of essentially any protease.

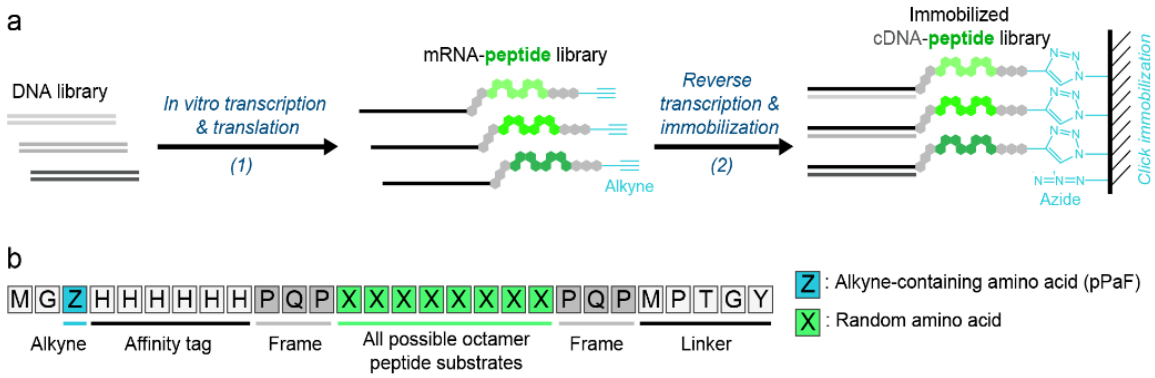

Figure 2. Overview of mRNA display and click immobilization of octamer peptide substrates for exploring protease specificity. (a) The strategy used for mRNA display and immobilization. (1) DNA library is transcribed to mRNA library, modified with puromycin, and in vitro translated by the PURE system to display each peptide as covalent fusion with its own encoding mRNA. The translated peptide is modified with an alkyne during translation via incorporation of the alkyne-containing non-natural amino acid pPaF. (2) The mRNA-peptide library is reverse transcribed, and the cDNA-peptide library is immobilized on agarose beads through Cu(I)-catalyzed alkyne-azide cycloaddition (CuAAC) click reaction. (b) Design of library containing all possible octamer peptide substrates with termini compatible for mRNA display.

Isolation of cleavable peptide substrates for three proteases by mRNA display

We chose three proteases comprising both narrow- and broad-specificity enzymes to validate our new method: factor Xa, ADAM17, and streptopain. Factor Xa is a narrow-specificity serine protease and a critical component of the coagulation cascade. The cleavage specificity of factor Xa has been well characterized by several methods[13]. ADAM17 is a broad-specificity transmembrane metalloprotease[14] that is involved in development, healthy physiology, and pathological mechanisms. Known targets of ADAM17 include Tumor Necrosis Factor alpha (TNFα) and substrates involved in tumor immune-surveillance, inflammation, and cancer, implicating the protease in diseases from rheumatoid arthritis and inflammatory bowel disease to large diffuse non-Hodgkin B-cell lymphoma[15]. The specificity of ADAM17 has been characterized by a peptide library approach[16], Q-PICS[17] and SPD-NGS[5c]. Finally, streptopain is a broad-specificity cysteine protease produced by *S. pyogenes*. Streptopain, also known as streptococcal pyrogenic exotoxin B (SpeB), has potent immunomodulating effects on the human immune system, cleaving multiple immunoglobulins,

critical components of both the classic and alternative pathways of complement activation, and numerous chemokines[18]. Further, streptopain can cleave a variety of host proteins unrelated to the immune system, such as fibrin/fibrinogen, plasmin, and host matrix metalloproteases, as well as a wide range of bacterial proteins[18]. Despite the importance of streptopain, its cleavage specificity has only been poorly characterized to date. Only about 25 quenched fluorescent peptide substrates have been used to probe positions P3-P1 and P1′, which resulted in a cleavage specificity map[19] that is entirely inadequate to explain streptopain's broad activity. The limited information on streptopain in the MEROPS database[6] (28 substrates) is insufficient to reliably predict its behavior even towards known cleavable substrates.[19]

We synthesized and immobilized a complete library of mRNA-displayed octamer substrate peptides to agarose beads via click reaction for each protease experiment (Fig. 2a), and washed the immobilized library with the respective digestion buffer. The subsequent incubation with protease cleaved only susceptible octamers and thereby released the cDNA/mRNA-fusions that encoded for those cleavable peptide substrates.

The released cDNA was subjected to high-throughput sequencing in order to analyze the region that encoded the cleavable protease substrates. In addition, we also prepared a control sample of the library of immobilized sequences, identical to the sequences immediately before incubation with protease. This sample enabled us to assess any potential bias that might occur during translation, transcription, or immobilization. To generate this control library, octamer library fusions were immobilized and washed identically to the above protocol, and the resin was used directly as a template for PCR amplification. The control library of peptides before cleavage and the original chemically synthesized library were also analyzed by high-throughput sequencing.

Analysis of high-throughput sequencing data to identify cleavable sequence motifs

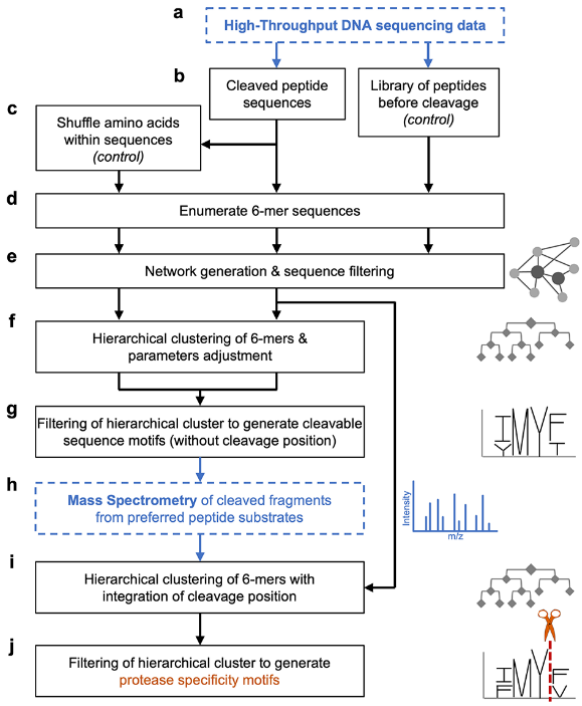

Figure 3. Computational motif mining pipeline. Automated processing of the high-throughput DNA sequencing data from the peptide substrate library before and after protease cleavage, and the subsequent incorporation of mass spectrometry cleavage data for preferred peptide substrates ultimately yields the protease specificity motifs for a given protease. The two data entry points are highlighted by dashed blue boxes. A flowchart with all steps in full detail can be found in Supplementary Fig. 1.

The flowchart in Fig. 3 shows the entire process of the protease specificity motif computation. Using this bioinformatics pipeline, we first analyzed the high-throughput DNA sequencing data to create cleavable sequence motifs (Fig. 3a-g), and then incorporated mass spectrometry data to yield the final protease specificity motifs, including the

cleavage position (Fig. 3h-j). Filtering, trimming, and translating the NGS raw data yielded 13-16 million octamer peptide sequences that conformed to the library design (Fig. 3b, Supplementary Table 2). To estimate the false positive rate of the detection method, a control dataset was generated *in silico* by shuffling the amino acid order of the sequences in the selected pool (Fig. 3c). Next, the k-mers frequencies of the datasets were enumerated (k-mers are subsequences of the length k). A k-mer length of six was selected after observing that when longer lengths were used, specific k-mers showed up too infrequently to reliably evaluate k-mer presence and absence (Fig. 3d). The processed 6-mer sequences of the cleaved peptides were compared to the library of peptides before cleavage (control) to calculate an enrichment score for each 6-mer sequence. The 6-mer sequences below the frequency and enrichment cutoffs were filtered out (see methods section for details). A network of the surviving 6-mer sequences was generated. Sequences were pruned from the network if their number and strength of connections to others were below the specified cutoffs. This pruning was performed to consistently obtain large complex networks with the true datasets, while yielding sparse or no networks with randomized datasets (Fig. 3e). Consensus motifs were then generated by agglomerative hierarchical clustering of remaining 6-mers (Fig. 3f). The motifs were filtered by the fraction of total unique sequences they contained and overall similarity of last merge. The resulting parent motifs in each branch represented our cleavable sequence motifs (Fig. 3g, Supplementary Fig. 2). These cleavable sequence motifs are shown for each of the three proteases in Figs. 4a, 5a, and 6a. The positions of those motifs within the hierarchical tree are listed in Supplementary Table 3.

Determining cleavage positions for select peptides by mass spectrometry for motif refinement

To refine the motifs generated from the first stage of automated motif mining (Fig. 3g), we determined the cleavage position for a subset of peptides by a cleavage assay. For this purpose, we chose the five most highly enriched 6-mer sequences from each of the identified cleavable sequence motifs (Fig. 3g). In addition, we also chose, out of all 6-mer sequences, the five most highly enriched and the two least enriched sequences as controls (Fig. 3e). For each of these 6-mers, we extracted the associated octamer sequences from the un-processed dataset. A four amino acid long tag was added to the N-terminus of the octamers to facilitate subsequent mass spectrometric analysis by increasing charge and hydrophilicity. The peptides were chemically synthesized and combined into groups of 20 or fewer peptides each. We digested these mixtures with the respective protease. The resulting cleaved fragments were analyzed by LC-MS/MS to identify the position of cleavage for this subset of peptides (Fig. 3h; Figs. 4-6, panel b). We also confirmed the cleavage position of a known substrate for each protease by adding that substrate to the digest as positive control. The mass spectrometry results revealed that most of the peptides tested had multiple cleavage sites, which is reasonable since a peptide with multiple cleavage sites is an even better protease substrate than a peptide with only a single site.

The cleavage sites information was fed back into the software pipeline (Fig. 3h) to refine the cleavable sequence motifs and generate the final protease specificity motifs. For that purpose, the step (f) was repeated with 6-mer sequences that now included the position of cleavage (Fig. 3i). The motifs in the hierarchical cluster were filtered by the fraction of total nodes and overall similarity of last merge, and only the parent motif in a branch was kept. The filtering by similarity used in Fig. 3g was adjusted to ensure that at least one tested peptide with a known cleavage position was supporting the protease specificity motifs obtained (Fig. 3j). The three proteases factor Xa, ADAM17 and streptopain were subjected to this method. We successfully assigned the cleavage positions and obtained high-resolution cleavage specificity profiles for each protease as shown in panel c of Figs. 4-6.

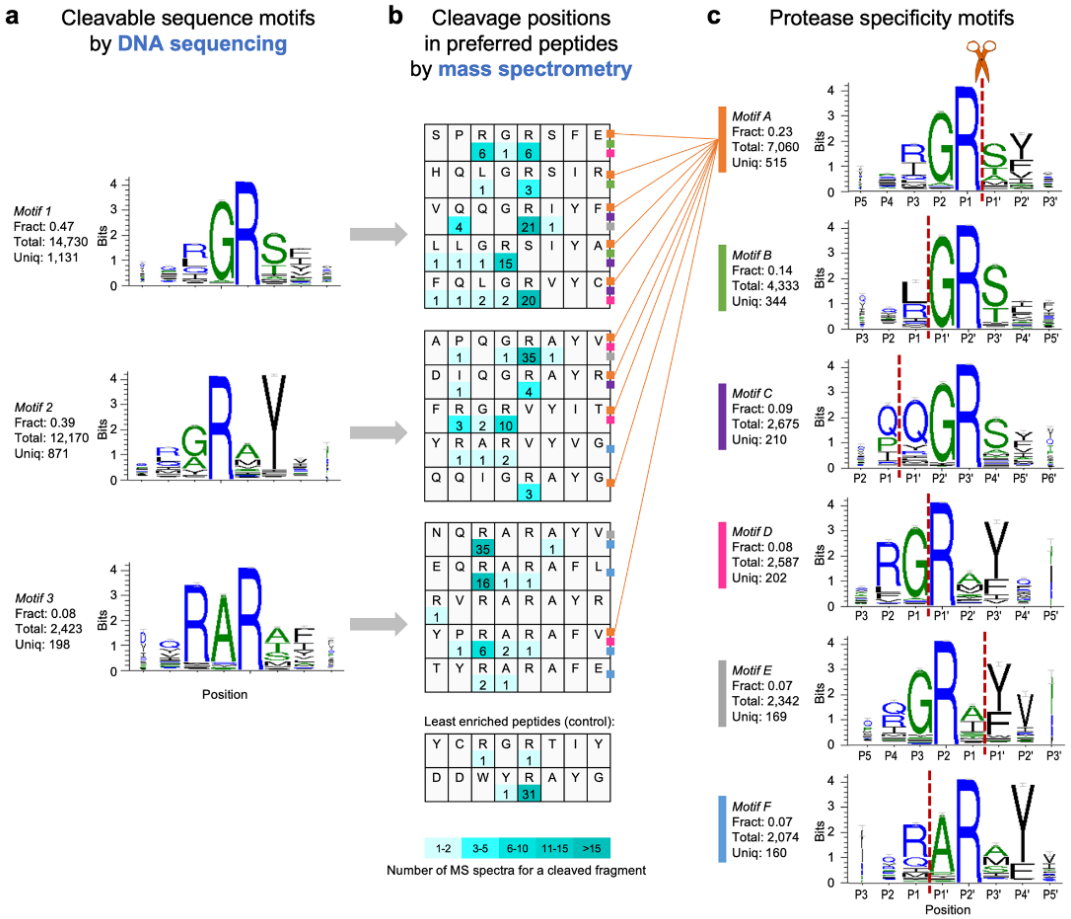

Figure 4. Protease specificity motif mining from high-throughput sequencing and mass spectrometry data for factor Xa. (a) DNA sequencing data of cleavable peptides yielded the three most enriched cleavable sequence motifs. Total: Number of sequences supporting this motif. Fract: Sequences supporting this motif as fraction of all sequences enriched for this protease. Uniq: Number of unique 6-mer sequences supporting this motif. (b) Cleavage positions in preferred peptides identified by mass spectrometry (MS). Numbers represent the number of MS spectra obtained (in shades of teal). Grey arrows indicate from which motif the group of preferred peptides originated. (c) Protease specificity motifs with cleavage position generated by combining cleavable sequence motifs seen in (a) with MS results seen in (b). Red dashed line indicates cleavage position. Only the six highest-ranking motifs are shown. Color match between squares after peptide sequence and vertical bar to the left of a motif indicates that the motif is supported by cleavage information from this sequence (connected by orange lines for motif A as example). Motifs are ordered from top to bottom by number of unique sequences supporting a motif. Sequence logos are generated from the aligned sequences using WebLogo 3.5.0: Overall height of each stack indicates sequence conservation at that position with height of an amino acid within the stack reflecting relative frequency. Error bars are the 95% confidence interval for the reported conservation. Width of amino acid letters corresponds to the relative frequency of 'valid' symbols at that position (the degree of support for amino acid frequencies at that position). The color of the amino acids represents their hydrophobicity (blue - hydrophilic, green – neutral, black – hydrophobic).

Protease specificity motifs of factor Xa

The narrow specificity serine protease factor Xa was used as a proof of principal application to verify the accuracy of our method against this very well-studied protease. The comprehensive results for factor Xa are shown in Fig. 4. The top three enriched cleavable motifs (Fig. 4a) were supported by 2,200 unique sequences and show the well-known GR or AR pattern[5d, 13b]. Mass spectrometry of the digested preferred peptide substrates clearly confirmed the known preferential cleavage after the R residue (Fig. 4b). Even the two peptides with the lowest copy number among the enriched peptides (least enriched peptides) were digested, which confirmed that the enrichment of cleavable sequences was successful. Fig. 4c shows the factor Xa protease specificity motifs including the cleavage site. The top six specificity

motifs were supported by 1,600 unique sequences (Fig. 4c). The top motif A alone was supported by 515 sequences. In contrast, less than 86 sequences were used previously to generate the specificity motifs for factor Xa by the PICS assay[13b]. Further, while the canonical specificity of R in P1 position and G or A in P2 were identified by PICS, it was unable to clearly conclude the specificity in positions P3 and P2′, based on the results from the two different peptide substrate libraries. Interestingly, most highly enriched and reactive sequences PQGR, IQGR, QQGR from a previously published tetramer library using phage display[5d] were also confirmed in our list of highly enriched sequences (Fig. 4b).

All of our final protease specificity motifs (Fig. 4c) contained the well-characterized GR or AR pattern[5d, 13b]. The main difference between the motifs were the cleavage positions, which were all supported by mass spectrometry. We will focus the following discussion on motif A, because the highest number of unique sequences supported this motif. Motif A confirmed the canonical specificity for R at position P1, and for the small amino acids G and A at position P2. The specificity for P3 were R, I, Q and L, among which the R in P3 agreed with four previous studies[5d, 13c, 13e, 20], and the L and Q in P3 agreed with two of them[5d, 20]. However, similar to the result from PICS, the cleavage at K at P1 described by a previous study[6] (collected by MEROPS motif) was not confirmed.

A unique and previously unreported Q-R (P2-P1) pattern was identified in both motif B and motif F with different prime side preference. The Q in P2 position, which we identified, was not found in the MEROPS database[6] nor in the motifs from the PICS method. The cleavage after QR was confirmed by mass spectrometry (Fig. 4b), and the QR pattern had also been found previously in the LLQR sequence that was highly enriched by phage display[5d].

Few previous studies focused on the specificity of prime side for factor Xa. The motif A shows the P1′ preference for S, T and A. The S and T preference had been described in the MEROPS database (based on 63 cleavages)[6], while S and A had been found by PICS. Furthermore, motif A clearly demonstrates a specificity at P2′ for Y, F, V and I. This finding confirms the F and V preferences reported in the MEROPS database[6] with several occurrences, as well as the preferences of Y and I reported in MEROPS[6] with only a single occurrence each. In summary, the motifs obtained by our method were supported by an order of magnitude more sequences compared to the motifs generated by the PICS method, or those described in the MEROPS database[6]. Just our top motif A already described most of the previously published specificity information, while our five additional motifs reported on further cleavage occurrences. The motifs D and E showed the cleavage after G or RA, which was confirmed by the minor fragment spectra in the LC-MS/MS results (7 and 6 spectra of cleavages after G or RA, respectively). Furthermore, our method also enables the search for possible minor specificity motifs by allowing the user to choose the minimum number of spectra required for the cleavage site alignment.

Protease specificity motifs of ADAM17

The broad specificity ADAM17 protease was investigated here to apply our method towards a protease that has been well-studied by the most powerful previous methods for understanding broad-specificity proteases. Our comprehensive results for ADAM17 are shown in Fig. 5. The top six enriched cleavable motifs (Fig. 5a) were supported by 609 unique sequences with the QAV or QRV patterns dominating. The mass spectrometry results of the digested preferred peptide substrates showed a trend of cleavage after the residue following the Q, or RP combination (Fig. 5b). The digestion was confirmed for even the least enriched peptides, which verified again the successful enrichment of cleavable sequences. Fig. 5c showed the ADAM17 protease specificity motifs including the cleavage site. The top four specificity motifs were supported by 363 unique sequences (Fig. 5c).

The motifs A, B and C showed some similarity with subsets of amino acid preferences at different positions. The total number of unique sequences comprised in the three motifs alone were 299. For comparison, the PICS assay had to rely on only 256 sequences to generate all ADAM17 specificity motifs[17].

The motifs A, B and C all showed a specificity for Q at P2 position, which agreed with the MEROPS database[6], but the PICS method did not identify this preference. The amino acids A and R were preferred in the P1 position for the

three motifs, which are also the top two amino acids at P1 position in MEROPS database[6]. The P1 specificities of the three motifs also agreed with the PICS results. The V preference in the P1′ position in the three motifs is the same as found by PICS and in the MEROPS database[6]; the L and I preference in P1′ in motif C also agreed with PICS results and one other previous study[16]. The canonical substrate of the ADAM17 protease, TNFα (PLAQA↓VRSSS, with ↓ indicating the cleavage position)[21], was also corroborated by our motif A and motif C for the four positions P2-P2′. However, the PICS results only covered the three positions (P1-P2′). ADAM17 has been reported to cleave the human ACE2 (SARS-CoV-2 receptor) after R708 (RMSR↓SRIN), which we confirmed in our motif B in P1 position[22]. The strong preference for VQR in positions P3-P1 of our motif B agrees with another native substrate of ADAM17, Cadherin-3 (EVQR↓LTVT), that was suggested by a recent secretomes analysis[23] but was not found by the PICS method. The VQA motif in our motif C agrees well with a cleavable site in the EPHA2 receptor (QVQA↓LTQE) found in that same analysis[23]. The human IgG Fc receptor III (CD16) is another well-characterized substrate of ADAM17, which has previously been shown to be cleaved three times *in vivo* (ITQGLA↓V↓ST↓ISSF)[24]. We also observed multiple cleavages for several of our tested peptides (Fig. 4b). Two of the three *in vivo* cleavage positions in CD16 were represented already by our top four cleavage motifs. When CD16 is cleaved between A and V, the P1 and P1′ positions agree with our motifs A and C. When the cleavage happens between V and S, the motif D agrees with the P2 and P1 positions, however the SPD-NGS method did not show the preference of V at P1 position[5c]. Our motif A also had Q preferred in the P1 position, which is found in another native substrate, the Interleukin 6 receptor (IL-6R) (SLPVQ↓DSSSV)[25]. In contrast, the PICS results did not agree with this natural substrate sequence. The Ebola virus surface glycoprotein GP is also a native ADAM17 substrate (KTLPD↓QGDND)[26], and the Q in P1′ position agreed with our motif B, but again the PICS method did not identify a match at any position. The SPD-NGS method did not show a high confidence score for cleavage prediction for Q at either P1 or P1′ positions[5c].

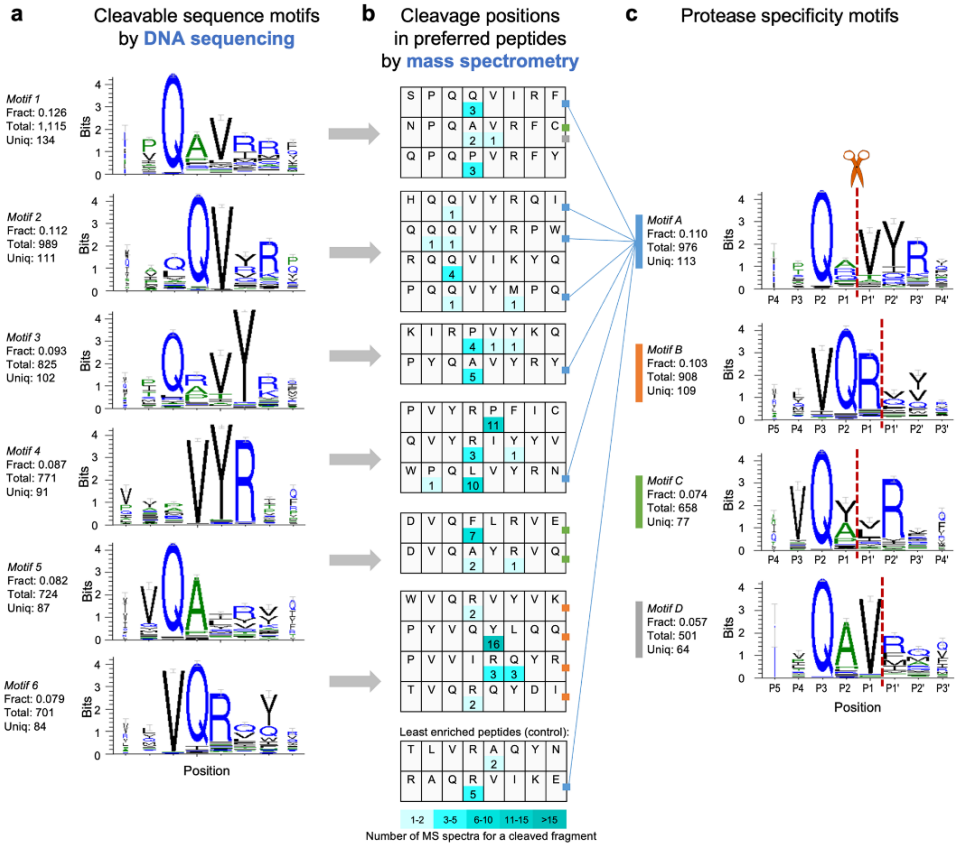

Figure 5. Protease specificity motif mining for ADAM17. (a) DNA sequencing data of cleavable peptides yielded the six most enriched cleavable sequence motifs. Total: Number of sequences supporting this motif. Fract: Sequences supporting this motif as fraction of all sequences enriched for this protease. Uniq: Number of unique 6-mer sequences supporting this motif. (b) Cleavage positions in preferred peptides identified by mass spectrometry (MS). Numbers represent the number of MS spectra obtained (in shades of teal). Grey arrows indicate from which motif the group of preferred peptides originated. (c) Protease specificity motifs with cleavage position generated by combining cleavable sequence motifs seen in (a) with MS results seen in (b). Red dashed line indicates cleavage position. Only the four highest-ranking motifs are shown. Color match between squares after peptide sequence and vertical bar to the left of a motif indicates that the motif is supported by cleavage information from this sequence (connected by blue lines for motif A as example). Motifs are ordered from top to bottom by number of unique sequences supporting a motif. Sequence logos are generated from the aligned sequences using WebLogo 3.5.0: Overall height of each stack indicates sequence conservation at that position with height of an amino acid within the stack reflecting relative frequency. Error bars are the 95% confidence interval for the reported conservation. Width of amino acid letters corresponds to the relative frequency of 'valid' symbols at that position (the degree of support for amino acid frequencies at that position). The color of the amino acids represents their hydrophobicity (blue - hydrophilic, green – neutral, black – hydrophobic).

A crucial role of ADAM17 is the proteolytic processing of tumor necrosis factor alpha (TNFα), which is important in the regulation of immune cells[15a]. Interestingly, the exact P2-P2′ sequence of TNFα (QAVR) was found in one of our highly enriched 6-mer sequences: PQAVRF. To probe the human proteome for additional matches with our enriched 6-mer sequences, we searched the human proteome database using the peptide search tool UniProt[27] for a small subset of our >1,000 enriched sequences, namely the 20 enriched sequences with MS-confirmed cleavage position (Fig. 5b). The entire 6-mers, and their 5-mer subsequences that retained the cleavage site were used for the searches, returning 8 and 529 matches, respectively. These matches included 35 cell surface transmembrane proteins. ADAM17 is known to cleave transmembrane-anchored proteins proximal to the membrane[28]; therefore, we further analyzed the 6-mer and 5-mer matches in known transmembrane proteins for their potential protease accessibility (defined as being located within 30 amino acids proximal to the membrane). The sequence LVRA↓Q is located in the putative olfactory receptor 52L2 (UniProt ID: Q8NGH6), which is responsible for the detection of odorants in olfactory receptor neurons[29]. Another sequence PQQ↓VI is found within the facilitated glucose transporter member 2 (UniProt ID: P11168), which facilitates glucose movement across cell membranes[30]. These cleavage sites are located on the protein surface according to their predicted structures by AlphaFold[31] and should therefore be protease-accessible. These examples indicate that the highly enriched cleavable sequences identified by our method and their matches to the human proteome could help inform future research on the physiological function of ADAM17. Expanding this search for proteome matches beyond these 20 enriched peptide sequences will likely yield additional hits of potentially interesting ADAM17 substrate candidates for further studies. These likely substrate candidates will have to be confirmed by testing the respective full-length proteins for cleavage under physiologically relevant conditions. Such an analysis is beyond the scope of the current study, not least because of added challenges due to many of the candidates being membrane proteins.

Protease specificity motifs of streptopain

The broad specificity cysteine protease streptopain was assessed by our method to further understand a protease with important clinical applications whose specificity has been notoriously difficult to analyze. Fig. 6 shows the comprehensive results for streptopain. The top three enriched cleavable motifs (Fig. 6a) were supported by 461 unique sequences, and the hydrophobic amino acids were enriched in these motifs. The mass spectrometry results of digested preferred peptide substrates revealed that multiple cleavages (>3 for most sequences) occurred in the enriched sequences (Fig. 6b). The successful cleavage of the two least enriched peptides confirmed the effective enrichment of all cleavable sequences. Fig. 6c showed the streptopain protease specificity motifs including the cleavage site. The top four specificity motifs were supported by 601 unique sequences in total (Fig. 6c). A previous study showed that hydrophobic residues are preferred at P2, but little preference was reported for positions P3 or P1′[19]. Another analysis by PICS also showed preference for hydrophobic residues at P2, but inconsistent results were obtained for other positions (e.g., P1 and P1') from different proteome substrates[2i]. Our motifs preferred Y and I (motif A) or M, F, and V (motif B) at the P2 position,

agreeing with the P2 specificity in the previous findings and the MEROPS database[6]. Our motifs showed a clear preference for F at the P1′ position, which had not been reported before. The specificity motifs and the mass spectrometry results suggested that sequences containing consecutive hydrophobic amino acids (I, M, F and Y) are highly cleavable by streptopain, since each hydrophobic amino acid can be recognized as a P2 position resulting in multiple cleavage sites in a single sequence. A subsite cooperativity change was also observed among our motifs. In motif B, when Y was in the P1 position, M was preferred in P2 over other hydrophobic amino acids. In contrast in motif C, when M was in the P1 position, F was preferred in P2 over other hydrophobic amino acids. A recent study reported the crucial cleavage of the GSDMA protein by streptopain after Q246 (ILIQ↓ASDV) which triggers pyroptosis[32]. This cleavage after IQ was also supported by our data as demonstrated by high numbers of LC-MS/MS spectra for two peptides (Fig. 6b).

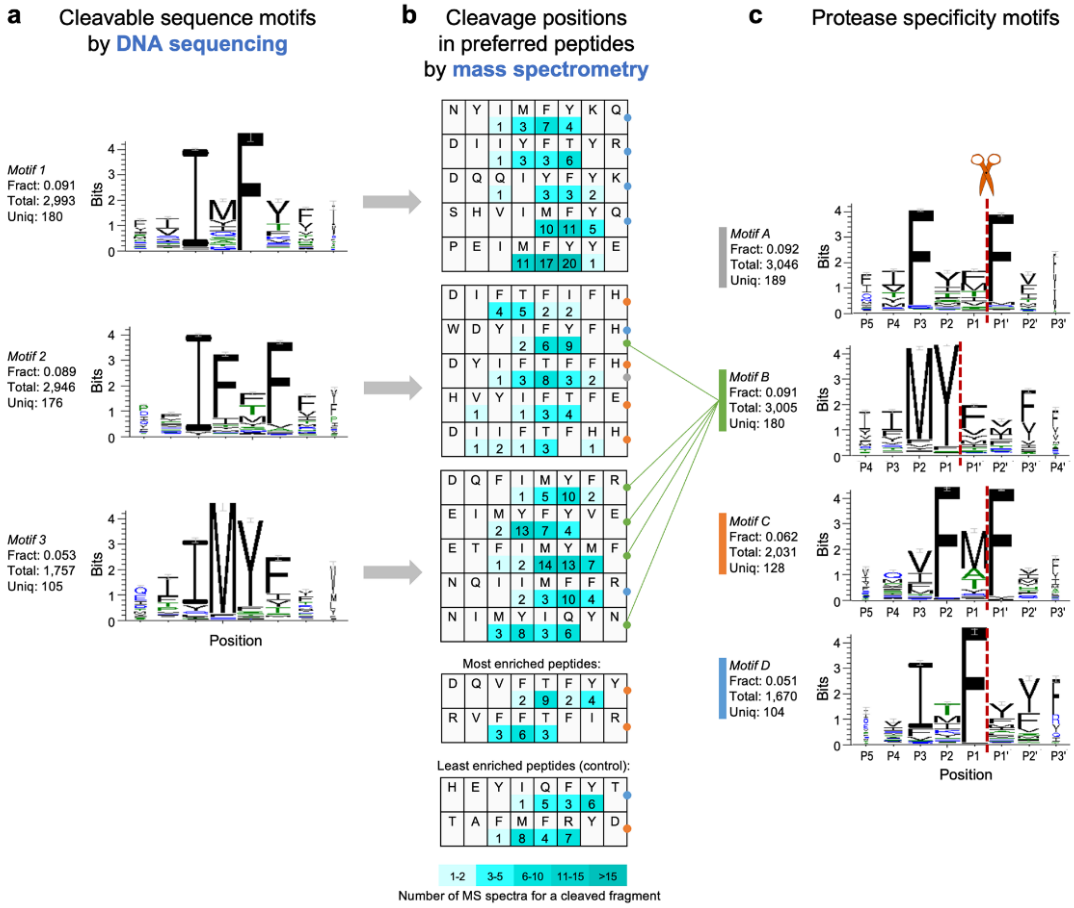

Figure 6. Protease specificity motif mining for streptopain. (a) DNA sequencing data of cleavable peptides yielded the three most enriched cleavable sequence motifs. Total: Number of sequences supporting this motif. Fract: Sequences supporting this motif as fraction of all sequences enriched for this protease. Uniq: Number of unique 6-mer sequences supporting this motif. (b) Cleavage positions in preferred peptides identified by mass spectrometry (MS). Numbers represent the number of MS spectra obtained (in shades of teal). Grey arrows indicate from which motif the group of preferred peptides originated. (c) Protease specificity motifs with cleavage position generated by combining cleavable sequence motifs seen in (a) with MS results seen in (b). Red dashed line indicates cleavage position. Only the four highest-ranking motifs are shown. Color match between squares after peptide sequence and vertical bar to the left of a motif indicates that the motif is supported by cleavage information from this sequence (connected by a green line for motif B as example). Motifs are ordered from top to bottom by number of unique sequences supporting a motif. Sequence logos are generated from the aligned sequences using WebLogo 3.5.0: Overall height of each stack indicates sequence conservation at that position with height of an amino acid within the stack reflecting relative frequency. Error bars are the 95% confidence interval for the reported conservation. Width of amino acid letters corresponds to the relative frequency of 'valid' symbols at that position (the degree of support for amino acid frequencies at that position). The color of the amino acids represents their hydrophobicity (blue - hydrophilic, green – neutral, black – hydrophobic).

Since the producer of streptopain, *S. pyogenes*, causes a wide range of infections in human, we searched for matches of the 19 enriched sequences with MS-confirmed cleavage sites (Fig. 6b) in the human proteome database. Streptopain

is known to target the proteins of the immune system and signaling proteins or systems of the host[18]. The entire 6-mers and the 5-mer subsequences with confirmed cleavage sites, returned 3 and 230 matches, respectively. These matches included 37 cell surface proteins and 5 secreted proteins, such as complement C1r subcomponent (loop region), ion channels, receptors and glutathione peroxidase.

In summary, these results vastly expand the previously published specificity information for this important protease. As with both streptopain and ADAM17, performing proteome-wide searches for sequences matching with more of our thousands of enriched peptide sequences, or against proteomes from other organisms, would likely generate additional hits for further studies.

Conclusion

The existing state-of-the-art methods have become adept at identifying canonical preferences of narrow-specificity proteases screens[2a-h, 3, 33], yet, the cleavage preferences for broad-specificity proteases have been challenging to fully elucidate. The new specificity analysis approach described here has the following advantages over existing technologies: (1) We eliminated amino acid bias that usually results from using a proteomic substrate library isolated from an organism. Proteomic libraries do not present all possible permutations of amino acids[2a-e, 2g, 13b, 17]. Instead, we used a chemically synthesized library containing all possible octamer substrates rendering our method organism-independent and our results applicable to all organisms. The library was built from nucleic acid trimers instead of the more common degenerate NNS or NNK codons resulting in a more even amino acid distribution, free from detrimental stop codons. Analyzing the precise distribution of amino acids present in the library immediately before the protease cleavage allowed us to accurately compare enrichment of cleaved substrates. (2) Our library did not require protease processing prior to selection[2a, 2c, 2g, 13b, 17]. PICS and some PICS-related methods require pre-processing with trypsin, for example, to prepare short peptide fragments leading to biases. Our library evenly presents essentially all possible octamers. (3) We included all natural amino acids without any modifications. Several previous methods omit amino acids like cysteine or methionine[3], or modify amino acids, such as protecting free lysine and cysteine residues[2a, 2c, 2g, 13b, 17]. With mRNA display, we were able to assay all possible octamers as unmodified peptides. (4) By combining LC-MS/MS and high-throughput sequencing, we assessed both the prime and non-prime cleavage specificity simultaneously. We developed a computer script that incorporates the MS results into the alignment process and generates cleavage site-indicated motifs even without prior knowledge of the protease of interest. (5) We were able to test orders of magnitude more peptide substrates than most of the previous best methods[13b, 17]. (6) mRNA display is a purely *in vitro* technology, which, in principle, allows for modifications of cleavage conditions to easily assess changes in protease function with regards to temperature, pH, ionic strength, or inhibitors[8b].

We applied this method to three proteases. The resulting detailed cleavage motifs were supported by testing more individual substrates than other current best methods. The specificity motifs for factor Xa markedly improved upon the previous best analysis by providing increased resolution at the P3 and P2′ sites[13b]. The specificity motifs generated for ADAM17 better resolved both sides of the cleavage position (Q at P1 and P1′ positions), compared to previous analyses[17]. The new and detailed specificity motifs for streptopain, the previously least characterized of the three proteases tested here, will lead to a better understanding of this important protease than ever before.[19] Besides the utility of the final specificity motifs, this method identified several thousand individual cleavable peptides for each protease that can be also used directly to search proteomes for sequence matches. Potentially interesting matches could be the starting point for future research that would need to include a confirmation that cleavage is also happening in the context of the whole native protein.

While our method overcomes a wide range of limitations, biases and shortcomings of previous approaches, there are some aspects that could be further improved. For example, potential impurities in the chemically synthesized peptides could affect the accuracy of the MS/MS cleavage site analysis. However, this potential issue can be minimized by either using high-purity peptides, or by increasing the minimum required number of spectra used for the MS/MS data-guided

realignment. Furthermore, some octamer sequences in our comprehensive library may translate less efficiently than others and could therefore be underrepresented. We reduced this potential bias by avoiding rare codons in the library through trimer nucleotide synthesis[12] and by 30-fold oversampling of all possible octamers. However, some translation bias might persist for a few unusual octamer sequences. Similar to other cleavage specificity analysis methods, the detection of cleavage of a small peptide does indicate - but not guarantee - cleavage of the same sequence within the context of a larger, folded protein. Therefore, protease specificity analyses, including the method presented here, only yield likely predictions for protein cleavage that will ultimately have to be tested in native conditions.

In conclusion, the method described here yields specificity determinations with an unprecedented level of detail and can be applied to identify the specificity of potentially any protease. This advantage in throughput is particularly crucial to understand broad-specificity proteases. Since proteases are ubiquitous enzymes with critical roles in hormone activation, digestion, and immune response, and are the target of about 10% of all pharmaceuticals[34], this method has the potential to help address numerous medical problems, and provide a comprehensive analysis tool for protease engineering and potential inhibitor design studies.

Supporting Information

Experimental details, supplementary figures, and supplementary tables are available in supporting information. The authors have cited additional references within the Supporting Information.[35-47]

Data Availability

Data are available on reasonable request.

Author Contributions

M.D.L. performed mRNA-displayed peptide library design, preparation, and digestion; analysis of NGS data. B.Z. performed mRNA-displayed peptide library digestion; NGS samples preparation and analysis; subset peptide library design, preparation, digestion, and MS sample preparation; analysis of bioinformatics and MS data. J.A.B. developed and implemented bioinformatics approaches. T.J.G. helped conceive experiments for mass spectrometry and design data analysis workflow; P.D.J. helped analyze mass spectrometry data and interpret results. M.D.L., B.S., and B.Z. designed the research, analyzed the data, and wrote the paper. All authors reviewed and edited the paper.

Competing Interests

The authors declare no competing interests.

Acknowledgements

We thank M. F. Freeman, L. Higgins and T. W. Markowski for advice on mass spectrometry; Kun-Hwa Lee for optimizing the hammerhead cleavage protocol; A. J. O'Donoghue, P. M. Schlievert and B. Walcheck for scientific discussions; and J. D. Lipscomb for comments on the manuscript. This work was funded by grants from the US National Institute of Health (NIH) (AI113406; MSTP grant T32 GM008244 to M.D.L.; U24CA199347 to T.J.G. and P.D.J.), the Minnesota Medical Foundation (4036-9663-10 and 4088-9221-12), the University of Minnesota Biocatalysis Initiative, the Office of VP of Research at the University of Minnesota (Grant-in-Aid), University of Minnesota Informatics

Institute (Updraft Grant), the American Heart Association (12PRE11900061 to M.D.L.), the Graduate School Doctoral Dissertation Fellowship (M.D.L.), and the Organization for Fundamental Research of Institute of Innovative Research, Tokyo Institute of Technology (B.Z).

Keywords: protease • mRNA display • click chemistry • NGS • LC-MS/MS

References:

[1] M. Kawaguchi, K. Inoue, I. Iuchi, M. Nishida, S. Yasumasu, *BMC Evol Biol* **2013**, *13*, 231.

[2] aO. Schilling, C. M. Overall, *Nat Biotechnol* **2008**, *26*, 685-694; bM. L. Biniossek, D. K. Nagler, C. Becker-Pauly, O. Schilling, *J Proteome Res* **2011**, *10*, 5363-5373; cO. Schilling, P. F. Huesgen, O. Barré, U. Auf dem Keller, C. M. Overall, *Nat Protoc* **2011**, *6*, 111-120; dW. Ju, C. A. Valencia, H. Pang, Y. Ke, W. Gao, B. Dong, R. Liu, *Proc Natl Acad Sci U S A* **2007**, *104*, 14294-14299; eC. A. Valencia, S. W. Cotten, B. A. Dong, R. H. Liu, *Biotechnol Prog* **2008**, *24*, 561-569; fU. Eckhard, P. F. Huesgen, O. Schilling, C. L. Bellac, G. S. Butler, J. H. Cox, A. Dufour, V. Goebeler, R. Kappelhoff, U. A. Keller, T. Klein, P. F. Lange, G. Marino, C. J. Morrison, A. Prudova, D. Rodriguez, A. E. Starr, Y. Wang, C. M. Overall, *Matrix Biol* **2016**, *49*, 37-60; gM. L. Biniossek, M. Niemer, K. Maksimchuk, B. Mayer, J. Fuchs, P. F. Huesgen, D. G. McCafferty, B. Turk, G. Fritz, J. Mayer, G. Haecker, L. Mach, O. Schilling, *Mol Cell Proteomics* **2016**, *15*, 2515-2524; hO. Kleifeld, A. Doucet, U. auf dem Keller, A. Prudova, O. Schilling, R. K. Kainthan, A. E. Starr, L. J. Foster, J. N. Kizhakkedathu, C. M. Overall, *Nat Biotechnol* **2010**, *28*, 281-288; iC. Blöchl, C. Holzner, M. Luciano, R. Bauer, J. Horejs-Hoeck, U. Eckhard, H. Brandstetter, C. G. Huber, *International Journal of Molecular Sciences* **2022**, *23*, 412; jF. Uliana, M. Vizovišek, L. Acquasaliente, R. Ciuffa, A. Fossati, F. Frommelt, S. Goetze, B. Wollscheid, M. Gstaiger, V. De Filippis, U. auf dem Keller, R. Aebersold, *Nat Commun* **2021**, *12*, 1693.

[3] A. J. O'Donoghue, A. A. Eroy-Reveles, G. M. Knudsen, J. Ingram, M. Zhou, J. B. Statnekov, A. L. Greninger, D. R. Hostetter, G. Qu, D. A. Maltby, M. O. Anderson, J. L. Derisi, J. H. McKerrow, A. L. Burlingame, C. S. Craik, *Nat Methods* **2012**, *9*, 1095-1100.

[4] T.-A. Nguyen, T. F. Gronauer, T. Nast-Kolb, S. A. Sieber, K. Lang, *Angewandte Chemie International Edition* **2022**, *61*, e202111085.

[5] aC. A. Kretz, K. Tomberg, A. Van Esbroeck, A. Yee, D. Ginsburg, *Scientific Reports* **2018**, *8*, 2788; bB. I. Ratnikov, P. Cieplak, A. G. Remacle, E. Nguyen, J. W. Smith, *PLoS Comput Biol* **2021**, *17*, e1008101; cJ. Zhou, S. Li, K. K. Leung, B. O'Donovan, J. Y. Zou, J. L. DeRisi, J. A. Wells, *Proc Natl Acad Sci U S A* **2020**, *117*, 25464-25475; dH. J. Hsu, K. C. Tsai, Y. K. Sun, H. J. Chang, Y. J. Huang, H. M. Yu, C. H. Lin, S. S. Mao, A. S. Yang, *J Biol Chem* **2008**, *283*, 12343-12353.

[6] N. D. Rawlings, A. J. Barrett, A. Bateman, *Nucleic Acids Res* **2012**, *40*, D343-D350.

[7] aN. Nemoto, E. Miyamoto-Sato, Y. Husimi, H. Yanagawa, *FEBS Lett.* **1997**, *414*, 405-408; bR. W. Roberts, J. W. Szostak, *Proc. Natl. Acad. Sci. USA* **1997**, *94*, 12297-12302.

[8] aM. S. Newton, Y. Cabezas-Perusse, C. L. Tong, B. Seelig, *ACS Synth Biol* **2020**, *9*, 181-190; bM. V. Golynskiy, B. Seelig, *Trends Biotechnol* **2010**, *28*, 340-345.

[9] B. Seelig, *Nat Protoc* **2011**, *6*, 540-552.

[10] R. B. Quast, D. Mrusek, C. Hoffmeister, A. Sonnabend, S. Kubick, *FEBS Lett* **2015**, *589*, 1703-1712.

[11] S. Li, S. Millward, R. Roberts, *J Am Chem Soc* **2002**, *124*, 9972-9973.

[12] A. Yagodkin, A. Azhayev, J. Roivainen, M. Antopolsky, A. Kayushin, M. Korosteleva, A. Miroshnikov, J. Randolph, H. Mackie, *Nucleosides Nucleotides Nucleic Acids* **2007**, *26*, 473-497.

[13] aE. P. Bianchini, V. B. Louvain, P. E. Marque, M. A. Juliano, L. Juliano, B. F. Le Bonniec, *J Biol Chem* **2002**, *277*, 20527-20534; bO. Schilling, U. auf dem Keller, C. M. Overall, *Biol Chem* **2011**, *392*, 1031-1037; cD. N. Gosalia, C. M. Salisbury, D. J. Maly, J. A. Ellman, S. L. Diamond, *Proteomics* **2005**, *5*, 1292-1298; dJ. L. Harris, B. J.

Backes, F. Leonetti, S. Mahrus, J. A. Ellman, C. S. Craik, *Proc Natl Acad Sci U S A* **2000**, *97*, 7754-7759; eM. D. Scholle, U. Kriplani, A. Pabon, K. Sishtla, M. J. Glucksman, B. K. Kay, *ChemBioChem* **2006**, *7*, 834-838.

[14] aR. A. Black, C. T. Rauch, C. J. Kozlosky, J. J. Peschon, J. L. Slack, M. F. Wolfson, B. J. Castner, K. L. Stocking, P. Reddy, S. Srinivasan, N. Nelson, N. Boiani, K. A. Schooley, M. Gerhart, R. Davis, J. N. Fitzner, R. S. Johnson, R. J. Paxton, C. J. March, D. P. Cerretti, *Nature* **1997**, *385*, 729-733; bJ. J. Peschon, J. L. Slack, P. Reddy, K. L. Stocking, S. W. Sunnarborg, D. C. Lee, W. E. Russell, B. J. Castner, R. S. Johnson, J. N. Fitzner, R. W. Boyce, N. Nelson, C. J. Kozlosky, M. F. Wolfson, C. T. Rauch, D. P. Cerretti, R. J. Paxton, C. J. March, R. A. Black, *Science* **1998**, *282*, 1281-1284.

[15] aM. L. Moss, D. Minond, *Mediat Inflamm* **2017**, 21; bH. K. Mishra, J. Ma, B. Walcheck, *Front Cell Infect Microbiol* **2017**, *7*, 138.

[16] C. I. Caescu, G. R. Jeschke, B. E. Turk, *Biochem J* **2009**, *424*, 79-88.

[17] J. Tucher, D. Linke, T. Koudelka, L. Cassidy, C. Tredup, R. Wichert, C. Pietrzik, C. Becker-Pauly, A. Tholey, *J Proteome Res* **2014**, *13*, 2205-2214.

[18] D. C. Nelson, J. Garbe, M. Collin, *Biol Chem* **2011**, *392*, 1077-1088.

[19] M. Nomizu, G. Pietrzynski, T. Kato, P. Lachance, R. Menard, E. Ziomek, *J Biol Chem* **2001**, *276*, 44551-44556.

[20] S. T. Furlong, R. C. Mauger, A. M. Strimpler, Y. P. Liu, F. X. Morris, P. D. Edwards, *Bioorg Med Chem* **2002**, *10*, 3637-3647.

[21] M. J. Mohan, T. Seaton, J. Mitchell, A. Howe, K. Blackburn, W. Burkhart, M. Moyer, I. Patel, G. M. Waitt, J. D. Becherer, M. L. Moss, M. E. Milla, *Biochemistry* **2002**, *41*, 9462-9469.

[22] Z. W. Lai, I. Hanchapola, D. L. Steer, A. I. Smith, *Biochemistry* **2011**, *50*, 5182-5194.

[23] K. Tsumagari, C.-H. Chang, Y. Ishihama, *iScience* **2021**, *24*, 102259.

[24] Y. Jing, Z. Ni, J. Wu, L. Higgins, T. W. Markowski, D. S. Kaufman, B. Walcheck, *PloS one* **2015**, *10*, e0121788.

[25] J. Mullberg, W. Oberthur, F. Lottspeich, E. Mehl, E. Dittrich, L. Graeve, P. C. Heinrich, S. Rose-John, *J. Immunol.* **1994**, *152*, 4958-4968.

[26] O. Dolnik, V. Volchkova, W. Garten, C. Carbonnelle, S. Becker, J. Kahnt, U. Ströher, H. D. Klenk, V. Volchkov, *EMBO J* **2004**, *23*, 2175-2184.

[27] T. U. Consortium, *Nucleic Acids Res* **2018**, *47*, D506-D515.

[28] I. Lorenzen, J. Lokau, Y. Korpys, M. Oldefest, C. M. Flynn, U. Künzel, C. Garbers, M. Freeman, J. Grötzinger, S. Düsterhöft, *Sci Rep* **2016**, *6*, 35067-35067.

[29] S. Zozulya, F. Echeverri, T. Nguyen, *Genome Biol* **2001**, *2*, research0018.0011.

[30] H. Fukumoto, S. Seino, H. Imura, Y. Seino, R. L. Eddy, Y. Fukushima, M. G. Byers, T. B. Shows, G. I. Bell, *Proc Natl Acad Sci U S A* **1988**, *85*, 5434-5438.

[31] aJ. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, *Nature* **2021**, *596*, 583-589; bM. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, *Nucleic Acids Research* **2021**, *50*, D439-D444.

[32] D. L. LaRock, A. F. Johnson, S. Wilde, J. S. Sands, M. P. Monteiro, C. N. LaRock, *Nature* **2022**, *605*, 527-531.

[33] aM. Vizovisek, R. Vidmar, M. Drag, M. Fonovic, G. S. Salvesen, B. Turk, *Trends Biochem Sci* **2018**, *43*, 829-844; bM. A. Pethe, A. B. Rubenstein, S. D. Khare, *Proc Natl Acad Sci U S A* **2019**, *116*, 168-176; cP. Van Damme, L. Martens, J. Van Damme, K. Hugelier, A. Staes, J. Vandekerckhove, K. Gevaert, *Nat Methods* **2005**, *2*, 771-777; dO. Schilling, O. Barre, P. F. Huesgen, C. M. Overall, *Nat Methods* **2010**, *7*, 508-511; eQ. Li, L. Yi, K. H. Hoi, P.

Marek, G. Georgiou, B. L. Iverson, *ACS Chem Biol* **2017**, *12*, 510-518; fK. T. Boulware, P. S. Daugherty, *Proc Natl Acad Sci U S A* **2006**, *103*, 7583-7588; gP. Schlage, F. E. Egli, U. Auf dem Keller, *Methods Mol Biol* **2017**, *1579*, 185-198.

[34] M. D. Lim, C. S. Craik, *Bioorg Med Chem* **2009**, *17*, 1094-1100.

[35] S. H. Hong, I. Ntai, A. D. Haimovich, N. L. Kelleher, F. J. Isaacs, M. C. Jewett, *ACS Synth Biol* **2014**, *3*, 398-409.

[36] C. Albayrak, J. R. Swartz, *Nucleic Acids Res* **2013**, *41*, 5949-5963.

[37] S. Horiya, J. K. Bailey, J. S. Temme, Y. V. Guillen Schlippe, I. J. Krauss, *J Am Chem Soc* **2014**, *136*, 5407-5415.

[38] M. D. Lane, B. Seelig, *Protein Expr Purif* **2016**, *121*, 66-72.

[39] D. Defays, *Comput J* **1977**, *20*, 364-366.

[40] G. E. Crooks, G. Hon, J. M. Chandonia, S. E. Brenner, *Genome Res.* **2004**, *14*, 1188-1190.

[41] J. Kyte, R. F. Doolittle, *Journal of molecular biology* **1982**, *157*, 105-132.

[42] J. Rappsilber, Y. Ishihama, M. Mann, *Anal Chem* **2003**, *75*, 663-670.

[43] Y. Lin-Moshier, P. J. Sebastian, L. Higgins, N. D. Sampson, J. E. Hewitt, J. S. Marchant, *J Biol Chem* **2013**, *288*, 355-367.

[44] M. C. Chambers, B. Maclean, R. Burke, D. Amodei, D. L. Ruderman, S. Neumann, L. Gatto, B. Fischer, B. Pratt, J. Egertson, K. Hoff, D. Kessner, N. Tasman, N. Shulman, B. Frewen, T. A. Baker, M.-Y. Brusniak, C. Paulse, D. Creasy, L. Flashner, K. Kani, C. Moulding, S. L. Seymour, L. M. Nuwaysir, B. Lefebvre, F. Kuhlmann, J. Roark, P. Rainer, S. Detlev, T. Hemenway, A. Huhmer, J. Langridge, B. Connolly, T. Chadick, K. Holly, J. Eckels, E. W. Deutsch, R. L. Moritz, J. E. Katz, D. B. Agus, M. MacCoss, D. L. Tabb, P. Mallick, *Nat Biotechnol* **2012**, *30*, 918-920.

[45] aH. Barsnes, M. Vaudel, *J Proteome Res* **2018**, *17*, 2552-2555; bM. Vaudel, H. Barsnes, F. S. Berven, A. Sickmann, L. Martens, *Proteomics* **2011**, *11*, 996-999.

[46] M. Wilhelm, J. Schlegl, H. Hahne, A. M. Gholami, M. Lieberenz, M. M. Savitski, E. Ziegler, L. Butzmann, S. Gessulat, H. Marx, T. Mathieson, S. Lemeer, K. Schnatbaum, U. Reimer, H. Wenschuh, M. Mollenhauer, J. Slotta-Huspenina, J. H. Boese, M. Bantscheff, A. Gerstmair, F. Faerber, B. Kuster, *Nature* **2014**, *509*, 582-587.

[47] M. Vaudel, J. M. Burkhart, R. P. Zahedi, E. Oveland, F. S. Berven, A. Sickmann, L. Martens, H. Barsnes, *Nat Biotechnol* **2015**, *33*, 22-24.