High Throughput Substrate Specificity Profiling of Serine and Cysteine Proteases Using Solution-phase Fluorogenic Peptide Microarrays*

Dhaivat N. Gosalia†‡§, Cleo M. Salisbury¶§, Jonathan A. Ellman¶¶, and Scott L. Diamond†‡**

Proteases regulate numerous biological processes with a degree of specificity often dictated by the amino acid sequence of the substrate cleavage site. To map protease/substrate interactions, a 722-member library of fluorogenic protease substrates of the general format Ac-Ala-X-X-(Arg/Lys)-coumarin was synthesized (X = all natural amino acids except cysteine) and microarrayed with fluorescent calibration standards in glycerol nanodroplets on glass slides. Specifically of 13 serine proteases (activated protein C, plasma kallikrein, factor VIIa, factor IXa, factor XIa and factor α XIia, activated complement C1s, C1r, and D, trypptase, trypsin, subtilisin Carlsberg, and cathepsin G) and 11 papain-like cysteine proteases (cathepsin B, H, K, L, S, and V, rhodesain, papain, chymopapain, ficin, and stem bromelain) were obtained from 103,968 separate microarray fluorogenic reactions (722 substrates × 24 different proteases × 6 replicates). This is the first comprehensive study to report the substrate specificity of rhodesain, a papain-like cysteine protease expressed by Trypanosoma brucei rhodesiense, a parasitic protozoan responsible for causing sleeping sickness. Rhodesain displayed a strong P2 preference for Leu, Val, Phe, and Tyr in both the P1 and Arg libraries. Solution-phase microarrays facilitate protease/substrate specificity profiling in a rapid manner with minimal peptide library or enzyme usage. *Molecular & Cellular Proteomics 4:626–636, 2005.

Because of their critical roles in biological pathways like hormone activation, proteasome degradation, and apoptosis, proteases are essential for cellular function and viability. Proteases regulate hormonal activation, cellular homeostasis, apoptosis, and coagulation and play an important role in the pathogenicity and progression of many diseases (1). Proteases comprise one of the largest protein families in organisms from Escherichia coli to humans (2–4). Improved understanding of proteases will provide insight into biological systems and will likely provide a number of important new therapeutic targets (1).

To properly function, proteases must preferentially cleave their target substrates in the presence of other proteins. While many factors impact protease substrate selection, one of the key aspects is the complementary nature of the enzyme-active site with the residues surrounding the cleaved bond in the substrate. As such, determination of the residues that comprise the preferred cleavage site of a protease provides critical information regarding substrate selection. Furthermore, determination of substrate specificity also provides a framework for the design of potent and selective inhibitors.

Here we exploit solution-phase substrate nanodroplet microarrays (5), in which fluorogenic substrates suspended in glycerol droplets are treated with aerosolized aqueous enzyme solutions, to provide protease substrate specificity profiles (6). These arrays allow high throughput characterization of the preferred residues on the P side (7) of the substrate in a highly parallel and miniaturized format. We report the use of these arrays here to map the substrate specificity of 24 serine and cysteine proteases in a rapid and efficient manner.

EXPERIMENTAL PROCEDURES

Materials—Purified human activated protein C (APC),1 human plasma kallikrein, human factor VIIa, human factor IXa β, human factor XIa, and human factor α XIia were purchased from Enzyme Research Laboratories (South Bend, IN). Human two-chain activated complement C1r, human two-chain activated complement C1s, human complement factor D, human trypptase (lung), bovine trypsin (pancreas, high-purity endotoxin-free), subtilisin Carlsberg (Bacillus licheniformis), human cathepsin G (neutrophil), human cathepsin B (liver), human cathepsin H (liver), human cathepsin K (recombinant, E. coli), human cathepsin L (liver), human cathepsin S (spleen), human

1 The abbreviations used are: APC, activated protein C; ACC, 7-amino-4-carbamoylmethylcoumarin; FRET, fluorescence resonance energy transfer; PS-SCL, positional scanning-synthetic combinatorial library.
Microarrays for Profiling Protease Substrate Specificity

cathepsin V (recombinant, NSO cells), papain (carica papaya), and stem bromelain (pineapple) were purchased from Calbiochem (La Jolla, CA). Chymopapain (papaya latex) and ficain (fig tree latex) were purchased from Sigma Aldrich (St. Louis, MO). The rhodesain, a gift from the L. S. Brinen laboratory at University of California, San Francisco, was expressed from Pichia pastoris as previously reported (8). Lipidated recombinant human tissue factor was purchased from American Diagnostica (Stamford, CT). All enzymes were stored according to manufacturers’ instructions.

Methods—The fluorogenic substrate library was synthesized and printed according to protocols previously described (5, 6). The P$_1$ = Arg and P$_1$ = Lys sublibraries, along with calibration standards (unacylated 7-amino-4-carbamoylmethylcoumarin (ACC), acetylated ACC, and blanks), were printed at either 50 or 100 μM in a 16 × 24 format equivalent to a 384-well plate on polylysine-coated glass slides (Erie Scientific, Portsmouth, NH) using a 1 × 1 pin (Telechem, Sunnyvale, CA) protocol on an OmniGrid Accent (Gene Machines, San Carlos, CA) microarrayer. Calibration standards were printed on each array, to enable quantification and normalization of fluorescence intensity between slides.

The proteases were reconstituted and diluted in buffers recommended by the manufacturers. The enzyme solutions were delivered to the microarrays at the following concentrations: 10 μM APC in 20 mM Tris-HCl, 100 mM NaCl (pH 7.4); 1 μM plasma kallikrein in 4 mM NaOAc-HCl, 150 mM NaCl (pH 5.3); 5 μM factor VIIa in 20 mM Tris-HCl, 100 mM NaCl, 1 mM CaCl$_2$ (pH 7.4), and 15 mM lipoprotein lipase recombinant human tissue factor; 10 μM factor Xa in 20 mM Tris-HCl, 100 mM NaCl (pH 7.4); 100 μM factor Xla in 4 mM NaOAc-HCl, 150 mM NaCl (pH 5.3); 10 μM factor α, Xila in 4 mM NaOAc-HCl, 150 mM NaCl (pH 5.3); 10 μM two-chain activated complement C1r in 47 mM NaH$_2$PO$_4$ buffer, 5 mM EDTA, 1 mM benzamidine (pH 7.4); 10 μM two-chain activated complement C1s in 47 mM NaH$_2$PO$_4$ buffer, 5 mM EDTA, 1 mM benzamidine (pH 7.4); 4 μM complement factor D in 15 mM NaOAc-HCl, 135 mM NaCl (pH 7.2); 1 μM trypsin in 50 mM NaOAc, 1 M NaCl, 50 μM heparin, 0.01% NaN$_3$ (pH 5.0); 1 μM trypsin in a 0.9% NaCl aqueous solution; 1 μM subtilisin Carlsberg in water; 10 μM cathepsin G in 50 mM NaOAc, 150 mM NaCl (pH 5.5); 10 μM cathepsin B in 20 mM NaOAc buffer, 1 mM EDTA (pH 5.0); 5 μM cathepsin H in 100 mM Na$_2$HPO$_4$, 2 mM EDTA, 2 mM DTT (pH 6.0); 500 nm cathepsin K in 50 mM NaOAc, 100 mM NaCl, 2.5 mM EDTA, 2.5 mM DTT (pH 5.5); 5 μM cathepsin L in 20 mM malonate buffer, 400 mM NaCl, 1 mM EDTA (pH 5.5); 5 μM cathepsin S in 100 mM Na$_2$HPO$_4$, 2 mM EDTA, 2 mM DTT, 0.4% Triton X-100 (pH 6.0); 500 nm cathepsin V in 50 mM NaOAc, 100 mM NaCl, 2.5 mM EDTA, 2.5 mM DTT (pH 5.5); 5 μM rhodesain in 100 mM NaOAc, 2.5 mM DTT (pH 5.5); 10 μM papain in PBS (pH 6.0); 10 μM chymopapain in 50 mM NaOAc, 2.5 mM cysteine, 0.5 mM EDTA (pH 6.2); 10 μM ficain in 1 mM KH$_2$PO$_4$ buffer (pH 7.0) and 10 μM stem bromelain in PBS (pH 6.0).

The 100 μM substrate arrays were used for the plasma kallikrein and factor Xla β assays; all other assays employed the 50 μM substrate arrays. The proteases were delivered to the substrate arrays as described previously and underwent ~30-fold dilution from the initial delivery concentration after mixing in the micropost (5, 6). The slides (n = 6 per enzyme) were incubated at 37 °C for 30 min to 6 h. This incubation time resulted in ~5–25% cleavage of the best substrate on each array, assuring that the assays were run within the linear range. The activated slides were scanned using a cooled CCD-based image scanner (Novaray; Alpha Innotech, San Leandro, CA), at Ex = 405/40 nm and Em = 475/40 nm with integration times of 2,500 msec and 15-μm pixel resolution. Images were acquired in a 16-bit format, and the analysis and presentation of the data was performed using ArrayVision (Imaging Research, Ontario, Canada) and Cluster and Treewiew (10). For the most preferred substrates, the signal/background ratio was typically >60 with coefficient of variance <5%.

RESULTS AND DISCUSSION

A 722-member, spatially separated ACC library of the format Ac-P$_4$-P$_3$-P$_2$-P$_1$-ACC-NH$_2$ was prepared with Ala at the P$_2$ site, all combinations of proteinogenic amino acids (except Cys) at the P$_2$ and P$_3$ sites, and a Lys or Arg residue at the P$_1$ site (Fig. 1) (6). The substrate specificities of 13 serine proteases and 11 cysteine proteases (Table I, Figs. 2 and 3) were profiled using these substrates in a microarray-based format. The proteases profiled here span three evolutionarily unrelated clans (PA, SB, and CI). Multiple statistically homologous proteases from the PA clan (the S1 family) and the C1 clan (the CA family) were chosen to show that sequence homology need not correlate with substrate specificity.

Serine Proteases—All the serine proteases profiled are from the S1 family of clan PA except subtilisin Carlsberg, which is from the S8 family of clan SB. The two clans differ in the order of their overall fold and catalytic residues, with clan PA having double β barrels with His–Asp–Ser as the catalytic triad and clan SB having parallel β sheets and Asp–His–Ser as the catalytic triad (3, 11).

The proteases profiled from the S1 family showed a preference for P$_1$ = Lys or Arg due to the presence of the highly conserved negatively charged Asp$^{389}$ (chymotrypsinogen numbering) present at the bottom of the S$_1$ pocket (12, 13). P$_1$ Arg residues are able to form a direct ionic bond with Asp$^{389}$ whereas the corresponding Lys interaction is water-bridged because of the shorter side-chain of Lys. The similarities in active site composition and geometry of the serine proteases profiled can be attributed to convergent evolution, whereas the differences in the substrate specificities can be attributed to changes in the identities of residues present in the substrate binding clefts (14).

Blood Coagulation Proteases—APC is responsible for regulating the blood coagulation pathway by proteolytic inactivation of coagulation cofactors Va and VIIIa, thereby inhibiting the generation of thrombin. Thrombotic risk is associated with the inherited Leiden mutation, which results in APC-resistant
Microarrays for Profiling Protease Substrate Specificity

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>P&lt;sub&gt;1&lt;/sub&gt;</th>
<th>P&lt;sub&gt;2&lt;/sub&gt;</th>
<th>P&lt;sub&gt;3&lt;/sub&gt;</th>
<th>Observed specificity</th>
<th>Natural action</th>
<th>EC number</th>
<th>Clan</th>
<th>Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human APC</td>
<td>K</td>
<td>K</td>
<td>K</td>
<td>K→Q&gt;T,V,P</td>
<td>Inhibits factors Va and VIIIa</td>
<td>EC 3.4.21.69</td>
<td>PA</td>
<td>S1</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>K</td>
<td>R</td>
<td>K&gt;T,P,O&gt;V</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q&gt;M,R,S,T</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human plasma kallikrein</td>
<td>K</td>
<td>Y,F,K&gt;N,T</td>
<td>K</td>
<td>K&gt;E,M</td>
<td>Indicates physiologic substrates of APC: factor Va</td>
<td>EC 3.4.21.34</td>
<td>PA</td>
<td>S1</td>
</tr>
<tr>
<td></td>
<td>K</td>
<td>F,L&gt;N&gt;T,Y</td>
<td>R</td>
<td>R&gt;H&gt;A,E,L,G</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factor VIIa + TF</td>
<td>K</td>
<td></td>
<td></td>
<td>V→T</td>
<td>Initiates extrinsic coagulation</td>
<td>EC 3.4.21.21</td>
<td>PA</td>
<td>S1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Q,R&gt;N,P</td>
<td>Cascade; activates factors IX and X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factor IXa</td>
<td>K</td>
<td></td>
<td></td>
<td>F,P,K,Y→L,N</td>
<td>Activates factor X; may also activate factor VII</td>
<td>EC 3.4.21.22</td>
<td>PA</td>
<td>S1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>L,M,Q,R</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factor Xla</td>
<td>R</td>
<td>N→T&gt;S</td>
<td>S</td>
<td>F&gt;K,M&gt;E,H,L</td>
<td>Activates factor IX in contact phase of activation</td>
<td>EC 3.4.21.27</td>
<td>PA</td>
<td>S1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N→S&gt;T&gt;A</td>
<td>A</td>
<td>F&gt;E,W</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factor Xla</td>
<td>R</td>
<td>F,T&gt;G,N,S</td>
<td>M→Q</td>
<td>Activates factor XI and plasma kallikrein precursor</td>
<td>EC 3.4.21.38</td>
<td>PA</td>
<td>S1</td>
<td></td>
</tr>
<tr>
<td>Complement C1r</td>
<td>K</td>
<td>A,S,V,I</td>
<td>R</td>
<td>K→Y</td>
<td>Activates complement C1s</td>
<td>EC 3.4.21.41</td>
<td>PA</td>
<td>S1</td>
</tr>
<tr>
<td>Complement C1s</td>
<td>K</td>
<td>S&gt;A,G</td>
<td>Broad</td>
<td>Complement activation through cleavage of C2 and C4</td>
<td>EC 3.4.21.42</td>
<td>PA</td>
<td>S1</td>
<td></td>
</tr>
<tr>
<td>Complement factor D</td>
<td>K</td>
<td>A,S,T,V</td>
<td>R,K</td>
<td>Hydrolizes factor B to cleave</td>
<td>EC 3.4.21.46</td>
<td>PA</td>
<td>S1</td>
<td></td>
</tr>
<tr>
<td>Complement factor D</td>
<td>R</td>
<td>K</td>
<td>R</td>
<td>Intestinal digestion of food proteins and activation of other digestive</td>
<td>EC 3.4.21.4</td>
<td>PA</td>
<td>S1</td>
<td></td>
</tr>
<tr>
<td>Trypsin</td>
<td>R</td>
<td>S→N&gt;H,T</td>
<td>K,M&gt;N→Q,R,S,T</td>
<td>Peptidase zymogens</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trypsin</td>
<td>R</td>
<td>N→S,T</td>
<td>R,K</td>
<td>Serine endopeptidase with no cysteine residues</td>
<td>EC 3.4.21.62</td>
<td>SB</td>
<td>S8</td>
<td></td>
</tr>
<tr>
<td>Subtilisin Carlsberg</td>
<td>K</td>
<td>A→I,S,V→T</td>
<td>Q,R,S,T→E,M</td>
<td>Proteolysis, elastin degradation, MHC immune response</td>
<td>EC 3.4.21.25</td>
<td>CA</td>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>Cathepsin G</td>
<td>K</td>
<td>T→V→A</td>
<td>I→E→A,V</td>
<td>Activates complement C1s</td>
<td>EC 3.4.21.20</td>
<td>PA</td>
<td>S1</td>
<td></td>
</tr>
<tr>
<td>Cathepsin H</td>
<td>K</td>
<td>V→A,I,K,T</td>
<td>K,P&gt;L</td>
<td>Intrasubcellular lysosomal exopeptidase</td>
<td>EC 3.4.22.1</td>
<td>CA</td>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>Cathepsin V</td>
<td>R</td>
<td>T,V→K,J</td>
<td>K,L,P,R</td>
<td>Lysozyme activity, elastin degradation, MHC immune response</td>
<td>EC 3.4.22.17</td>
<td>CA</td>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>Cathepsin V</td>
<td>R</td>
<td>L</td>
<td>Broad</td>
<td>Lysozyme activity, elastin degradation, MHC immune response</td>
<td>EC 3.4.22.17</td>
<td>CA</td>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>Rhodesain</td>
<td>R</td>
<td>L→V</td>
<td>Broad</td>
<td>Degradation of the invariant chain of the MHC class II complex</td>
<td>EC 3.4.22.43</td>
<td>CA</td>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>Papain</td>
<td>K</td>
<td>V</td>
<td>P</td>
<td>Proteolytic component of the latex of papaya</td>
<td>EC 3.4.22.2</td>
<td>CA</td>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>Papain</td>
<td>R</td>
<td>V</td>
<td>P</td>
<td>Proteolytic component of the latex of papaya</td>
<td>EC 3.4.22.2</td>
<td>CA</td>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>Chymopapain</td>
<td>K</td>
<td>V→S,T</td>
<td>K→R</td>
<td>Proteolytic component of the latex of papaya</td>
<td>EC 3.4.22.6</td>
<td>CA</td>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>Chymopapain</td>
<td>R</td>
<td>V→I&gt;T,A</td>
<td>A→V,R→P</td>
<td>Proteolytic component of the latex of papaya</td>
<td>EC 3.4.22.3</td>
<td>CA</td>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>Ficain</td>
<td>K</td>
<td>L,V→T,Y</td>
<td>P</td>
<td>Proteolytic component of the latex of fig; may provide resistance to pests</td>
<td>EC 3.4.22.32</td>
<td>CA</td>
<td>C1</td>
<td></td>
</tr>
<tr>
<td>Stem bromelain</td>
<td>K</td>
<td>R</td>
<td>P</td>
<td>Most abundant cysteine endopeptidase from the stem of pineapple, may</td>
<td>EC 3.4.22.32</td>
<td>CA</td>
<td>C1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>R</td>
<td>P</td>
<td>provide resistance to pests in host plants</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

factor V (15). APC demonstrated a preference for Thr, Lys, Gln, Val, and Pro in the P<sub>2</sub> position (Fig. 2A). These data are consistent with the known physiologic substrates of APC: factor Va (cleaved at multiple sites, including Thr<sup>305</sup>Arg<sup>306</sup>—Asn<sup>307</sup> and Thr<sup>678</sup>Arg<sup>679</sup>—Lys<sup>680</sup>, where ~ represents the cleaved bond) and factor VIIa (Gln<sup>562</sup>Arg<sup>563</sup>—Gly<sup>564</sup> and Pro<sup>739</sup>Arg<sup>740</sup>—Ser<sup>741</sup>) (3). Furthermore, in a study with 25 synthetic peptidyl coumarin substrates, Ohno et al. found Boc-Leu-Ser-Thr-Arg-coumarin to be the most efficient APC substrate (16). At the P<sub>3</sub> position, both the Arg and Lys sublibraries revealed a prefer-
ence for Gln and Met. In addition to these shared preferences, the P1/H11005 Lys sublibrary also revealed efficient cleavage of substrates with P3 Lys and Ile residues, whereas the P1/H11005 Arg sublibrary showed a preference for P3 Ser and Thr residues.

Human plasma kallikrein plays an important role in contact activation of clotting, fibrinolysis, blood pressure regulation, and neutrophil stimulation. Deficiency of the zymogen, plasma prekallikrein, results in Fletcher trait, an autosomal recessive disorder associated with myocardial infarction, thromboembolism, and other vascular diseases (11). Human plasma kallikrein showed a preference for the aromatic residues Phe and Tyr, along with Lys, Asn, Thr (and, for the P1 = Arg sublibrary, Leu) in the P2 position (Fig. 2B). The enzyme showed weaker P3 specificity (Lys, Glu, and Met for P1 = Lys sublibrary and Arg, His, Ala, Glu, Leu, and Gly for P1 = Arg sublibrary). The data are consistent with the physiological substrates of plasma kallikrein: factor XII (Thr352Arg353/H11011 Val354), kininogen (Phe388Arg389/Se390), the urokinase plasminogen activator precursor (Phe157Lys158/H11011 Ile159), and human prorenin (Lys42Arg43/H11011 Leu44). These findings are also consistent with previous studies of plasma kallikrein specificity (3, 17–20). In a panel of 20 coumarin substrates, Iwanaga et al. found

---

**Fig. 2. Serine proteases.** Characterization of the specificity of serine proteases using the Ac-Ala-P3-P2-Lys-ACC-NH2 and Ac-Ala-P3-P2-Arg-ACC-NH2 substrate microarrays. Each square in the grid is colored in proportion to the average (n = 6) quantitated fluorescence intensity of the corresponding substrates after treatment with the enzyme, indicating the relative amount of cleavage. The vertical axis indicates the P2 residue and the horizontal axis indicates the P3 residues, as shown in N.
substrates with a P_2 Phe residue to be more efficiently cleaved than those with P_2 Gly or Pro residues (21). Whereas it has been noted that extending substrate length beyond the P_2 site has little effect on turnover efficiency for some synthetic substrates (20), extension to at least the P_4 site is critical for the physiological functioning of plasma kallikrein (17).

The tissue factor/factor VIIa complex initiates the extrinsic pathway of the blood coagulation cascade. Elevated levels of VIIa have been implicated in heart disease, although deficiency results in severe bleeding (22). Factor VIIa functions physiologically in complex with Ca^{2+} and tissue factor. Ca^{2+} is required by factor VIIa to become conformationally active (23), and tissue factor enhances the amidolytic efficiency of factor VIIa by 60- to 100-fold (23, 24) by causing a conformational change in factor VIIa (23, 25). As expected, we found that in the absence of tissue factor and CaCl_2, factor VIIa showed essentially no cleavage with our sublibraries (data not shown), consistent with literature reports of poor activity in the absence of these activators (11, 20, 23, 26, 27). With addition of tissue factor and CaCl_2, factor VIIa showed enhanced cleavage of the P_1 = Arg sublibrary (very low cleavage of the P_1 = Lys sublibrary was observed). A strong preference was observed for the branched amino acids Val and Thr in the P_2 position and Gln, Arg, Asn, and Pro in the P_3 position for the P_1 = Arg library (Fig. 2C). These data are consistent with the physiologic substrates of factor VIIa: factor IXa (Thr_179Arg_180Val_181 and Thr_144Arg_145Ala_146) and factor X (Thr_150Arg_151Ile_152) (3).

Factor IXaβ is involved in thrombin generation, and deficiency results in hemophilia (11). Factor IXaβ has low proteolytic activity in the absence of its physiological cofactor, coagulation factor VIIa (13, 19, 20, 26). In the absence of factor VIIa, we observed cleavage of the P_1 = Arg sublibrary (very low cleavage of the P_1 = Lys sublibrary was observed). This difference in activity was likely due to the increased stability of

---

**Fig. 3. Cysteine proteases.** Characterization of the specificity of cysteine proteases using the Ac-Ala-P_3-P_2-Lys-ACC-NH_2 and Ac-Ala-P_3-P_2-Arg-ACC-NH_2 substrate microarrays. Each square in the grid is colored in proportion to the average (n = 6) quantitated fluorescence intensity of the corresponding substrates after treatment with the enzyme, indicating the relative amount of cleavage. The vertical axis indicates the P_2 residue, and the horizontal axis indicates the P_3 residues, as shown in L.
the enzyme active site when binding substrates containing a P1 Arg residue. Within the S1 subsite, the shorter side-chain of the P1 Lys forms a water-bridged interaction with Asp189, whereas the longer side-chain of the P1 Arg stabilizes the S1 pocket via a salt bridge formed between the Arg guanidium group and the carboxylate of Asp189. The stronger salt bridge interaction results in a 30- to 40-fold preference for Arg over Lys at the P1 position (13). Factor IXa showed a preference for Phe, Pro, Tyr, Leu, Lys, and Asn in the P2 position and fairly broad specificity at the P3 position for the P1 = Arg sublibrary (Fig. 2D). These results are consistent with a previous report by Iwanaga et al., which found a preference for substrates with a P2 Pro residue over those with a P2 Ser, Thr, or Gly residue when studying a panel of 15 peptidyl coumarin substrates (21).

Human factor Xla showed a strong preference for Thr, Ser, and Asn in the P2 position for both the Arg and Lys sublibraries (Fig. 2E). A diverse set of residues was accepted at the P3 site. These results are consistent with the physiologically relevant substrate coagulation factor IX (Thr14^Arg15^Ala16 and Thr17^Arg180^Val181) (3) and with previously reported synthetic substrate studies: Iwanaga et al. reported highly efficient cleavage of P2 Ser and Thr substrates from a panel of 15 peptidyl coumarin substrates (21).

As with factor VIIa and IXa, human factor α Xlla showed poor proteolytic activity toward the P1 Lys substrates, as had been observed previously (28). The P1 = Arg sublibrary showed preference for Phe, Thr, Ser, Gly, and Asn in the P2 position and Met and Gln in the P3 position (Fig. 2F). These data correlate well with the physiologic substrates of factor Xlla: coagulation factors VII (Gly193^Arg194^Ile195) and XII (Thr352^Arg353^Val354) and the precursor of plasma kallikrein (Thr370^Arg371^Ile372) (3) and with single substrate studies that report a preference for Phe and Gly in the P2 position and disfavor Glu in the P3 position (19, 20, 29).

Complement Pathway Proteases—The complement pathway plays an important role in host defense against pathogens and infections. The complement pathway consists of multiple functionally linked proteins that mediate inflammatory responses, clear foreign pathogens, and kill susceptible cells (30, 31). These enzymes have also been implicated in prion aggregation (32) and Alzheimer disease (33).

Activated human complement factor C1s preferred the small polar residues Ser, Ala, and Gly in the P2 position and had broad specificity at the P3 position (Fig. 2G). This is consistent with the physiological substrate of C1s, C2 (Gly222^Arg223^Lys224) and with previous studies employing peptidyl substrates. McRae et al. report the fastest cleavage of substrates of the format Cbz-P2-Arg-thioester when P2 = Ser, Ala, or Gly, and Keough et al. report efficient cleavage of P2 = Ser and Gly p-nitroanilide substrates (34–36).

Compared with complement C1s, complements factor D and C1r were relatively inactive against the peptide libraries. This disparity is due to the fact that these enzymes become catalytically active through an induced fit mechanism in the presence of the physiologic substrate (for factor D) (37) or in a larger protein complex (for C1r) (38). As a consequence, peptide substrates are not sufficient to reveal the full catalytic potential of the enzymes (35). Interestingly, while cleavage of tri-, tetra-, and pentapeptides have been reported to be slow, the dipeptidase activity of these two enzymes is relatively efficient (35, 37, 39). For both factor D and C1r, we observed efficient cleavage of substrates with P3 = Arg or Lys (Fig. 2, H and I). While P3 basic residues are found in the natural substrates for activated C1r (C1r and C1s precursor) (3), this apparent specificity would also be expected if the cleavage efficiency of dipeptides is faster than that of longer substrates. In such a case, the efficient dipeptidase activity of the enzyme would result in fast cleavage of peptides containing a basic amino acid two residues from the N terminus (for the tetrapeptidyl substrates as used in this study, this would result in cleavage between a basic P3 residue and the P2 site). A second dipeptide cleavage could then occur between the P1 basic residue and the coumarin moiety, resulting in an increase in fluorescence intensity and an apparent P3 basic preference. Given the dipeptidase activity for these two enzymes, we cannot discount that the observed specificity results from off target cleavage.

Other Serine Proteases—Tryptases are serine proteases stored in mast cell granules and are known to mediate inflammatory and allergic responses. Secreted as catalytically active, noncovalently bound tetramers, there are at least four closely related tryptases: α, β, II, and III, sharing at least 93% identical amino acid sequences (40). Human lung tryptase (α) showed a preference for the polar amino acids Asn, Ser, and Thr in the P2 position for both the P1 = Arg and Lys sublibraries (Fig. 2J). A strong preference for basic residues Arg and Lys was seen in the P3 position. These results are in agreement with data from positional scanning libraries (PS-SCLs) of peptidyl coumarin substrates for α trypstatin (41, 42).

The porcine pancreatic trypsin used in our study preferred polar residues such as Ser, Asn, and His at the P3 position (Fig. 2K). Additionally, we observed cleavage of P2 Thr substrates for the P1 = Arg sublibrary and P2 Ala substrates for P1 = Lys sublibrary. This is consistent with findings in single substrate experiments reported in literature and natural substrates: chymotrypsinogen A (Ser1^Arg1^Ile19), pancreatic elastase 2A precursor (Thr11^Arg12^Val13), and pancreatic elastase 2B precursor (Ser1^Arg12^Val13) (3, 19, 20, 41, 43). Broader specificity was seen in the P3 position, with Lys, Gln, Arg, Met, Asn, Ser, and Thr all accepted for the P1 = Arg sublibrary. In general, Pro was least preferred at the P3 site. The broad acceptance of substrates has also been recently observed in PS-SCL studies described by Furlong et al. (41).

While the serine protease subtilisin Carlsberg from B. licheniformis reportedly does not prefer substrates with P1 basic residues (more commonly reported as accepting Glu,
Human cathepsin G, along with human neutrophil elastase and proteinase 3, is found in the azurophilic granules of neutrophils and is responsible for the degradation of foreign organisms, activation of CD 11b/18-bound Xa, and platelet activation (11, 47–49). Human neutrophil cathepsin G is a unique serine protease because of its chymotrypsin-like (P1 hydrophobic) and trypsin-like (P1 basic) substrate specificities (47, 48, 50–52). This dual specificity can be attributed to Glu226 residing at the bottom of the S1 pocket capable of interacting with both basic and aromatic/hydrophobic residues (47, 48, 53). For the trypsin-like activity, we observed a preference for Thr and Val at the P2 position. This is different from the reported preference for Pro or Met in the P2 position (Fig. 2L). Cleavage of tetrapeptide substrates with P1 = Lys and Arg were confirmed in conventional solution-phase well-plate assays (data not shown).

Cathepsin H, the only other exopeptidase profiled, showed a clear preference for the aliphatic hydrophobic residue Leu in the P2 position (Fig. 3B). The other residues showing a minor preference at this position were Phe, Ile, and Val (and Lys for the P2 = Arg sublibrary). Single substrate experiments with bovine cathepsin H showed a preference for aliphatic hydrophobic residues over aromatic residues (67). A broad specificity was observed for the S3 substrate.

Cathepsin K also revealed a clear preference for Leu in the P2 position and a broad specificity in the P3 position (Fig. 3C), consistent with literature reports (68, 69). We observed a minor preference for P2 Lys (for the P3 = Arg sublibrary), consistent with a previous report of cathepsin K accepting P2 basic residues (69). Unlike the homologous cathepsins V, L, and H, cathepsin K also accepted Pro at the P2 position. This P2 Pro acceptance is required for the unique collagenase activity of cathepsin K (68) and has been noted previously in both PS-SCLs (68) and fluorescence resonance energy transfer (FRET) studies (69). While cooperative interactions between substrate subsites cannot be detected using PS-SCLs, the study by Alves et al. using FRET substrates had noted that the acceptance of P2 Pro residues was dependent on the neighboring substrate residues. Of the six P2 Pro substrates tested in the FRET study, the most efficient cleavage was observed with substrates containing a P3 Lys residue (69). We found that for our Ac-Ala-P3-Pro-(Arg/Lys)-ACC-NH2 substrates, P3 Lys substrates were among the most efficient (along with P3 = Pro and Gly for the P1 = Arg sublibrary and P3 = Gly, Ile, and Leu for the P1 = Lys sublibrary).

Cathepsin L showed a preference for the bulky hydrophobic residues Val, Phe, and Leu in the P2 site (Fig. 3D), with considerably less acceptance of P2 Pro substrates. Prior studies have attributed the preference for large aromatic and bulky residues at the P2 site to the presence of the small residue Ala1705 (papain numbering) at the bottom of the S2 pocket, which would suggest the presence of a large open-ended pocket (54, 58). In cathepsin K, this residue is a Leu, a change that is implicated in the differential acceptance of P2 Pro residues between the two enzymes (68). The P2 specificity observed in the microarray assay is in accordance with Maciewicz et al. who found cathepsin L preferred the substrate Cbz-Phe-Arg-coumarin (62). Similarly, using a panel of FRET substrates to study cathepsin L, Puzer et al. report a preference for substrates with P2 Leu and Phe (70). At the P3 position, basic residues were preferred, consistent with literature reports (68, 70).
Microarrays for Profiling Protease Substrate Specificity

Cathepsin S is homologous to the other endopeptidase cathepsins as demonstrated by the superposition of its structure onto those of cathepsins K, V, and L with root mean square deviation values of <1 Å (54). Cathepsin S showed a preference for Leu in the P2 position in both sublibraries with a broad specificity for the P3 position (Fig. 3E). Similar P2 specificity for hydrophobic residues has been observed in single substrate studies for bovine and human cathepsin S (58, 71). The diminished acceptance of P2 Val residues relative to cathepsin V and cathepsin L may be attributed to the presence of a Val residue that more greatly restricts the entrance to the S2 pocket in cathepsin S relative to cathepsin V (which contains a Leu residue at this position) and Cat L (which has a Met residue at this position).

Cathepsin V, like cathepsin L, shows preference for Leu and Val in the P2 position with broad specificity in the P3 position (Fig. 3F), consistent with literature reports (70, 72). This similarity can be attributed to the fact that both cathepsins V and L contain an Ala205 in the S2 pocket, allowing for accommodation of large residues at position 205, resulting in a deeper S2 subsite pocket (72). Puze et al. reported that cathepsin V preferred Leu over Phe in the P2 position in studies with a panel of dipetidyl-coumarin substrate studies (70, 73), showing that cathepsin V has an intermediate specificity between cathepsin L and S.

Rhodesain—Rhodesain is a papain-like cysteine protease expressed by Trypanosoma brucei rhodesiensis, a parasitic protozoa. The parasite is the cause of sleeping sickness, which infects nearly 25,000–50,000 people yearly, and if left untreated, is fatal. The current chemotherapeutic treatments are poorly effective for the disease. In fact, the only treatment effective against T. b. rhodesiensis, melarsoprol, results in potentially fatal encephalopathy in 5–10% of treated patients (8). As new inhibitors are required to combat the disease, elucidating the substrate specificity is of prime importance. While a comprehensive study of the substrate specificity of rhodesain has not previously been reported, an initial study by Caffrey et al. employing four dipetidyl coumarin substrates found efficient cleavage of Cbz-Val-Arg-coumarin and Cbz-Phe-Arg-coumarin (8). We show here that the protease shows a very strong preference for hydrophobic and aromatic residues (Leu, Val, Phe, Tyr, and a minor preference for Ile, Trp, Met) in the P2 position and has a broad specificity for non-acidic residues in the P3 position (Fig. 3G).

Plant Cysteine Proteases—Papain, a plant enzyme from the papaya fruit, is the archetypal cysteine protease. Papain showed a strong preference for Val in the P2 position and Pro in the P3 position (Fig. 3H). These data are consistent with PS-SCL and quenched combinatorial library studies reported in the literature (74, 75).

Chymopapain from papaya latex, ficain from fig tree latex, and stem bromelain from pineapple are fruit enzymes belonging to the papain family of cysteine proteases. Despite their homology, each displayed unique substrate specificity in the microarray assay. Chymopapain showed strong P2 specificity for the branched amino acids Val, Thr, and Ile (Fig. 3I). Ficain preferred Leu, Val, and Thr in the P2 position (Fig. 3J), and stem bromelain showed a strict preference for Arg in the P2 position (Fig. 3K). This strong preference of stem bromelain for Arg at both the P1 and P2 sites has previously been noted for other synthetic substrates; Cbz-Arg-Arg-coumarin and Cbz-Arg-Arg-p-nitroanilide substrates were found to be the most efficiently cleaved substrates in a panel of 18 potential substrates by Rowan et al. (76). In the P3 position, both ficain and stem bromelain, like papain, showed a preference for substrates with Pro at this site.

CONCLUSIONS

Microarray-based methodologies are becoming increasingly important proteomic tools. However, to date most microarray-based methodologies have relied on covalent attachment of the library members (compound, peptides, proteins) under investigation to the array surface. Direct attachment introduces the potential for solid-phase effects, such as inaccessibility and nonspecific binding (77, 78). The solution-phase microarray approach discussed here minimizes these risks because the glycerol used to form the array microspots is a water mimic capable of forming multiple hydrogen bonds. As such, the risk of protein aggregation/denaturation, compound precipitation, and nonspecific interactions is reduced (79–81). Furthermore, the glycerol provides the necessary lubrication for enzymatic catalysis (79–81). While glycerol has been noted to result in increased proteolytic efficiency through oligomerization and structural stabilization for some proteases (82–84), correlation studies between microarray and well-plate data indicate no dramatic variations in protease specificity (6). Here we have demonstrated the broad generality of this solution-phase microarray approach by efficiently profiling the specificity of 13 serine proteases and 11 cysteine proteases with a spatially separated, 722-member fluorogenic substrate microarray.

The high throughput nature of the solution-phase microarray enabled us to map substrate specificity in a highly parallel and miniaturized format with minimal consumption of biological samples (5, 6). At the high density that the fluorogenic substrate library was arrayed, a complete Ac-P4-P3-P2-P1-ACC-NH2 library with one member held fixed (193 compounds) can be accommodated on one or two microarrays, allowing rapid elucidation of specificities of hundreds of distinct enzymes. Also, differential proteolytic activities of cellular lysates or other complex biological fluids (e.g. blood) can be characterized or phenotyped with potential for diagnostic applications. The method enables us to decipher the cooperative interactions between substrate subsites and provides access to functional evolutionary data (6). The assay method is amenable to drug discovery, metabolic/toxicity profiling, and drug design applications.

---

Microarrays for Profiling Protease Substrate Specificity

and may be useful for engineering potent and specific substrates and inhibitors (5, 6).

Acknowledgment—We thank Dustin J. Malý for assistance with the preparation of the P$_{1} = $ Lys library.

* This work was supported by the National Institutes of Health Grants GM54051 (to J. A. E.) and HL56621 (to S. L. D.) and the University of California Cancer Research Coordination Committee. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

§ These authors contributed equally to this work.

To whom correspondence should be addressed: Jonathan A. Ellman, Dept. of Chemistry, University of California, Berkeley, CA 94720-1460. Tel.: 510-642-4488; Fax: 510-642-8369; E-mail: jellman@uclink.berkeley.edu. Scott L. Diamond, Institute for Medicine and Engineering, University of Pennsylvania, 1024 Vagelos Research Laboratory, 3340 Smith Walk, Philadelphia, PA 19104. Tel.: 215-573-5702; Fax: 215-573-7227; E-mail: sld@seas.upenn.edu.

REFERENCES


Microarrays for Profiling Protease Substrate Specificity

complex, *EMBO J.* 21, 231–239


Microarrays for Profiling Protease Substrate Specificity