THE JOURNAL OF BIOLOGICAL CHEMISTRY VOL. 281, NO. 18, pp. 12824–12832, May 5, 2006
© 2006 by The American Society for Biochemistry and Molecular Biology, Inc. Printed in the U.S.A.

Substrate Profiling of Cysteine Proteases Using a Combinatorial Peptide Library Identifies Functionally Unique Specificities*[S]

Received for publication, December 15, 2005, and in revised form, February 10, 2006 Published, JBC Papers in Press, March 6, 2006, DOI 10.1074/jbc.M513331200

Youngchool Choe[‡], Francesco Leonetti[§], Doron C. Greenbaum[‡], Fabien Lecaille[¶], Matthew Bogyo[‡], Dieter Brömme[¶], Jonathan A. Ellman[§], and Charles S. Craik[‡1]

From the [‡]Department of Pharmaceutical Chemistry, University of California at San Francisco, California 94143, the [§]Chemistry Department, University of California at Berkeley, California 94720, and the [¶]Department of Human Genetics, Mount Sinai School of Medicine, New York, New York 10029

The substrate specificities of papain-like cysteine proteases (clan CA, family C1) papain, bromelain, and human cathepsins L, V, K, S, F, B, and five proteases of parasitic origin were studied using a completely diversified positional scanning synthetic combinatorial library. A bifunctional coumarin fluorophore was used that facilitated synthesis of the library and individual peptide substrates. The library has a total of 160,000 tetrapeptide substrate sequences completely randomizing each of the P1, P2, P3, and P4 positions with 20 amino acids. A microtiter plate assay format permitted a rapid determination of the specificity profile of each enzyme. Individual peptide substrates were then synthesized and tested for a quantitative determination of the specificity of the human cathepsins. Despite the conserved three-dimensional structure and similar substrate specificity of the enzymes studied, distinct amino acid preferences that differentiate each enzyme were identified. The specificities of cathepsins K and S partially match the cleavage site sequences in their physiological substrates. Capitalizing on its unique preference for proline and glycine at the P2 and P3 positions, respectively, selective substrates and a substrate-based inhibitor were developed for cathepsin K. A cluster analysis of the proteases based on the complete specificity profile provided a functional characterization distinct from standard sequence analysis. This approach provides useful information for developing selective chemical probes to study protease-related pathologies and physiologies.

Proteases hydrolyze amide bonds in proteins and peptides and represent one of the largest and most important protein families known. They comprise over 2% of the human genome and play diverse physiological roles (merops.sanger.ac.uk) (1). The substrate specificity of a protease enables the enzyme to preferentially cleave its substrates in the presence of other peptides or proteins. Therefore, specificity information can provide clues about the biological function of the protease and aid in the design of efficient substrates and potent, selective inhibitors. Various methods including both biological and chemical-based approaches to study protease specificity have been developed and were recently reviewed (2).

Positional scanning synthetic combinatorial libraries (PS-SCLs)[2] of fluorogenic substrates have emerged as useful reagents for the rapid and exhaustive determination of protease specificity (3). A peptide-based PS-SCL is composed of sublibraries in which one peptide position is fixed with an amino acid, whereas the remaining positions contain an equimolar mixture of amino acids. Assaying proteases with these sublibraries rapidly establishes the amino acid preferences at the defined position. Initially, the substrate specificities of caspases and granzyme B were profiled using PS-SCLs with the P1 position fixed as an aspartic acid.

The limitations of the original P1 fixed libraries were overcome through the development of a modified coumarin, 7-amino-4-carbamoylmethylcoumarin (ACC) fluorogenic leaving group. The bifunctional nature of this enables straightforward solid-phase synthesis of libraries containing any amino acid at the P1 position. Early applications involved the use of a P1-diverse PS-SCL in combination with several P1-fixed PS-SCLs to study P1 and P2-P3-P4 specificity, respectively (4–7). We report the preparation of a completely diversified PS-SCL of ACC-based substrates. This library permits the determination of P1-P2-P3-P4 specificity of proteases regardless of their P1 specificity. Using this complete diverse library, numerous proteases from various sources including humans, parasites, bacteria, and viruses have been profiled. As a representative family, we present a study of the substrate specificity of papain-like cysteine proteases.

The papain-like cysteine proteases, which include plant enzymes papain and bromelain, human cysteine cathepsins (B, H, L, S, C, K, O, F, V, X, W), and parasite proteases cruzain and falcipains, have been characterized as key enzymes in many biological and pathological events (8–11). As a result, many of them represent particularly attractive drug targets. Of particular interest to these studies is cathepsin K, a cysteine protease implicated in osteoporosis and other diseases (12). The substrate binding pocket of these proteases can be divided into seven substrate binding subsites, S4 to S3′, that interact with P4 to P3′ residues of substrates (13). Hydrolysis occurs at the scissile bond between P1 and P1′. Among these, S3 and S2′ subsites interact with substrates through only side chain contacts, and their interactions spread over a relatively wide area. In contrast, the S2, S1, and S1′ subsites involve both main chain and side chain contacts. These recognition properties in combi-

* This work was supported by National Institutes of Health Grants CA72006 (to C. S. C. and J. A. E.), AI35707 (to C. S. C.), GM54051 (to J. A. E.), and AR46182 and AR48669 (to D. B.). 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.

[S] The on-line version of this article (available at http://www.jbc.org) contains two supplemental figures.

[1] To whom correspondence should be addressed: Dept. of Pharmaceutical Chemistry, University of California at San Francisco, Box 2280, San Francisco, CA 94143. Tel.: 415-476-8146; Fax: 415-502-8298; E-mail: craik@cgl.ucsf.edu.

[2] The abbreviations used are: PS-SCL, positional scanning synthetic combinatorial libraries; ACC, 7-amino-4-carbamoylmethylcoumarin; DMF, *N,N*-dimethyl formamide; DICI, diisopropylcarbodiimide; HOBt, 1-hydroxybenzotriazole; Ac, acetyl; Z, benzyloxycarbonyl; AMC, 7-amino-4-methylcoumarin; HPLC, high pressure liquid chromatography; AOMK, acyloxylmethyl ketone; Fmoc, *N*-(9-fluorenyl)methoxycarbonyl.

nation with the well conserved structure of the family result in broad and similar specificities for the papain-like proteases (14).

In this study, however, the complete diverse PS-SCL and the cluster analysis of the resulting specificity information have revealed distinctive differences between the members of this class. The utility of this library and the specificity information obtained using it is well exemplified by the development of specific ACC-based substrates and an acyloxymethyl ketone inhibitor for cathepsin K.

EXPERIMENTAL PROCEDURES

Materials—Chemicals were obtained from commercial suppliers and used without further purification, unless otherwise stated. Rink amide AM resin and Fmoc-amino acids were purchased from Novabiochem. Anhydrous low amine content *N,N*-dimethyl formamide (DMF) was from EM Science. *O*-(7-azabenzotriazol-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate was from PerSeptive Biosystems. Diisopropylcarbodiimide (DICI), 1-hydroxybenzotriazole (HOBt), trifluoroacetic acid, and triisopropylsilane were from Aldrich. Synthetic substrates, Z-FR-AMC (7-amino-4-methylcoumarin) and Z-LR-AMC, were purchased from Bachem.

Papain, and pineapple stem bromelain were purchased from Sigma. Human cathepsin B was purchased from Cortex Biochem. Heterologous expression, purification, and active site titration were performed as described previously for human cathepsin F (15), K (12), L (16), S (15), and V (17). Five papain-like cysteine proteases of parasite origin were kind gifts from Drs. J. H. McKerrow, M. Sajid, and C. R. Caffrey. Rhodesain from *Trypanosoma brucei rhodesiense*, cruzain from *Trypanosoma cruzi*, a cathepsin L-like protease from *Leishmania mexicana*, and cathepsin B-like proteases 1 and 2 from *Schistosoma mansoni* were heterologously expressed and purified as described previously elsewhere (18, 19).

Synthesis of the Complete Diverse Tetrapeptide-ACC PS-SCL—The preparation of ACC and P1-substituted ACC was carried out as described previously using an Argonaut Quest 210 organic synthesizer (4, 20, 21). The substitution level of the resin (0.63 mmol/g) was determined by a spectrophotometric Fmoc quantitative assay (22). The synthesis of the library was performed using a MultiChem 96-well synthesis apparatus (Robbins Scientific). To prepare the P1 part of the P1 library, each of 20 Fmoc-amino acids (omitting cysteine and including norleucine) bound to ACC-resin (0.1 mmol) was added to the wells of the reaction apparatus. The use of norleucine in the amino acid pool is to increase the amount of information provided by the substrate specificity screen. Since norleucine contains the same number of carbons as leucine and isoleucine and has a similar unbranched chain structure as lysine, it provides additional information in probing the extended substrate specificity of proteases. For the P2, P3, and P4 libraries, an isokinetic mixture of 20 Fmoc-P1 amino acids bound to ACC-resin (2 mmol per each library) was prepared by shaking the slurry in DMF for 2 h (4, 23). After filtration, the resin was dried and then split in the wells of the reaction apparatus (0.1 mmol/well, 20 wells per each library). DMF was added to each well to solvate the Fmoc-P1 amino acid-ACC-resin, and gentle agitation for 30 min followed. To remove the Fmoc protection group, the DMF was drained, and a solution of 20% piperidine in DMF (4 ml/well) was added to the resin and agitated for 30 min. The piperidine solution was then removed by filtration, and the resin was thoroughly washed with DMF.

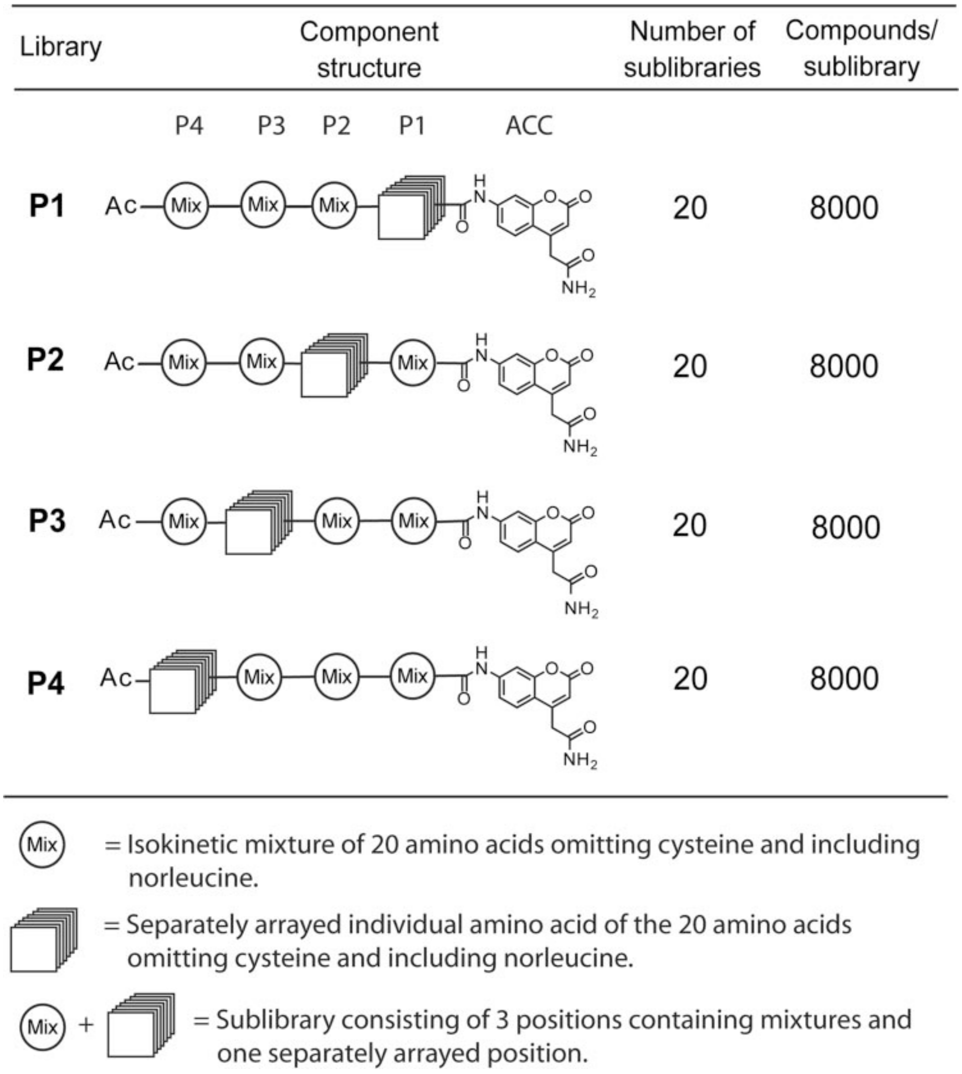
To install P2 amino acids to the P2 library, 20 individual Fmoc-amino acids (10 eq, 1 mmol) were preactivated in separate vials using HOBt (10 eq, 1 mmol) and DICI (10 eq, 1 mmol) in DMF and added to the wells for the P2 library. To couple P2 amino acids to the P1, P3, and P4 libraries, an isokinetic mixture of the 20 Fmoc-amino acids (20 mmol per each library, 10 eq/well) was preactivated with HOBt (20 mmol) and DICI (20 mmol) in DMF. The solution was then added to each well for the P1, P3, and P4 libraries, and a 3-h agitation for coupling followed. When finished, the solution was drained, and the resin was thoroughly washed with DMF. The P3 and P4 positions were installed in the same manner except using 20 individual preactivated Fmoc-amino acids for the P3 position of the P3 library and for the P4 position of the P4 library, whereas a preactivated isokinetic mixture was used for the remaining positions.

After the synthesis of the peptide portion was completed, the Fmoc blocking group of the P4 amino acids was removed, and the resin in each well, after washing with DMF, was treated with a capping solution consisting of AcOH (80 mmol), HOBt (80 mmol), and DICI (80 mmol) in DMF. After being agitated for 4 h, the resin was washed with DMF and then with CH_2Cl_2. The substrates were cleaved from the resin by treating for 1 h with a solution of trifluoroacetic acid:triisopropylsilane:H_2O (95:2.5:2.5, 3 ml/well), and the collected material was lyophilized. The final products were dissolved in Me_2SO to a concentration of 25 mM and stored at −20 °C until use.

Synthesis of Individual Substrates and an Irreversible Inhibitor—The synthesis of individual peptide substrates was carried out using the same method employed for the complete diverse PS-SCL until the trifluoroacetic acid cleavage step. The peptide-ACC substrates cleaved from the resin were precipitated with *t*-butyl methyl ether. After any residual ether was evaporated, the resulting products were subjected to reverse phase preparatory HPLC (Rabbit HPLC with a Vydac C18 column, 0–95% CH_3CN gradient with 0.01% trifluoroacetic acid). Matrix-assisted laser desorption/ionization mass spectrometry (Voyager, Applied Biosystems) was used to confirm the molecular weights of the purified substrates. The final purified substrates, after lyophilization, were dissolved in Me_2SO and stored at −20 °C until use.

An irreversible inhibitor, Ac-HGPR-acyloxylmethyl ketone (AOMK), was also designed based on the library assay results of cathepsin K and other cathepsins. The synthesis was carried out using conditions similar to those described elsewhere (24, 25). The inhibitor was purified, and its molecular weight was confirmed as described for the preparation of individual substrates.

PS-SCL Assay—The cysteine proteases were assayed at 25 °C in a buffer containing 100 mM sodium acetate (pH 5.5), 100 mM NaCl, 10 mM dithiothreitol, 1 mM EDTA, 0.01% Brij-35, and 1% Me_2SO (from the substrates). Aliquots of 25 nmol in 1 μl from each of 20 sublibraries of the P1, P2, P3, and P4 libraries were added to the wells of a 96-well Microfluor-1 U-bottom plate (Dynex Technologies). The final concentration of each compound of the 8,000 compounds/well was 31.25 nM in 100-μl final reaction volume. The assays were initiated by the addition of preactivated enzyme and monitored fluorometrically with a SpectraMax Gemini fluorescence spectrometer (Molecular Devices) with excitation at 380 nm, emission at 460 nm, and cutoff at 435 nm (4, 23). The excitation and emission maxima of the peptide-conjugated ACC substrates are 325 and 400 nm, respectively. Cleavage of the substrate by a protease to release the free ACC results in a shift of the excitation and emission maxima to 350 and 450 nm, respectively. An excitation of 380 nm and an emission at 460 nm is used to maximize the signal of the ACC group over the background signal of the uncleaved substrate. In addition, the ACC fluorophore has an ~2.8-fold higher fluorescence yield than AMC at the excitation and emission wavelengths of 380 and 460 nm. The enhanced fluorescence of the ACC group allows for the more sensitive detection of protease activity.



= Isokinetic mixture of 20 amino acids omitting cysteine and including norleucine.

= Separately arrayed individual amino acid of the 20 amino acids omitting cysteine and including norleucine.

= Sublibrary consisting of 3 positions containing mixtures and one separately arrayed position.

FIGURE 1. **The complete diverse PS-SCL.** The library was synthesized using ACC, a fluorogenic leaving group. *Mix* means an equimolar mixture of 20 amino acids (omitting cysteine and including norleucine). The library consists of P1, P2, P3, and P4 libraries, and each of them consists of 20 sublibraries. Each sublibrary is composed of 8,000 species of tetrapeptide-ACC substrates. The library totals 160,000 compound diversity.

Cluster Analysis of Specificity Data from the Complete Diverse PS-SCL Assays—To compare the specificity information with amino acid sequence information, the results from the library assays were clustered. First, the activity rates from the library assay were converted to values in a range from −1 to 1 by assigning a value of 1 to the amino acids that showed the strongest activity in each library (P1 to P4), whereas amino acids that showed no activity were assigned a value of −1. The results were analyzed with the program CLUSTER and displayed in a tree diagram by using TreeView (26). The P1, P2, P3, and P4 specificities were clustered together to compare with the sequence alignment and also separately for more detailed comparison. The structure-based amino acid sequence alignment of active protease domains was performed using the CLUSTAL_W program (MacVector, Accelrys Inc.). The primary sequences were taken from the SWISS-PROT or GenBank^{TM} databases.

Kinetic Analysis of Individual Peptide-ACC Substrates and an Irreversible Inhibitor—Michaelis-Menten steady state kinetic analysis was used to determine the kinetic constants of each substrate and protease pair. The final concentration of substrates ranged from 0.25 μM to 1 mM, and the concentration of Me_2SO in the assays was less than 2% (v/v). The concentrations of cathepsins K, L, and B were 20, 1.37, and 1 nM, respectively. All kinetic assays were performed at 25 °C in triplicate. The hydrolysis of ACC substrates was monitored fluorometrically using the assay conditions described for the complete diverse library assay. The Kaleidagraph program (Synergy software) and Equation 1 were used to analyze the results and calculate k_{cat}, K_m, and k_{cat}/K_m.

$$v_0 = \frac{k_{\text{cat}}[\text{E}]_0}{1 + (K_m/[S]_0)} \qquad \text{(Eq. 1)}$$

Kinetic characterization of the inhibitor Ac-HGPR-AOMK was performed as described previously (18). Briefly, residual activity of human cathepsin K (20 nM), cathepsin L (1.37 nM), and cathepsin B (1 nM) were assayed with a synthetic substrate Z-FR-AMC (final 50 μM) under the complete diverse library assay conditions. Values for the pseudo-first-order rate constant k_{obs} at each concentration of inhibitor $[I]_0$ were computed for individual curves by fitting the data to Equation 2 when $[I]_0 \geq 10$ times the enzyme concentration $[E]_0$, where $[P]$ is the concentration of product formed over time t, and v_o is the initial velocity of the reaction.

$$[p] = \frac{v_0}{k_{\text{obs}}}(1 - \exp^{-k_{\text{obs}} \cdot t}) \qquad \text{(Eq. 2)}$$

Non-linear regression analysis to determine the inactivation constant k_{inact} and the inhibition constant k_i was performed using the Kaleidagraph program and Equation 3. $[S]_0$ is the concentration of substrate.

$$k_{\text{obs}} = \frac{k_{\text{inact}}[\text{I}]_0}{([\text{I}]_0 + K_i(1 + [S]_0/K_m)} \qquad \text{(Eq. 3)}$$

Competition Labeling Assay to Assess the Selectivity of the Inhibitor—To compare the inhibitory activity and selectivity of Ac-HGPR-AOMK against human cathepsins L, B, and K, a competition labeling experiment was carried out as described previously (27). The intensity of bands inversely reflects the binding efficacy of Ac-HGPR-AOMK to the given proteases. This was measured using densitometry for quantitative comparison.

RESULTS

Synthesis of the Complete Diverse PS-SCL—A completely diversified PS-SCL with the general structure of acetyl-P4-P3-P2-P1-ACC was synthesized using ACC, a bifunctional fluorophore leaving group with chemically labile sites for peptide synthesis and attachment to solid support. The library consists of P1, P2, P3, and P4 libraries in which the corresponding P1, P2, P3, or P4 position is fixed with one of 20 amino acids (omitting cysteine and including norleucine), whereas the remaining three positions contain an equimolar mixture of these amino acids (Fig. 1). As a result, each of the P1–P4 libraries has 20 sublibraries that contain a mixture of 8,000 (=203) species of tetrapeptide fluorogenic substrates. As a whole, the complete diverse library contains 160,000 unique tetrapeptide substrates.

The library was functionally characterized using the enzymes trypsin, papain, and legumain, the substrate specificities of which are well known (Supplemental Data 1). The P1-diverse PS-SCL and various P1-fixed PS-SCLs were also used to profile these enzymes and human cysteine cathepsins (data not shown). The results obtained using these libraries were compared with those obtained from the complete diverse PS-SCL assays. The results were in good agreement with each other, providing confidence that the complete diverse library was not biased.

Specificity of Papain-like Cysteine Proteases—The specificities of papain, bromelain, and human cysteine cathepsins L, V, S, K, F, and B were determined using the complete diverse PS-SCL (Fig. 2). The specificities of cruzain, rhodesain, cathepsin L-like protease from *L. mexicana*, and cathepsin B-like proteases 1 and 2 from *S. mansoni* were similarly determined using the complete diverse PS-SCL. Nearly all of these proteases displayed a preference for hydrophobic amino acids at the P2 position except bromelain, which strongly favored basic amino acids such as arginine. All six human cathepsins generally matched papain in specificity by preferring arginine and lysine at the P1 position, strictly hydrophobic amino acids at the P2 position, and broader specificities at the P3 and P4 positions. However, the complete diverse PS-SCL assay demonstrated the differences in the chemical characteristics of the favored amino acids at the P2 position and more subtle P3 specificity.

At the P2 position, the substrate specificity profile of cathepsin L shows a preference for aromatic residues (phenylalanine, tryptophan, tyrosine) over aliphatic amino acids (valine, leucine), which distinguishes it from cathepsins K and S (Fig. 2, *a*, c, and *d*). Cathepsins K and S have been described to prefer branched hydrophobic residues at the P2 position, whereas cathepsin L has been shown to favor aromatic amino acids (12, 28). These previous studies are in good agreement with the complete diverse PS-SCL assay results, which showed that cathepsins K and S exclusively favored aliphatic amino acids (leucine, isoleucine, valine, methionine) at the P2 position. Cathepsin V, which is the closest to cathepsin L in terms of sequence identity, showed a preference similar to cathepsin L, favoring aromatic amino acids (tryptophan, tyrosine) over aliphatic amino acids (leucine, valine) (Fig. 2*b*). Cathepsin V accepted phenylalanine and leucine equally well at the P2 position, which is also consistent with previously published studies (17, 29). The P2 specificity of cathepsin F was similar to that of cathepsin K except for the proline preference of cathepsin K (Fig. 2*e*). It is noteworthy that cathepsin F accepted aspartic acid at the P2 and P3 positions, whereas none of the other cathepsins tolerate this acidic amino acid residue at either position.

The complete diverse library assay confirmed that cathepsin B has much broader P2 specificity. Cathepsin B also showed stronger activity with the P1 library than with the P2 library, which significantly contrasts with the cathepsin L group proteases cathepsins L, V, S, and K (Fig. 2*f*). The library assay also confirmed that cathepsin B accepts arginine well at the P2 position, whereas the other cathepsins did not show any noticeable activity with this amino acid, in agreement with previous studies (30).

At the P3 position, cathepsins L and S showed similarly broad specificity but also displayed noticeable preference for basic amino acids (lysine, arginine) and some aliphatic amino acids (norleucine, leucine, methionine, isoleucine), whereas cathepsin V favored proline and norleucine. The library assay also indicated that cathepsin B has a narrower P3 specificity (norleucine, leucine, methionine, lysine, arginine) than it was previously believed to have.

Specificity of Papain-like Proteases of Parasitic Origin—All the parasite proteases tested showed very similar P1 specificity as human cathepsins. Again, the interaction in the S2 subsite appears to be the predominant specificity-defining factor. Cruzain, rhodesain, and cathepsin L-like protease from *L. mexicana* showed specificity that is similar to that of human cathepsins L and V, whereas cathepsin B-like proteases 1 and 2 from *S. mansoni* showed much broader P2 specificity, accepting more amino acids than the aforementioned parasite proteases (Supplemental Data 2). It is noteworthy that cathepsin B-like protease 1 has less similar P2 and P3 specificities to human cathepsin B, although it has higher sequence identity to human cathepsin B than cathepsin B-like protease 2 has. This shows that a simple sequence comparison is not sufficient to deduce the specificity of an homologous protease.

Unique Substrate Specificity of Human Cathepsin K—With the complete diverse PS-SCL, cathepsin K displayed the most distinguishing substrate specificity among the human cathepsins tested. The protease exclusively favored aliphatic amino acids (leucine, isoleucine) at the P2 position, unlike cathepsins L and V, which accepted both aromatic and aliphatic amino acids. Most distinctively, cathepsin K favored proline and glycine at the P2 and P3 positions, respectively, whereas neither of those amino acids, especially proline, were preferred by the other human cathepsins (in Fig. 2*c*, the proline preference is designated with a *black bar*). This is in agreement with a previous study that showed that Z-GPR-AMC shows a partial selectivity for cathepsin K (31).

Comparison of Substrate Specificity by PS-SCL and Physiological Substrate Specificity of Cathepsins K and S—Cathepsins B, H, and L are ubiquitous, making it difficult to clearly identify their cognate natural substrates. However, the tissue and cellular distribution of cathepsins K and S is more restricted than that of cathepsins B, H, and L. As a result, more information on their physiological substrates is available. Table 1 highlights the similarities between the specificities determined by the PS-SCL and the physiological substrates for cathepsin K (32) and cathepsin S (33).

Cluster Analysis of Substrate Specificity—To compare the specificity data with the primary sequence information, 13 papain-like cysteine proteases including papain, bromelain, human cathepsins L, V, S, K, F, B, and parasitic proteases such as cruzain, rhodesain, cathepsin L (*L. mexicana*), and cathepsin B-like proteases 1 and 2 (*S. mansoni*) were clustered based on their specificity profiles using the program CLUS-

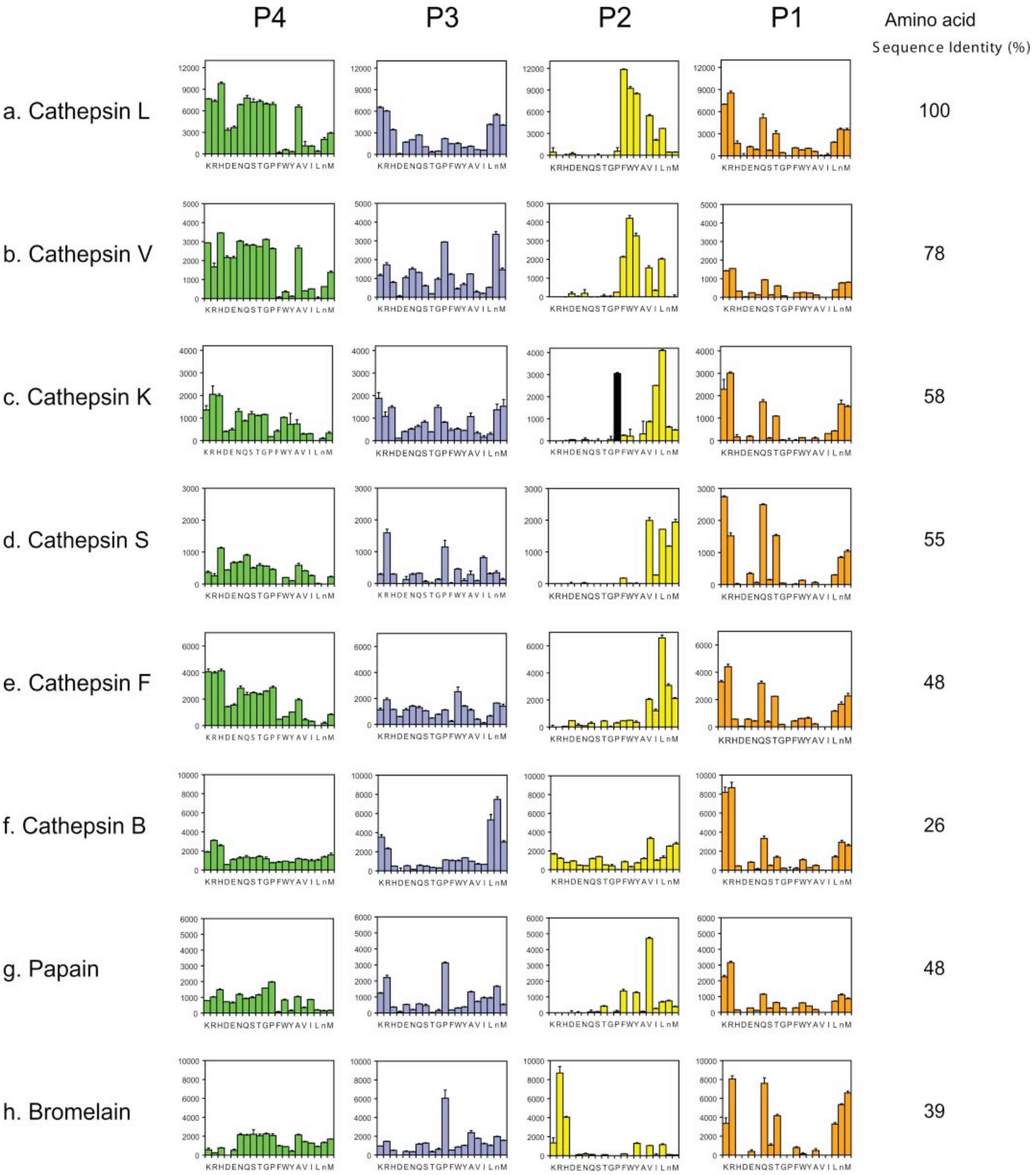


FIGURE 2. **Substrate specificity of human cysteine cathepsins determined using the complete diverse substrate library.** The results are presented in the order of amino acid sequence identity to human cathepsin L. All assays were performed in triplicate. The height of the bars and the error bars denotes mean ± S.D. The *y* axis is the picomolar fluorophore produced per second. The *x* axis indicates 20 amino acids held constant at each position, designated by the single-letter code (*n*, norleucine). The amino acids have been grouped based on their chemical characteristics of side chain residues (acidic, basic, polar, aromatic, and aliphatic amino acids). The unique P2 specificity of human cathepsin K for proline is marked as a *black bar*.

TER (26). For comparison, a structure-based amino acid sequence alignment of these proteases is also shown (Fig. 3*a*). The cluster analysis was performed by analyzing the entire P1–P4 specificity profile or each position separately for each protease.

The resulting tree diagrams of 13 proteases display specificity similarity and therefore represent functional similarity among the proteases. The clustering result of the entire P1–P4 specificity profile nearly matched the sequence alignment (Fig. 3*b*). Human cathepsin V and all

TABLE 1

Specificity and physiological substrate cleavage sequences of human cathepsins K and S

A comparison of the favored amino acids determined using the complete diverse PS-SCL and suggested physiological substrate cleavage sequences of the proteases is shown. Matching amino acids are highlighted in bold typeface. X denotes broad specificity.

	P4	P3	P2	P1
Cathepsin K				
Specificity by the complete diverse PS-SCL	X	**K**/**G**/H/M	**P**/I/ L	**Q**/R/K
Cleavage site sequences	P	**G**	**P**	A
Collagen Type-I, II triple helical domains	P	**G**	V	S
	G	**K**	**P**	G
	E	**G**	**P**	**Q**
Cathepsin S				
Specificity by the complete diverse PS-SCL	X	**R**/P/I	**M**/**L**/V/n	**K**/**R**/Q
Cleavage site sequences	L	**R**	**M**	**K**
Human MHC class II-associated invariant chain	E	N	**L**	**R**

the cathepsin L-like proteases of parasite origin were grouped together with human cathepsin L. Cathepsin S and papain formed the closest group to the cathepsin L group followed by the cathepsin F group. Cathepsin K was located far from the cathepsin L group due to its peculiar substrate specificity. Cathepsin B and cathepsin B-like proteases were clustered together as the farthest group from cathepsin L.

The cluster analysis of each P1, P2, P3, and P4 specificity data offered greater resolution (Fig. 3*c*). At the P1 position, the specificity of the cysteine proteases appeared to be highly conserved. Only cathepsin S was distinguished by its unique P1 specificity, favoring lysine and glutamine over arginine. Although there is little diversity at this position, cathepsin B-related proteases were still grouped separately from the cathepsin L group (Fig. 3*c*, *panel P1*). At the P2 position, two distinctive specificity groups were apparent. The first group that includes cathepsin L group proteases (cathepsins L and V, cruzain, rhodesain, and *L. mexicana* cathepsin L) accepted hydrophobic amino acids but preferred aromatic amino acids to aliphatic amino acids (Fig. 3*c*, *panel P2*). Cathepsins F and K were located relatively far from cathepsin L within

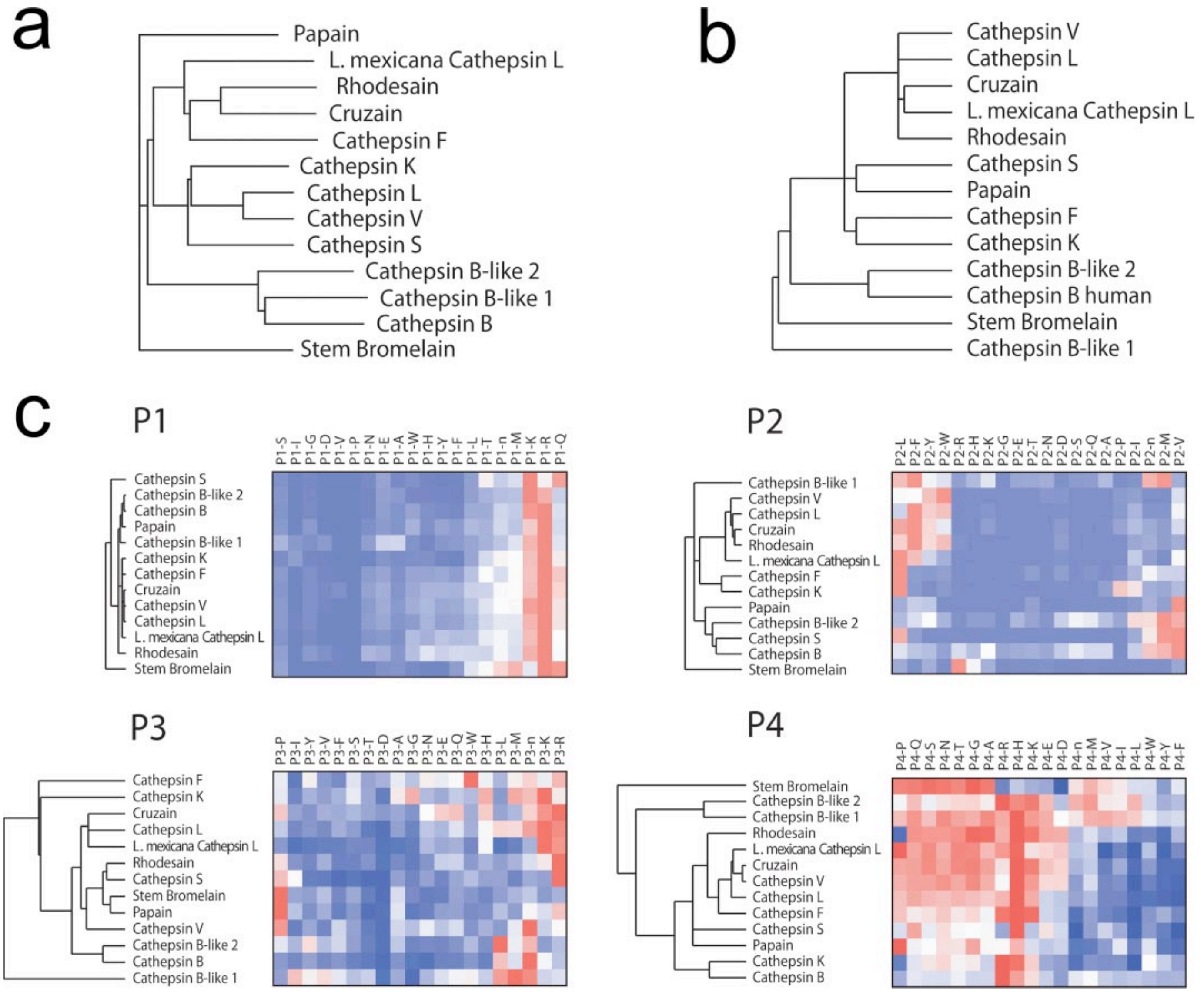


FIGURE 3. **Cluster analysis of the substrate specificity data of 13 papain-like cysteine proteases.** *a*, for comparison, sequence homology alignment of these cysteine proteases is presented. Clustal_W (v1.4) multiple alignment function of MacVector program was used. Alignment parameters are as follows: open gap penalty = 10.0; extended gap penalty = 0.1; delay divergent = 40%; gap distance = 8; similarity matrix = Blosum. *b*, entire P1–P4 specificity profiles were clustered together. *c*, each of the P1–P4 specificity profiles was clustered separately. The tree structure was obtained by hierarchical clustering and indicates the degree of similarity with the relative height of the lines connecting profiles.

TABLE 2

Kinetic analysis of peptide-ACC substrates with human cathepsins

Ac-HGPR-ACC and Ac-HGPN-ACC were designed based on the unique specificity of cathepsin K. Ac-R-ACC was not turned over by the cathepsins. Data represent the mean ± S.D. of a triplicate experiment. Amino acid sequence is indicated using the single-letter code. NA denotes no protease activity detected.

Substrate and kinetic constant	k_{cat}	K_m	k_{cat}/K_m
	sec^{-1}	μM	M^{-1} *sec*$^{-1}$
Z-FR-AMC			
Cathepsin K	2.1 ± 0.4	48.5 ± 1.9	4.4×10^4
Cathepsin L	7.5 ± 0.1	2.2 ± 0.2	3.4×10^6
Cathespin B	7.2 ± 0.1	38.1 ± 2.4	1.9×10^5
Ac-FR-ACC			
Cathepsin K	1.9 ± 0.1	37.4 ± 10.8	5.2×10^4
Cathepsin L	6.2 ± 0.2	11.4 ± 0.6	5.4×10^5
Cathespin B	0.3 ± 0.1	74.3 ± 10.4	4.3×10^3
Ac-LR-ACC			
Cathepsin K	13.5 ± 0.3	39.5 ± 1.6	3.4×10^5
Cathepsin L	10.9 ± 0.2	57.4 ± 0.8	1.9×10^5
Cathespin B	1.5 ± 0.1	634.5 ± 105.8	2.4×10^3
Ac-RLR-ACC			
Cathepsin K	9.2 ± 0.4	99.6 ± 0.6	9.3×10^4
Cathepsin L	11.9 ± 1.0	21.5 ± 3.2	5.6×10^5
Cathespin B	1.4 ± 0.1	347.1 ± 39.3	4.1×10^3
Ac-HRLR-ACC			
Cathepsin K	6.8 ± 0.5	57.4 ± 10.7	1.2×10^5
Cathepsin L	13.1 ± 1.8	7.6 ± 1.5	1.7×10^6
Cathespin B	5.1 ± 0.5	351.3 ± 45.9	1.5×10^4
Ac-HRYR-ACC			
Cathepsin K	0.5 ± 0.2	738.7 ± 431.2	7.6×10^2
Cathepsin L	20.8 ± 0.1	11.2 ± 0.4	1.9×10^6
Cathespin B	8.1 ± 0.1	475.0 ± 41.2	7.0×10^3
Ac-HGPR-ACC			
Cathepsin K	6.9 ± 0.5	197.9 ± 10.5	3.5×10^4
Cathespin S	NA	NA	NA
Cathepsin L	NA	NA	NA
Cathepsin V	NA	NA	NA
Cathespin B	NA	NA	NA
Ac-HGPN-ACC			
Cathepsin K	1.2 ± 0.2	215.0 ± 39.0	5.6×10^3
Cathespin S	NA	NA	NA
Cathepsin L	NA	NA	NA
Cathepsin V	NA	NA	NA
Cathespin B	NA	NA	NA

this group due to their unique specificities. Cathepsins B and S, cathepsin B-like 2, and papain formed the second group that showed broad specificity with a relative preference of aliphatic amino acids. The unique preference for arginine and histidine of bromelain is also well displayed. The P3 position has formerly not drawn much interest, partly due to the lack of a method to exhaustively compare all possible amino acids. However, the complete diverse PS-SCL and the cluster analysis demonstrated that there is subtle yet interesting diversity among P3 specificities such as preference for basic amino acids, aliphatic amino acids, or proline, although it is not as apparent as P2 specificity (Fig. 3*c*, *panel P3*). The cluster analysis of the P4 specificity shows there is not much diversity at this position (Fig. 3*c*, *panel P4*).

Kinetics Assays of Individual ACC-based Substrates—To provide a more quantitative assessment of the specificity information obtained by the library assay, a series of peptide-ACC substrates of varying length and sequence were synthesized for kinetic studies with cathepsins K, L, and B (Table 2). With the commercially available substrate Z-FR-AMC, cathepsin L showed the strongest activity among these cathepsins due to its low K_m. With Ac-R-ACC, none of the proteases show any significant activity, which confirms that they require extended substrate-enzyme interactions (data not shown). Comparing Ac-FR-ACC and Ac-LR-ACC, cathepsin K displayed much stronger activity with Ac-LR-ACC, whereas cathepsin L displayed stronger activity with Ac-FR-ACC. These results are consistent with the complete diverse PS-SCL assay results (Fig. 2, *a* and *c*). With the tetrapeptide substrates Ac-HRLR-ACC and Ac-HRYR-ACC, cathepsin K again displayed a preference for leucine, whereas it showed negligible activity with Ac-HRYR-ACC, clearly due to the difference of the P2 amino acid residue. By contrast, cathepsin L displayed comparable activity with both substrates, a result also predicted by the complete diverse PS-SCL assays. Cathepsin L showed increased activity with each progressively longer substrate, suggesting that longer substrates ensure stronger interactions between the protease and its substrates. Overall, the more quantitative results from the kinetic study using individual ACC-based substrates were in good agreement with the more qualitative substrate specificity data from the library assays.

Preparation of Specific Substrates and a Selective Inhibitor for Cathepsin K—Combining the unique specificities of cathepsin K determined by the library assay (proline and glycine at the P2 and P3 positions, respectively), two individual substrates Ac-HGPR-ACC and Ac-HGPN-ACC were synthesized and tested for their selectivity for cathepsin K. They were assayed against cathepsins K, S, L, V, and B, and both proved to be selective for cathepsin K (Table 2). Further confirming the library assay result, cathepsin K showed weaker activity with Ac-HGPN-ACC (with asparagine, one of the least favored amino acids at the P1 position) than with Ac-HGPR-ACC. Although P1-S1 is not the main specificity interaction for the papain-like cysteine proteases, it clearly plays an important role in affinity and efficient catalysis of substrates, suggesting why these proteases share an almost identical P1 substrate specificity.

An irreversible inhibitor Ac-HGPR-AOMK was then prepared based on the same glycine-proline specificity of cathepsin K to test its selectivity for the protease (Fig. 4*a*). Although cathepsin B was also inhibited by this inhibitor contrary to the library assay result, inhibitor kinetic analysis and competition labeling experiments showed that inhibitor selectivity was achieved for cathepsin K over cathepsins L and V (Table 3, Fig. 4*b*). It appears that the substrate profiling results can be used to design selective substrates successfully but not necessarily to design entirely selective inhibitors, particularly for cathepsin B.

DISCUSSION

The cysteine cathepsins have been implicated in a variety of diseases and highly specialized cellular events (8, 34–36). However, the exact physiological or pathological roles of each cathepsin remain largely unclear due to their highly conserved structure and specificity along with their diverse yet often overlapping tissue and cell distributions. The complete diverse PS-SCL assay confirmed their broad and overall similar substrate specificity. However, despite their resemblance in specificity, the substrate library profiles also highlighted distinguishing preferences, particularly at the P2 and P3 positions.

Cathepsin K has been suggested to play a key role in osteoclast-mediated osteoporosis, which results in destruction of bone matrix through type I collagen degradation (37). For this reason, the protease has been considered an important drug target. To develop selective substrates or inhibitors for cathepsin K, it is important to understand the specificity of the closely related human proteases including other cathepsins as well as cathepsin K itself. Cathepsins S, L, and K have been described to be difficult to distinguish, particularly cathepsins S and K, which show high P2 specificity similarity (38). The complete diverse PS-SCL, however, successfully identified the differences between these closely related proteases by exhaustively comparing 20 amino acids at the P1 to P4 positions. Particularly, the library assay showed that cathepsin K has a unique preference for proline and glycine at the P2 and P3 positions, respectively. It is interesting that type I collagen, a well known substrate of cathepsin K, contains multiple glycine-proline sequences.

FIGURE 4. **Inhibitory activity and selectivity of Ac-HGPR-AOMK against human cathepsins L, B, and K.** Competitive labeling experiment using 125I-DCG-04, an activity-dependent affinity label for papain-like cysteine proteases, was performed to compare the inhibitory activity and selectivity of the inhibitor that was designed based on the library assay. Human cathepsins L, B, and K (1 μg of protein in 100 μl of 50 mM Tris (pH 5.5), 1 mM dithiothreitol, 5 mM $MgCl_2$, 250 mM sucrose) were incubated with Ac-HGPR-AOMK (0, 0.1, 1, 10, 100 μM) for 1 h at room temperature. Samples were then labeled by the addition of ^{125}I-DCG-04 followed by a further incubation at room temperature for 1 h. Samples were quenched by the addition of 45 μl of sample buffer followed by boiling for 5 min. Samples were analyzed by SDS-PAGE followed by autoradiography. *a*, structure of the inhibitor, Ac-HGPR-AOMK. *b*, comparison of inhibition by Ac-HGPR-AOMK against cathepsins L, B, and K. For quantitative comparison, relative inhibition was measured using densitometry. The intensity of bands and percentage of labeling (by ^{125}I-DCG-04) inversely reflect the binding efficacy of Ac-HGPR-AOMK to the proteases.

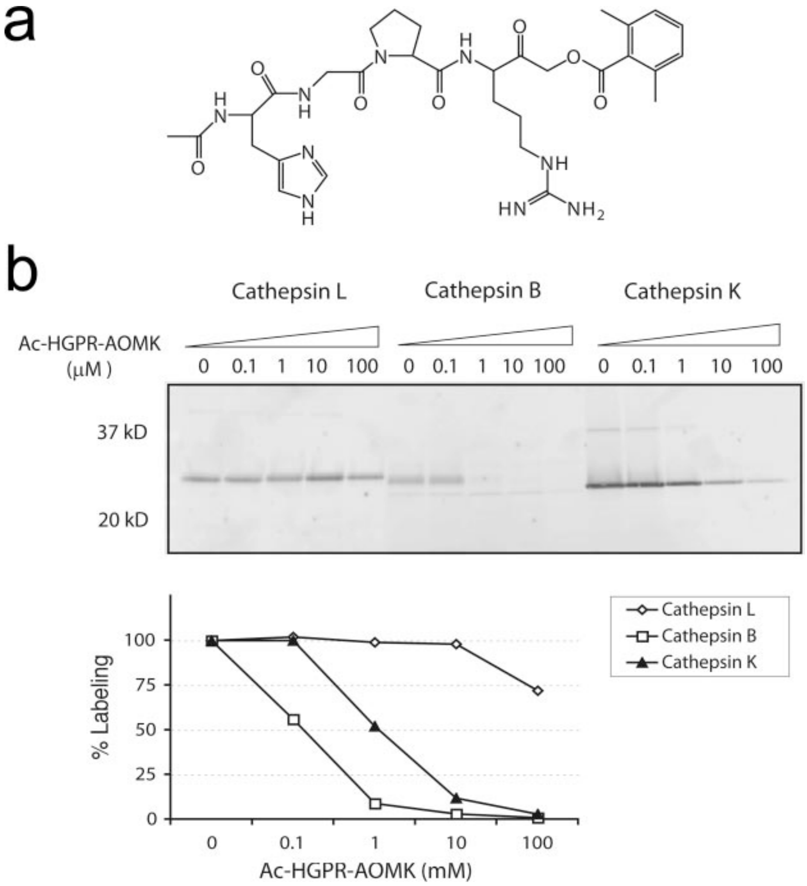


TABLE 3

Kinetic analysis of an irreversible inhibitor, Ac-HGPR-AOMK

The selectivity of the inhibitor was tested for human cathepsins K, B, L, and V. Data represent the mean ± S.D. of a triplicate experiment. NI denotes no inhibitory activity detected.

	k_{inact}	K_i	k_{inact}/K_i
	10^{-3} sec^{-1}	μM	M^{-1} sec^{-1}
Cathepsin K	6.5 ± 0.3	1.3 ± 0.0	5.1×10^3
Cathepsin B	12.9 ± 0.2	1.1 ± 0.3	1.2×10^4
Cathepsin L	NI	NI	NI
Cathepsin V	NI	NI	NI

Based on this information, selective tetrapeptide-ACC substrates and an irreversible inhibitor of cathepsin K with glycine-proline at the P3-P2 positions were developed, corroborating the usefulness of the ACC-based PS-SCL and the information obtained using it.

The specificity-based cluster analysis of papain-like cysteine proteases predictably reflected primary sequence identity. However, distinctive specificity differences were displayed among these proteases, providing a more detailed functional comparison at each subsite. This information will particularly benefit structural studies, which often rely heavily on primary sequence-based protein modeling. Most importantly, the information is expected to aid the exploration of specificity determinants in substrate binding pockets of proteases by helping to find closely related proteases based on their specificity. As more specificity data are accumulated, more information on protease function and structure will be generated, eventually providing the basis for predicting the substrate specificity of a protease from its primary sequence.

With the emergence of newly identified proteases, there is an urgent need for efficient methods to characterize their substrate specificity. Until recently, specificity studies of proteases had to rely on kinetic-based analysis that is limited by the number of available substrates. Traditionally, substrates for protease assays have been developed based on the physiological substrate cleavage sequences or consensus substrate sequences of homologous proteases or through screening of commercially available substrates. The complete diverse PS-SCL can rapidly establish the specificity of a protease without any prior knowledge of its physiological role, structure, or homology to other proteases. Its use is not limited by P1 specificity, which is an advantage over our previous PS-SCLs (4).

The complete diverse PS-SCL assay utilizes a simple rapid microtiter plate format to provide exhaustive and reproducible specificity information of proteases at the P1, P2, P3, and P4 positions. In most cases, submicrogram level of an active papain-like protease was sufficient to determine its specificity within minutes. The ACC substrate library assay is also sensitive and continuous, which gives it the potential to be easily adapted to high throughput screening or automated assay systems.

By using a tetrapeptide substrate library completely diversified at the P1–P4 positions to determine the substrate specificity of various papain-like cysteine proteases, we identified functionally unique specificities. The specificity information generated by the complete diverse PS-SCL enabled: (i) quick development of efficient and specific assays for the closely related proteases; (ii) design of selective inhibitors and labeling reagents; (iii) facilitation of the identification of specificity determinants through cluster analysis; and (iv) exploration of the protein data base for physiological substrates. Although the PS-SCL described is limited to P1–P4 subsites, it can readily be used in conjunction with other methods to examine the extended specificity pockets. Clearly, the complete diverse PS-SCL is expected to be useful to the development of selective reagents to control the activity of proteases and to help determine the physiological consequences of proteolytic activity.

Acknowledgments—We thank the members of the Craik laboratory for many useful discussions.

REFERENCES

1. Rawlings, N. D., Tolle, D. P., and Barrett, A. J. (2004) *Nucleic Acids Res.* **32,** D160–D164
2. Marnett, A. B., and Craik, C. S. (2005) *Trends Biotechnol.* **23,** 59–64
3. Thornberry, N. A., Rano, T. A., Peterson, E. P., Rasper, D. M., Timkey, T., Garcia-Calvo, M., Houtzager, V. M., Nordstrom, P. A., Roy, S., Vaillancourt, J. P., Chapman, K. T., and Nicholson, D. W. (1997) *J. Biol. Chem.* **272,** 17907–17911
4. Harris, J. L., Backes, B. J., Leonetti, F., Mahrus, S., Ellman, J. A., and Craik, C. S. (2000) *Proc. Natl. Acad. Sci. U. S. A.* **97,** 7754–7759
5. Takeuchi, T., Harris, J. L., Huang, W., Yan, K. W., Coughlin, S. R., and Craik, C. S. (2000) *J. Biol. Chem.* **275,** 26333–26342
6. Harris, J. L., Niles, A., Burdick, K., Maffitt, M., Backes, B. J., Ellman, J. A., Kuntz, I., Haak-Frendscho, M., and Craik, C. S. (2001) *J. Biol. Chem.* **276,** 34941–34947
7. Salter, J. P., Choe, Y., Albrecht, H., Franklin, C., Lim, K. C., Craik, C. S., and McKerrow, J. H. (2002) *J. Biol. Chem.* **277,** 24618–24624
8. Lecaille, F., Kaleta, J., and Brömme, D. (2002) *Chem. Rev.* **102,** 4459–4488
9. Brömme, D., and Kaleta, J. (2002) *Curr. Pharm. Des.* **8,** 1639–1658
10. Sajid, M., and McKerrow, J. H. (2002) *Mol. Biochem. Parasitol.* **120,** 1–21
11. Rosenthal, P. J. (2004) *Int. J. Parasitol.* **34,** 1489–1499
12. Brömme, D., Okamoto, K., Wang, B. B., and Biroc, S. (1996) *J. Biol. Chem.* **271,** 2126–2132
13. Berger, A., and Schechter, I. (1970) *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **257,** 249–264
14. Turk, D., Guncar, G., Podobnik, M., and Turk, B. (1998) *Biol. Chem.* **379,** 137–147
15. Brömme, D., and McGrath, M. E. (1996) *Protein Sci.* **5,** 789–791
16. Smith, S. M., and Gottesman, M. M. (1989) *J. Biol. Chem.* **264,** 20487–20495
17. Brömme, D., Li, Z., Barnes, M., and Mehler, E. (1999) *Biochemistry* **38,** 2377–2385
18. Caffrey, C. R., Hansell, E., Lucas, K. D., Brinen, L. S., Alvarez Hernandez, A., Cheng, J., Gwaltney, S. L., II, Roush, W. R., Stierhof, Y. D., Bogyo, M., Steverding, D., and McKerrow, J. H. (2001) *Mol. Biochem. Parasitol.* **118,** 61–73
19. Eakin, A. E., Mills, A. A., Harth, G., McKerrow, J. H., and Craik, C. S. (1992) *J. Biol. Chem.* **267,** 7411–7420
20. Backes, B. J., Harris, J. L., Leonetti, F., Craik, C. S., and Ellman, J. A. (2000) *Nat. Biotechnol.* **18,** 187–193
21. Maly, D. J., Leonetti, F., Backes, B. J., Dauber, D. S., Harris, J. L., Craik, C. S., and Ellman, J. A. (2002) *J. Org. Chem.* **67,** 910–915
22. Bunin, B. A. (1998) *The Combinatorial Index*, p. 6, Academic Press, San Diego, CA
23. Ostresh, J. M., Winkle, J. H., Hamashin, V. T., and Houghten, R. A. (1994) *Biopolymers* **34,** 1681–1689
24. Wagner, B. M., Smith, R. A., Coles, P. J., Copp, L. J., Ernest, M. J., and Krantz, A. (1994) *J. Med. Chem.* **37,** 1833–1840
25. Brömme, D., Smith, R. A., Coles, P. J., Kirschke, H., Storer, A. C., and Krantz, A. (1994) *Biol. Chem. Hoppe-Seyler* **375,** 343–347
26. Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998) *Proc. Natl. Acad. Sci. U. S. A.* **95,** 14863–14868
27. Greenbaum, D., Medzihradszky, K. F., Burlingame, A., and Bogyo, M. (2000) *Chem. Biol.* **7,** 569–581
28. McGrath, M. E., Palmer, J. T., Brömme, D., and Somoza, J. R. (1998) *Protein Sci.* **7,** 1294–1302
29. Puzer, L., Cotrin, S. S., Alves, M. F., Egborge, T., Araujo, M. S., Juliano, M. A., Juliano, L., Brömme, D., and Carmona, A. K. (2004) *Arch. Biochem. Biophys.* **430,** 274–283
30. Cotrin, S. S., Puzer, L., de Souza Judice, W. A., Juliano, L., Carmona, A. K., and Juliano, M. A. (2004) *Anal. Biochem.* **335,** 244–252
31. Xia, L., Kilb, J., Wex, H., Li, Z., Lipyansky, A., Breuil, V., Stein, L., Palmer, J. T., Dempster, D. W., and Brömme, D. (1999) *Biol. Chem.* **380,** 679–687
32. Kafienah, W., Brömme, D., Buttle, D. J., Croucher, L. J., and Hollander, A. P. (1998) *Biochem. J.* **331**, 727–732
33. Villadangos, J. A., Riese, R. J., Peters, C., Chapman, H. A., and Ploegh, H. L. (1997) *J. Exp. Med.* **186,** 549–560
34. Turk, B., Turk, D., and Turk, V. (2000) *Biochim. Biophys. Acta* **1477,** 98–111
35. Turk, V., Turk, B., and Turk, D. (2001) *EMBO J.* **20,** 4629–4633
36. Nakagawa, T., Roth, W., Wong, P., Nelson, A., Farr, A., Deussing, J., Villadangos, J. A., Ploegh, H., Peters, C., and Rudensky, A. Y. (1998) *Science* **280,** 450–453
37. Yasuda, Y., Kaleta, J., and Brömme, D. (2005) *Adv. Drug Delivery Rev.* **57,** 973–993
38. Barrett, A. J., Rawlings, N. D., and Woessner, J. F. (2004) *Handbook of Proteolytic Enzymes*, Second Ed., p. 1105, Academic Press, London