Substrate specificity of schistosome versus human legumain determined by P1–P3 peptide libraries

Mary A. Mathieu a,⁎, Matthew Bogyo b, Conor R. Caffrey a, Youngchool Choe c, Jewok Lee d, Harold Chapman d, Mohammed Sajid a, Charles S. Craik c, James H. McKerrow a

a Department of Pathology, UCSF, San Francisco, CA 94143, USA
b Department of Biochemistry and Biophysics, UCSF, San Francisco, CA 94143, USA
c Department of Pharmaceutical Chemistry, UCSF, San Francisco, CA 94143, USA
d Department of Medicine, UCSF, San Francisco, CA 94143, USA

Received 8 November 2001; accepted in revised form 30 January 2002

Abstract

Asparaginyl endopeptidases, or ‘legumains’ have been identified and characterized in plants, the blood fluke parasite Schistosoma, and mammals. The legumains are a novel family of cysteine proteases and display restricted specificity for peptide hydrolysis on the carboxyl side of asparagine residues. Two forms of recombinant asparaginyl endopeptidase from Schistosoma mansoni (C197 Sm32 and N197C Sm32), expressed in Pichia pastoris, have been analyzed for substrate specificity using a positional-scanning synthetic combinatorial library (PS-SCL). We first screened Sm32 using a P1-diverse library. This library demonstrated the absolute specificity of Sm32 for asparagine at P1. To determine the P2/C1/P3 preferences of Sm32, we constructed a library with asparagine fixed at P1, and the P2/C1/P3 positions randomized. The library was screened using the two forms of Sm32, human asparaginyl endopeptidase, and to confirm its diversity, cruzain from Trypanosoma cruzi. The schistosome legumain showed a preference for P3: Thr > Ala > Val > Ile, and P2: Ala > Thr > Val > Asn, with an overall broader specificity at P3 than at P2. Both human and schistosome legumain can accommodate Thr and Ala at P2 and P3. However, optimal substrate sequences differ, with Sm32 preferring Thr-Ala-Asn, and human legumain preferring Pro-Thr-Asn. Predictions of substrate specificity from the library screen were confirmed using single peptide substrates for kinetic assays. © 2002 Published by Elsevier Science B.V.

Keywords: Schistosoma mansoni; Asparaginyl endopeptidase; Legumain; Specificity; Positional scanning

1. Introduction

Asparaginyl endopeptidases (EC 3.4.22.34), or ‘legumains’ belong to the C13 family of cysteine proteases. They have been assigned to clan CD, along with caspases, clostripains, gingipains, and GPI: protein transamidases based on sequences spanning the catalytic dyad in the motif His-Gly spacer-Ala-Cys [1,2]. The legumain family members are believed to have a protein fold closely related to the caspases [1,3]. Legumains were first characterized in the seeds of leguminous plants [1,4] and later discovered in the parasitic blood fluke Schistosoma mansoni [5] and in mammalian cells [6]. There is evidence that the plant enzymes process and degrade proenzymes in storage vacuoles [7,8]. The mammalian enzyme functions in bacterial antigen processing for MHC class II presenting cells [9] and in the inhibition of osteoclast formation and bone resorption [10]. S. mansoni asparaginyl endopeptidase (Sm32) was first identified as a diagnostic marker for schistosomiasis [11] and was proposed as a candidate in the protease pathway of host hemoglobin degradation by the parasite [12]. Monoclonal antibodies have localized

⁎ Corresponding author. Tel.: +1-415-514-3052; fax: +1-415-637-4338.
E-mail address: mmathie@itsa.ucsf.edu (M.A. Mathieu).
Sm32 to the gut epithelium of the worm where it may
\textit{trans}-process and activate proteases that are involved in
hemoglobin digestion [13]. Because several of the gut-
associated proteases of schistosome parasites contain an
asparagine at or near cleavage site for removal of the
prodomain, schistosome legumain is a prime candidate
for an upstream activating enzyme in a protease
cascade. Mapping substrate specificity, manifested by
which amino acids are preferred around the P1 aspar-
agene, would help to validate the proposed role of
legumain in schistosome digestion of host proteins.

Although its role in parasite nutrition may be indirect,
it is an attractive target for inhibition of ‘downstream’
metabolic processes catalyzed by other proteases.
Broad-spectrum inhibitors against cysteine proteases
likely to be activated by legumain have in fact been
used to treat \textit{S. mansoni}-infected mice, resulting in
reduced worm burden and parasite egg production [14].

Two distinct cDNA’s with 97.2\% identity have been
determined that encode for the asparaginyl endopeptidase
from \textit{S. mansoni}. The first cDNA (N197) encodes an
asparagine residue at position 197 (\textit{S. mansoni} number-
ing), where a crucial active site cysteine should reside
[15]. Caffrey et al. have identified a second distinct gene
for Sm32 containing the requisite C197 residue (C197
Sm32). While autocatalytic processing to fully acti-
vate Sm32 containing the requisite C197 residue (C197
Sm32) occurs at acid pH, protease activ-
it is an attracti-

The unique specificity of Sm32 for asparagine at P1
may be advantageous for the elucidation of its precise
physiological role. Little is known about the substrate
specificity of the legumain family at positions other than
P1. Knowledge of specific side chains optimal for
binding would help to define possible biological func-
tions and serve as a basis for inhibitor design. Posi-
tional-scanning synthetic combinatorial libraries (PS-
SCL) have been successfully used to characterize
protease specificity [17]. Fluorogenic peptide substrates
can be used in a rapid and sensitive assay for selecting
optimal peptide sequences preferred by a given enzyme.

In this report, we have recombinantly expressed
N197C and C197 Sm32 using the \textit{Pichia pastoris} yeast
expression system and used a combinatorial library
approach to determine the substrate specificity of
Sm32 at P2 and P3. After confirming a strict preference
for asparaginyl at P1 using a P1 diverse library, we
synthesized a tripeptide positional library where aspar-
agine is fixed in P1 and the P2 and P3 positions are
randomized, respectively. The schistosome enzymes
were profiled for P2 and P3 specificity compared to
the corresponding human legumain. Kinetic analysis
using single substrates was carried out to validate the
results from the library screen.

### 2. Materials and methods

#### 2.1. Materials

Rink Amide resin (0.80 meq g$^{-1}$), PyBOP and Fmoc
amino acids were purchased from Advanced Chemtech
(Louisville, KY) \{Fmoc-Ala-OH, Fmoc-Arg(Pbf)-OH,
Fmoc-Asn(Trt)-OH, Fmoc-Asp(O-t-Bu)-OH, Fmoc-
Glu(O-t-Bu)-OH, Fmoc-Gln(Trt)-OH, Fmoc-Gly-OH,
Fmoc-His(Trt)-OH, Fmoc-Ile-OH, Fmoc-Nle-OH,
Fmoc-Phe-OH, Fmoc-Pro-OH, Fmoc-Ser(O-t-Bu)-
OH, Fmoc-Thr(O-t-Bu)-OH, Fmoc-Trp(Boc)-OH,
Fmoc-Tyr(O-t-Bu)-OH, Fmoc-Val-OH\}. Anhydrous
\textit{N,N}-dimethylformamide was purchased from EM
Science (Hawthorne, NY) and \textit{O-}(7-azabenzotriazole-
1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate
(HATU) from PerSeptive Biosystems (Foster City,
CA). Diisopropylcarbodiimide (DIC), 1-hydroxybenzo-
triazole (HOBr), trifluoroacetic acid (TFA), triisopro-
pylsilane (TIS), \textit{N,N}-diethyl-\textit{p}-phenylenediamine
(DIPEA) and collidine were purchased from Aldrich.
The FlexChem Solid Phase Chemistry System used for
library synthesis was purchased from Robbins Scientific
(Sunnyvale, CA). Recombinant \textit{S. mansoni} asparaginyl-
endopeptidase (Sm32) was prepared as described pre-
viously [16]. Sm32 was activated at 37 °C for 3 h in 100
mM sodium acetate at pH 4.5 containing 1 mM DTT.

For comparison to the parasite legumain, human
asparaginyl-endopeptidase (hAE) was recombinantly
expressed in insect cells (Lee and Chapman, unpublished
results). Cruzain, the major cysteine protease of \textit{Trypa-
nosoma cruzi}, was prepared as described previously and
used to confirm diversity of the library [18].

#### 2.2. Synthesis of Fmoc-Asp (OtBu)-OH

H-Asp (OtBu)-OH (3.0 g, 7.3 mmol) and HATU (5.4
g, 14.3 mmol), were dissolved in 15 ml DMF. Collidine
was added (1.9 ml, 14.3 mmol) and the mixture was
stirred for 10 min. 4-methylcoumarin-7-amide (MCA,
1.3 g, 7.3 mmol) was added as a solid and the mixture
was stirred at room temperature overnight. DMF was
removed by rotoevaporation. Water was added and the
aqueous phase was extracted three times with ethyl
acetate, washed twice each with sodium carbonate and
sodium chloride, and extracted with 0.1 N HCl. The
crude product was dried over magnesium sulfate. After
rotoevaporation, the product was purified by silica gel
chromatography with a solvent of EtOH:hexane, 1:3.
The eluent was dried to give 2.8 g (4.9 mmol, 67\% yield)
of cream-colored product. The identity of the product
was confirmed by mass spectrometry. The major
product was observed at 513.26 amu, the sodium salt
of the Asp-MCA compound of mass 512.16. Side chain
deprotection was accomplished by adding 5.0 ml of a
TFA ( trifluoroacetic acid)/H$_2$O/triisopropylsilane mix-
ture (95:2.5:2.5) to the product and allowing it to stand for 1 h. Toluene was added (15 ml) and the mixture was concentrated by rotoevaporation. Cold ether was added to precipitate the product. The ether was poured off and the solid precipitate collected.

2.3. Synthesis of P1-substituted MCA-resin

Rink amide resin (70 μmol) was swelled in 0.5 ml DMF and deprotected with 20% piperidine in DMF with agitation for 15 min. The resin was washed with DMF (5 × with 0.5 ml) and combined with Fmoc-Asp-MCA (140 μmol), PyBOP (140 μmol), HOBT (140 μmol), and DIPEA (280 μmol). This reaction mixture was agitated overnight, filtered, washed with DMF, and the coupling reaction procedure repeated a second time. The product was washed (DMF, 3 × with 0.5 ml; MeOH, 3 × with 0.5 ml). The identity of the product was confirmed by mass spectrometry. The efficiency of first residue attachment was estimated at 4.2 mmol g⁻¹ from Fmoc absorbance at 290 nm.

2.4. Synthesis of P1-Asn-MCA library

P1-Asn-MCA-resin was added to 38 wells of a Flexchem 96 well reaction apparatus (≈20 mg, 20 μmol). The resin was swelled with 0.5 ml DMF, filtered, and deprotected with 20% piperidine in DMF for 20 min. The wells were then filtered and washed (DMF, 3 × with 0.5 ml, MeOH, 2 × with 0.5 ml) in preparation for the P2 coupling reaction. Each of nineteen amino acids (cysteine was omitted and norleucine was substituted for methionine) was added to wells 1–19 to introduce a fixed P2 position (5 eq each of Fmoc-amino acid, DIC, and HOBt in DMF). In wells 20–34, an isokinetic mixture of 19 amino acids was added to randomize the P2 position in these reactions.

Coupling was achieved using 5 eq of HOBt and DIC per well following a previously described protocol [19]. The reaction block was agitated for 4 h, filtered, and washed with DMF and MeOH as above. Deprotection of each reaction was repeated using 20% piperidine in DMF, followed by washing with DMF and MeOH as above. The randomization of P3 in wells 1–19 and fixing of P3 in wells 20–38 was achieved by reversing the protocol described above. The reaction block was agitated for 4 h and the wells were washed with DMF, MeOH, and methylene chloride (2 × with 0.5 ml). The resulting tripeptides were capped using a solution (150 μl well⁻¹) of acetic anhydride (400 μl), pyridine (315 μl) and DMF (6.0 ml). The block was agitated for 4 h and then washed with DMF, MeOH and CH₂Cl₂ as above. To cleave each acetylated tripeptide from the resin, a solution (95:2.5:2.5 TFA/TIS/H₂O, 0.5 ml well⁻¹) was added and the block was agitated for 2 h. The reaction block was then transferred to a deep-well collection chamber and the tripeptides were removed by filtration under reduced pressure with the addition of the remaining cleavage solution. The contents from each well were transferred to 15 ml falcon tubes and precipitated with ether. Ether was poured off and the remaining contents were resuspended in 50% acetonitrile and lyophilized. Individual substrates were dissolved in DMSO to provide 10 mM stocks for enzyme assays.

2.5. Kinetic assay of library

Enzyme assays were carried out in opaque 96-well microtiter plates (Corning Inc., Corning, NY) using Labsystems Fluoroskan II and Deltasoft 3 software. Each member of the substrate library was assayed in triplicate at a final concentration of 50 μM well⁻¹ against activated enzyme (20 μl well⁻¹) in 100 mM sodium acetate at pH 4.5 containing 1 mM DTT and 1 mM EDTA. Hydrolysis of the substrate was monitored fluorimetrically with excitation at 355 nm and emission at 460 nm based on the fluorescence properties of free AMC.

2.6. P1-diverse library synthesis and assay

Individual P-1 substituted Fmoc amino acid resin was prepared as described previously [19]. 7-Amino-4-carboxymethylcoumarin (AMC) was used as a fluorescent leaving group in place of AMC. Randomized P3, P5, and P4 positions were incorporated by addition of the isokinetic mixture of 19 amino acids as described above. Substrates from the P1-diverse library were added to 19 wells of a 96-well Microfluor plate (Dynex Technologies, Chantilly, VA) for a final concentration of approximately 0.1 μM. Hydrolysis reactions were initiated by addition of enzyme and monitored on a Perkin–Elmer LS50B luminescence spectrometer with excitation at 380 nm and emission at 460 nm.

2.7. Single substrate kinetic assays

Individual tripeptide substrates were obtained as custom syntheses from Enzyme Systems Products, Livermore, CA. The following custom substrates were used for kinetic analysis: Z-Ala-Ala-Asn-MCA, Ac-Thr-Ala-Asn-MCA, Ac-Pro-Thr-Asn-MCA and Ac-Phe-Tyr-Asn-MCA. Enzyme activity was monitored at 25 °C in an assay buffer of 50 mM sodium acetate, pH 5.5, containing 2 mM DTT and 2 mM EDTA. The final concentration of substrate ranged from 0.006 to 0.4 mM. The DMSO concentration in the assay was less than 5%. Pre-activated enzyme solution was added to each well (100 μl total per well). Release of AMC was monitored fluorimetrically with an excitation wavelength of 380 nm and emission wavelength at 460 nm. GraphPad Prism software (San Diego, CA) was used to
fit the data to a non-linear regression curve. $V_{\text{max}}$ and $K_m$ were determined using a Michaelis–Menten plot of initial enzyme velocity versus substrate concentration.

### 3. Results

#### 3.1. Design and construction of the peptide library

As a member of the asparaginyl endopeptidase group, Sm32 was predicted to have specificity for cleavage on the carboxyl side of asparagine residues [20,21]. We confirmed the absolute specificity for cleavage after asparagine using a P1 divergent library (Fig. 1). The fluorogenic tri-peptide substrate library with the general structure Ac-X-X-Asn-amidomethylcoumarin (Fig. 2) was therefore generated to determine the substrate specificity of Sm32. Randomization of P2 and P3 positions was achieved using a method that ensures equal representation of all amino acids of an isokinetic mixture [22]. As some amino acids typically provide low coupling yields, each coupling was repeated to increase substitution levels.

#### 3.2. Screening of the peptide library

Cleavage by a protease of Ac-X-X-Asn-amidomethylcoumarin at the scissile bond liberates the fluorescent leaving group, AMC, for sensitive and convenient detection of proteolytic activity by fluorimetry. Successful activation of schistosome legumain was first confirmed by its ability to cleave the peptidyl substrate (Z-Ala-Ala-Asn-MCA), and by inhibition of enzyme activity using iodoacetic acid [16].

The schistosome enzyme showed a preference for P3: Thr > Ala > Val > Ile, and P2: Ala > Thr > Val > Asn, with an overall broader specificity at P3 than at P2 (Fig. 3). Neither schistosome nor human legumain accommodated His or Tyr in P2 or P3, and activity is low against peptides containing Asp, Glu, Gln, Phe or Trp.

![Fig. 1. Strict specificity for Asn in P1 of *S. mansoni* asparaginyl endopeptidase is confirmed by screening a P1 variable library. The library consists of 19 wells, each containing a compound with a different amino acid represented at P1, but randomized at P2–P4 with isokinetic mixtures of amino acids. The $x$-axis provides the special address of the amino acid as represented by the one-letter code (i.e. A = alanine). Enzyme was activated for 2.5 h at pH 4.5 in 100 mM sodium acetate buffer containing 10 mM DTT. Enzyme solution (100 µg) was added to each of the ten wells. The $y$-axis represents the rate of AMC production expressed as a percentage of the maximum rate observed in each experiment. Production of AMC was monitored continuously on a Perkin–Elmer LS50B with excitation at 380 nm and emission at 460 nm.](image1)

![Tripeptide PS-SCL](image2)

![Fig. 2. Tri-peptide-AMC combinatorial positional-scanning library. The library consists of two sublibraries of 361 compounds each (10 wells of 19 compounds). Each library is acylated at the amino terminus, contains asparagine at P1, and is tagged with an AMC fluorophore. $X$ represents a spatially addressed amino acid and $Y$ represents an isokinetic mixture of the 1–19 amino acids described previously.](image3)
(Leu > Lys, Arg, Gly > Phe, His, Ala). These results are consistent with previously reported data using peptide substrates [23], as well as structural data with peptide inhibitor co-crystals [24].

The library was also screened using hAE to compare its specificity with that of Sm32. Both human and schistosome enzymes accommodate threonine and alanine well in P2 and P3; however, there are some notable differences between their optimal substrate sequences. In particular, Thr-Ala-Asn is optimal for schistosome legumain, Pro-Thr-Asn, is optimal for human legumain.

### 3.3. Single substrate kinetic assays

To verify the results of the peptide library screen, single tripeptide-methylamidocoumarin (MCA) substrates predicted by PS-SCL to be good (Ala-Ala-Asn, Thr-Ala-Asn, Pro-Thr-Asn), or poor (Phe-Tyr-Asn), were assayed to compare schistosome and human legumains. The results of the kinetic analysis shown in Table 1 are consistent with the predictions made by the PS-SCL. Ac-Thr-Ala-Asn-MCA and Z-Ala-Ala-Asn-MCA, are the substrates most preferred by the schistosome enzyme ($K_m$ values, Table 1). The substrate Ac-

<table>
<thead>
<tr>
<th>Schistosome legumain</th>
<th>Human legumain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Z-Ala-Ala-Asn-MCA</td>
<td>$K_m$ (µM)</td>
</tr>
<tr>
<td>Ac-Thr-Ala-Asn-MCA</td>
<td>70 ± 13</td>
</tr>
<tr>
<td>Ac-Pro-Thr-Asn-MCA</td>
<td>286 ± 40</td>
</tr>
<tr>
<td>Ac-Phe-Tyr-Asn-MCA</td>
<td>n.a.</td>
</tr>
</tbody>
</table>
4. Discussion

Asparaginyl endopeptidases, or ‘legumains’ are a recently identified group of cysteine endopeptidases. They represent one of four protease families in clan CD. These families, C11, C13, C14 and C25 have been grouped together based on amino acid sequence and catalytic dyad motif. The protein fold for members of clan CD is predicted to be similar to the fold determined for the caspases in family C14. Compared with other cysteine proteases of broader specificity, members of clan CD have a strict requirement for a specific side chain of the P1 amino acid residue. Legumain is the only protease that is selective for asparagine at P1.

Plant, mammalian and S. mansoni enzymes have similar inhibition profiles, and similar biological roles in post-translational modification of other proteins or enzymes [4,8,20]. As a first step in the determination of substrate specificity for the human and S. mansoni asparaginyl endopeptidases, we confirmed the P1-Asn specificity by screening a P1 diverse library (Fig. 1). We then constructed a positional scanning combinatorial library (PS-SCL) with asparagine fixed at P1 and the respective P2 and P3 positions randomized using 19 amino acids. Although the results of the library screen suggested a similar specificity at P2 and P3, a notable difference was the preference for Pro at P3 for the human enzyme. Differentiation of the parasite enzyme from that of the host at P3 supports the possibility of selective inhibition.

To further validate the results of the PS-SCL, and confirm the utility of the peptide library, the P2–P3 substrate preference for cysteine protease cruzain from T. cruzi was determined [19,23]. Indeed, a very different and consistent profile for cruzain was found, suggesting that any bias in the peptide library is minimal (Fig. 3). Further confirmation of the PS-SCL was demonstrated by the $V_{\text{max}}$ and $K_m$ data obtained using purified single substrates (Table 1). The kinetic data is consistent with the predictions of good or poor peptide substrate sequences from the library screen. Finally, the results were also in agreement with previously reported data for porcine legumain [15].

The general method of using a PS-SCL to determine substrate specificity is rapid and efficient, and may provide information for predicting or confirming biological substrates. Manoury et al. [9] reported that microbial tetanus toxin antigen was processed by a mammalian asparaginyl endopeptidase (AEP). The digestion products showed that cleavage occurred after asparagine (Ile-Asp-Asn, Pro-Asn-Asn, and Phe-Asn-Asn). One of these cleavage sites (Pro-Asn-Asn) is consistent with an optimal sequence predicted by the PS-SCL screen. The other two are acceptable but would not be predicted to be optimal. No data is reported as to which site might be cleaved initially.

S. mansoni asparaginyl endopeptidase trans-processes and activates recombinant S. mansoni cathepsin B (SmCB1) (Sajid, personal communication). Cleavage occurs at an asparagine residue (Asp-Trp-Asn $\downarrow$ Val-Ile-Pro) between the pro-domain and mature regions of SmCB1. This is an acceptable substrate sequence, but again, not an optimal one.

Therefore, in the case of schistosome legumain, substrate specificity may also be determined by residues on the ‘prime side’ (carboxy terminus) of asparagine [24], or by cleavage site accessibility with the tertiary structure of the cathepsin B proenzyme target. Phage display is currently being employed to map the prime side specificity of S. mansoni asparaginyl endopeptidase to test this hypothesis.

Acknowledgements

This work was supported by a TDRU grant (AI35707) and the Sandler Family Support Foundation.

References