Cercarial Elastase Is Encoded by a Functionally Conserved Gene Family across Multiple Species of Schistosomes

Received for publication, March 11, 2002, and in revised form, April 18, 2002. Published, JBC Papers in Press, May 1, 2002, DOI 10.1074/jbc.M202364200

Jason P. Salter, Youngchool Cho, Hugo Albrecht, Christopher Franklin, Kee-Chong Lim, Charles S. Craik, and James H. McKerrow

From the §University of California San Francisco Graduate Program in Biomedical Sciences and the Departments of §Pathology and §Pharmaceutical Chemistry, University of California, San Francisco, California 94143 and ¶Discovery Partners International AG, Discovery Biology Division, Gewerbestrasse 16, F. O. Box 544, CH-4123 Allschwil, Switzerland

Water borne cercariae of the trematode genus Schistosoma rapidly penetrate host skin. A single serine protease activity, cercarial elastase, is deposited in advance of the invading parasite by holocystosis of vesicles from ten large acetabular gland cells. Cercarial elastase activity is a composite of multiple isoforms. Genes coding for the isoforms can be divided into two classes by amino acid and promoter sequence homology. Two of the five genes identified in Schistosoma mansoni account for over 90% of the activity and protein released. The remaining genes produce little protein or are silent. Positional scanning synthetic combinatorial substrate libraries demonstrate that the two major isoforms have similar substrate specificities and are, therefore, isoenzymes. The closely related Schistosoma hematobium and the distantly related Schistosomatium doualli also contain multiple orthologous cercarial elastase genes suggesting that gene duplication may have occurred after speciation in Schistosoma evolution and that this duplication has been conserved.

Cercariae, the aquatic infective larval stage of schistosomes, are highly adapted to rapidly penetrate the skin of the host upon contact. Enzymatic hydrolysis of host proteins is required for successful entry into the host vascular system (1). Two gland systems, the precacetabular and postacetabular glands, release proenzymes and comprise the majority of the volume of the cercarial head. Each gland cell releases proenzymes at the leading edge of the invading parasite through long, microtubule-lined cell processes or "ducts" that exit at the anterior head (2). The postacetabular glands are also responsible for depositing mucin, providing an adhesive surface on the skin for the parasite to initially attach. Considering the diverse nature of macromolecular barriers the cercariae must breach during invasion, we previously investigated the possibility that multiple enzyme activities were required. However, only a single serine protease activity, cercarial elastase, was found to be present in acetabular gland secretions and required for invasion (3).

* This work was supported by a Merit Grant from the Department of Veterans Affairs and the Sandler Family Foundation. 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.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EMBL Data Bank with accession number(s) AFS10347, AFS10338, AFS10339, AFS10340, AFS10341, AFS10342, AFS10343, AFS10344.

** To whom correspondence should be addressed: Dept. of Pathology and Pharmaceutical Chemistry, University of California San Francisco, 513 Parnassus, P. O. Box 0511, San Francisco, CA 94143. Tel.: 415-476-6906; Fax: 415-476-415; E-mail: jmck@scality.ucsf.edu.

Cercarial elastase is a trypsin family serine protease named because of its ability to cleave insoluble elastin, a major component of the dermis of skin (4, 5). Its PI substrate specificity (1) is for large hydrophobic side chains, but in contrast to chymotrypsin (3) cercarial elastase is more active against macromolecular substrates than synthetic tetrapeptides.

We examined the complement of genes coding for cercarial elastase in Schistosoma mansoni and found a family of isoforms that can be divided into two classes by amino acid and promoter sequence homology. This family of genes is also conserved in another schistosome species Schistosoma hematobium and Schistosomatium doualli. The two most highly expressed S. mansoni isoforms comprise >90% of the released activity and are virtually identical in biochemical properties.

EXPERIMENTAL PROCEDURES

Collection of Cercarial Secretions—Approximately 3–5 × 10⁶ S. mansoni cercariae (Puerto Rican strain) were collected in 750 ml of distilled water from 200–300 infected Biomphalaria glabrata snails using a light induction method previously reported (6). Secretions were collected by placing cercariae in Petri dishes coated with limicolic acid (to simulate skin contact) and floated in a 37 °C water bath to produce a thermal gradient. After 2 h the conditioned water was filtered using Swinnex-47 Grade 541 filters (Millipore, Bedford, MA) to remove cercarial bodies and debris. The secretions were then lyophilized and stored at 4 °C. Each batch of collected material constitutes one "snail shed".

Protein Chromatography—Cercarial secretions from four snail sheds were pooled and resuspended in 6 ml of gel filtration buffer (200 mM sodium-acetate, pH 6.5) and centrifuged for 30 min at 10,000 rpm in an SA-34 rotor (4 °C). The supernatant was loaded onto an SR 16/100 column (Amersham Biosciences) packed with Sepharacryl 200 (gel filtration resin; Amersham Biosciences). The column was run at a rate of 25 ml/hr, and 4-ml fractions were collected overnight at 4 °C. Fractions were assayed using 10 ml of sample and 100 ml of assay buffer (100 mM glycine, pH 9.0, 100 μm succinyl-ala-ala-pro-phe-p-nitroanilide (AAPF-pNA)). Absorbance kinetics were determined using a UV-Max spectrophotometer and SoftMax Version 2.02 software (Molecular Devices, Sunnyvale, CA).

Fractions with high AAPF-pNA activity were pooled and buffer-exchanged into running buffer (20 mM MIES, pH 6.8) using PD-10 buffer exchange columns (Amersham Biosciences). The sample was then loaded onto an HR 5/5 Mono-Q anion exchange column at 1 ml/min. Elution of protein from the column was accomplished with a combination of discontinuous steps or gradients of elution buffer (20 mM MIES, pH 6.8, 1.0 M NaCl). There were three elution steps separated by linear gradients (fractions 0–200—125 mM NaCl, fractions 21–2525—900 mM NaCl, fractions 26–39300—1000 mM NaCl). Fractions were assayed for protease activity as described above.

1 The abbreviations used are: AAPF-pNA, succinyl-ala-ala-pro-phe-p-nitroanilide; MIES, (2-O-morpholinoethanesulfonic acid), RT, reverse transcription; ACC, 7-amino-4-carboxymethylcoumarin; SWPL, serine-trypotphan-proline-lyeucine; TWPL, threonine-trypotphan-proline-lyeucine; RWPL, arginine-trypotphan-proline-lyeucine; RRPL, arginine-arginine-proline-lyeucine.
Peaks of activity from the ion exchange column were analyzed by SDS-PAGE. NuPage-polyacrylamide 4–12% bis-Tris gradient gels (Invitrogen) were used according to the manufacturer’s instructions. Gels were stained with Coomassie Blue to visualize protein and determine purity.

Recombinant SmCE-1a and Antibody Production—An Escherichia coli expression construct, pET-21a-SmCE-1a-125M-6xHis, was assembled by inserting the active portion of SmCE-1a, amplified by PCR, into the restriction sites NotI (Munich Reagents, Munich, Germany) and NdeI (Novagen, Madison, WI) sites of the forward primer 5'-TGGCCTAGTGAATAGCCCTGGA-3' created a blunt end to match with the blunt NdeI site of the vector changing the first amino acid to a start codon (Il-28-Met). The reverse primer 5'-CAGATGGCTGAAATGGACAGCATACAAAAC-3' included an Ascl site (underlined) and removed the 3' AdoII crossing an open reading frame into the pET21a vector that added a 6x histidine tag to the C terminus of the protein. Purified recombinant His-tagged protein from E. coli was then used to produce rabbit antisera using standard commercial procedures (Covance, Richmond, CA).

Immunodot Analysis. Antigenic proteins were separated by SDS-PAGE, as described above and transferred onto polyvinyldene difluoride membranes using the Novex transfer system (Invitrogen). Polyclonal rabbit sera was used to probe the blot. Anti-rabbit horseradish peroxidase-conjugated secondary antibody was used for detection with standard ECL reagents (Amer sham, Buckinghamshire, UK). Precautions were taken to avoid cross-contamination of adjacent blots.

Promoter Identification by PCR-based Genomic DNA Walking—Genomic libraries were constructed with the overnight digestion of 10 μg of genomic DNA and 3 μg of pUC19 (Stratagene, La Jolla, CA) plasmid using individual restriction enzymes (ApaI, BamHI, BclI, EcoRI, KpnI, KpnII, NotI, PstI, SalI, XhoI, and XbaI, New England Biolabs, Beverly, MA). Matching digests of genomic DNA and plasmid DNA acting as a universal linker) were mixed and purified using a spin prep column (Qiagen, Valencia, CA). Purified samples were digested overnight at 14°C to link the plasmid to the genomic fragment. PCR products of the digested reactions were set up in volumes of 50 μl using 2 μl of library DNA as template and the primer sets (5'-GATTACGCAAGCTTGCGATGCG, 5'-TGGATTCTAGGGAATTCGACG) for ellipsarase or (5'-GATTACGCAAGCTTGCGATGCG, 5'-TGGATTCTAGGGAATTCGACG) for GAPDH). Thermocycling conditions were 40 s at 94°C, 1 min at 60°C, and 3 min at 72°C for 30 cycles. Second round PCR products were set up in volumes of 100 μl using 0.5 μl of first round PCR product as template and the primer set (5'-GATTACGCAAGCTTGCGATGCG, 5'-TGGATTCTAGGGAATTCGACG) for ellipsarase or (5'-GATTACGCAAGCTTGCGATGCG, 5'-TGGATTCTAGGGAATTCGACG) for GAPDH). Thermocycling conditions were as described above.

Isoform Identification by RT-PCR—A primer was designed to bridge the deletion region within the group 2 promoters (see Fig. 3A). This primer was used for group 2 transcripts and was in use in an RT-PCR to amplify the downstream expressed genes. Pol(A) mRNA was isolated by poly(A) affinity chromatography. A PCR was performed on cDNA from the hepatopancreas of live infected snails. Cerebratulidae develop in this region of the host snail, and transcripts for cercarial elastase are only present during this developmental stage. RNA was converted to cDNA using avian myeloblastosis virus reverse transcriptase (Invitrogen) as described by the manufacturer. PCR was performed using 5 μl of first strand cDNA as template and the primer set (5'-TTGCAACTCCATCGATTACGACG, 5'-GGTCTGCGAATTCGACG). Thermocycling conditions were 30 s at 94°C, 1 min at 55°C, and 1 min at 72°C for 30 cycles. Cerebratulidae were sequenced using the Taq polymerase. PCR products were resolved in 1% agarose gels. Individual DNA bands were excised, purified with gel extraction columns (Qiagen), and cloned using the TOPO TA cloning vector (Invitrogen). Plasmids containing inserts of correct size were sequenced.

Identification of Isoform Promoters Reveals Two Classes of Cercarial Elastase—A PCR strategy was adopted to clone regions of genomic DNA upstream of cercarial elastase genes (Fig. 2). Multiple libraries of genomic DNA, digested with individual restriction enzymes and ligated with universal linkers, were constructed. The libraries were made about the protease gene family. First, a reverse primer homologous to the N-terminal portion of the gene family would bind all elastase isoforms. Second, restriction sites in the non-coding upstream regions of genomic DNA were used to identify cercarial elastase homologues as previously described. Genomic DNA or cDNA was used as normal template. PCR was performed using 0.5 μg of template DNA and the primer set (5'-CCATTGTTCATCCTTGAATTTTGAGGAATTCGACG). Thermocycling conditions were 30 s at 94°C, 1 min at 45°C, and 1 min at 72°C for 36 cycles using Taq polymerase. Individual DNA bands were excised, purified, and cloned. Plasmids containing inserts of the predicted size were sequenced.

Results

Purification of Cercarial Elastase Reveals Multiple Isoforms—Cercarial elastase purification requires two steps. Gel filtration of cercarial secretions produces a single broad peak of elastase activity and was detected with the tetrapeptidyl substrateAAPF-pNA (Fig. 1A). Anion exchange chromatography was then used to further purify cercarial elastase activity. A shallow salt gradient revealed three peaks of activity (Fig. 1D). Each peak was associated with a 25-kDa protein product (Fig. 1C). Polyclonal antiserum generated from pure, recombinant, non-glycosylated cercarial elastase of bacterial origin was used in an immuno blot analysis of these three peaks. The high level of reactivity for each of the bands suggested strong sequence similarity (Fig. 1D). The lower 16-kDa band observed in each lane was a degradation product resulting from autoproteolysis (5).
of the individual PCR products confirmed that the bands were upstream genomic DNA from multiple related elastase genes. As a control, reverse primers to the GAPDH gene were also used in a separate experiment. GAPDH has been previously shown to be a single copy gene (11, 12), and PCR reactions yielded single products for each library, consistent with this finding (Fig. 2B).

Two genomic clones of ceraciar elastase, designated EL1 and EL2, have been reported by Pierrot et al. (13). EL1 was described as the genomic clone of the previously published ceraciar elastase cDNA sequence (SmCE-1a; Table I). EL2 was reported to contain a D125A mutation in the catalytic triad, yielding an inactive protease and also having a predicted transcript that was undetectable. An alignment of the predicted cDNA from EL1 and the first published cDNA (SmCE-1a) revealed only 25 single base pair substitutions (of 795) suggesting that these genes may be isoforms. A unique SpeI restriction site is present in SmCE-1a but absent in SmCE-1b. Digestion of RT-PCR products with SpeI demonstrated that both transcripts are produced and confirmed that the genes are isoforms and not strain variants (data not shown). Exhaustive attempts by RT-PCR with parasite cDNA to identify a transcript of the predicted catalytically dead gene, SmCE-1c (EL2), were similarly unsuccessful. Previously submitted DNA sequences to public data bases and their nomenclature relative to this paper are summarized (Table I).

Alignment of the upstream non-coding DNA sequences revealed a high degree of conservation centering on the TATA box. This region extends 36 bp upstream, 55 bp downstream, and is 80% identical. One of the isoform promoters contained a 10-bp deletion just upstream of the translational initiation codon (Fig. 3A). Additionally, three amino acids in the open reading frame of this fragment did not match the known elastase sequences indicating another isoform. A forward primer bridging the deletion region allowed for the full-length cloning of the mRNA downstream of this promoter by RT-PCR. Two PCR products were generated and sequenced. Both products were genes coding for ceraciar elastase, but each was significantly divergent from previously known amino acid sequences. These results bring the total number of ceraciar elastase genes in S. mansoni to five (Fig. 4).

Not All Isoforms Are Transcribed and/or Translated—N-terminal sequencing of the purified enzymes has confirmed the transcription of three of the five ceraciar elastase genes (Fig. 3B). Isoforms SmCE-1a and SmCE-1b are the most abundant species present in ceraciar extracts comprising over 99% of the protein and activity. Isoform SmCE-2a constitutes a minor component of ceraciar secretions. Isoform SmCE-2b has detectable levels of mRNA transcript, but attempts to purify active protein of this isoform were unsuccessful and indicate that its abundance is lower than our methods are able to detect. The last predicted isoform, SmCE-1c (EL2), could not be detected by RT-PCR using specific primers.

Positional Scanning, Combinatorial Substrate Libraries Demonstrate That the Two Major Isoforms Are Isoenzymes—Substrate preferences between the two major species of ceraciar elastase, SmCE-1a and SmCE-1b, were compared using two combinatorial peptide substrate libraries that together scanned substrate positions P4, P3, P2, and P1. Both enzymes prefer the same sets of amino acids for each of the positions scanned indicating that the isoforms are enzymatic equivalents (Fig. 5).

Position P4 has a marked preference for serine or threonine but will tolerate 18 other amino acids. Position P3 showed little preference. The P2 position is the most selective position of the enzyme preferring only proline. The selectivity for P2 in
Table I
Cercarial elastase sequences in public data bases

<table>
<thead>
<tr>
<th>New nomenclature</th>
<th>Previous nomenclature</th>
<th>DNA accession number</th>
<th>Reference</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SmCE-1a</td>
<td>Cercarial elastase</td>
<td>J03946</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>SmCE-1b</td>
<td>EL1</td>
<td>U31768</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SmCE-1c</td>
<td>EL2</td>
<td>Z70206</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>SmCE-1a</td>
<td>HP1</td>
<td>U31769</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**DNA library source. First cercarial elastase sequence published.**
-
**Genomic DNA library source. Large non-coding upstream region. Correct exon regions are 1-230, 360-660, 935-1165.**
-
**Genomic PCR source. Contains intron structure. AT added after 669 restores correct exon regions 1-230, 360-660, 935-1165.**
-
**Source genomic PCR, contains intronic information for CE-1a, missing first 2 amino acid of sequence. Correct exon regions are 1-168, 310-627, 1255-1477.**

**A**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Organism</th>
<th>Transcribed (RT-PCR)</th>
<th>Translated (cDNA)</th>
<th>Relative Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>SmCE-1a</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SmCE-1b</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SmCE-1c</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**B**

| Group 1 cercarial elastases | | |
|-----------------------------|-----|-----|-----|-----|-----|
| CE-1a                       | ✓   | ✓   | 45  |     |     |
| CE-1b                       | ✓   | ✓   | 45  |     |     |
| CE-1c                       | ✓   | ✓   | 0   |     |     |

**Group 2 cercarial elastases**

| CE-2a | ✓ | ✓ | 10 |
| CE-2b | ✓ | ✓ | 5  |

*Shown in Figure 4. ** Shown in Figure 1.A.E. *** Shown in Figure 10 D & 1C

Fig. 3. S. mansoni cercarial elastases and their transcription/translation properties. A, alignment of Group 1 and Group 2 non-coding upstream DNA. Shaded regions indicate identity within the three sequences. The TATA box and ATG start codon are boxed, and the identifying Group 2 deletion is underlined. B, two groups of cercarial elastase have been assigned according to upstream DNA sequence and gene sequence homology. Thin lines represent non-coding regions and boxes represent coding regions. Gaps between blocks represent predicted intron positions. Line and block lengths are not proportional.

SmCE-1b is slightly broader than SmCE-1a with alamines also accepted in the position. The pins for SmCE-1a and SmCE-1b was scanned as an enzyme mixture due to the high enzyme requirements of the P1 diverge library. Both isoforms were then tested with individual tetrapeptide substrates for each P1 amino acid that yielded cleavage activity and were found to be similar (data not shown).

Individual Substrates Confirm Combinatorial Substrate Library Predictions—Four tetrapeptide substrates representing both favorable (S/T/D/W) and unfavorable (R long) amino acids at positions P4 and P3 (Fig. 5) were compared relative to each other for cleavage by cercarial elastase activity (Fig. 6). P2 and P1 were fixed with proline and leucine respectively. The predicted rankings of substrate kinases from the chemical library was confirmed (SWPL > TWPL > RWPL > RRPL). An 11-fold difference in substrate kinetics was observed between a favorable (serine) and unfavorable (arginine) amino acid in the P4 position. The P3 position also influenced substrate kinetics with the favorable amino acid (tryptophan) having three times more activity than the unfavorable amino acid (arginine). The best tetrapeptide substrate of the four was SWPL.

S. hematobium and S. douthitti Contain Orthologous Cercarial Elastase Genes—A variety of methods were used to identify gene sequences within other species of schistosomes. De-natured PCR using S. hematobium cDNA template produced two similar but unique sequences designated ShCE-1a and ShCE-1b. A phosphoamino acid library was also screened, using a radiolabeled SmCE-1a probe, and identified a single full-length mRNA for ShCE-1a (Fig. 4 and 7). The same PCR strategy using S. douthitti genomic DNA also produced two fragments, SmCE-1a and SmCE-1b, of which SmCE-1b contained a unique intron that
does not share a location with any other known introns present in *S. mansoni* cercarial elastase sequences (Fig. 7).

**Cercarial Elastase Similarity across Species Follows a General Genetic Trend**—The phylogenetic relationships of all nine cercarial elastase amino acid sequences reported here were calculated using the PHYLIP software package (Fig. 8). Each set of isoforms from each species clusters together, *S. mansoni* and *S. hematobium* are closely linked, whereas the *S. douhitti* enzymes branch from the *S. hematobium* cluster. The most homologous serine proteases from other organisms were tryp-
confirmed that it is the sole histolytic activity present in cercarial secretions (3). As such, one would expect it to be conserved across schistosome species. The studies reported here confirm that it is highly conserved between S. mansoni and S. hematobium, and it is identifiable in the more distantly related S. douiti.

Cercarial elastase activity is encoded by a family of closely related genes. The presence of gene duplication and the resultant protease isoforms may reflect in part selection for redundancy in a key gene required for transmission from small to human host. Comparison of the diversity of cercarial elastase isoforms in S. mansoni demonstrated a high degree of similarity in their preferred substrates. This suggests that these two genes are not providing substrate cleavage diversity but may increase the total amount of protein produced in the acetabular glands through duplication.

The protease isoforms fall into two families identifiable by differences in their promoters. Expression of cercarial elastase is strictly limited to the sporocyst lifecycle stage in which cercariae develop (4, 14). However, cercarial elastase in S. mansoni has been shown to be expressed in both pre- and postacetabular cells. These two cell compartments, while contributing to cercarial secretions, produce different products. The preacetabular cells have cercarial elastase co-localized in vesicles containing high concentrations of calcium, whereas the postacetabular glands contain protease and mucopolysaccharides (15, 16). Perhaps the two promoter families represent gene products that localize to these two separate compartments.

Alternatively, the gene redundancy could be explained by a “gene in waiting hypothesis.” The duplication and subsequent divergence of these genes provides a potential mechanism to alter the substrate specificity of the protease. Such a reservoir of genes would provide protection for the parasite against host adaptation or enhancement and expansion of host range. This hypothesis can explain why there are multiple isoforms yet only two highly conserved genes accounting for the majority of protein produced. This hypothesis is also testable. Additional data on schistosomes infecting a variety of hosts would be predicted to show differential levels of cercarial elastase gene transcripts and products with conservation of the genes themselves.

High sequence conservation around the TATA box makes it impossible to identify specific potential transcription factor binding sites but does suggest that the assembly of the transcription machinery would involve most of this region. A better understanding of the transcriptional regulation of schistosome will be possible as more upstream sequence information is obtained by comparisons between and within species.

We have shown cercarial elastase activity to be the sum of multiple protease isoforms. Protease gene duplications are observed in three schistosome species and provide an opportunity to document the evolution of a gene family in the genus Schistosoma. Positional scanning combinatorial substrate libraries were used to demonstrate that the proteases expressed from the gene family are isozymes.

REFERENCES
Schistosome Gene Family


