# Arresting Tissue Invasion of a Parasite by Protease Inhibitors Chosen with the Aid of Computer Modeling[†]

Fred E. Cohen,[‡,§] Lydia M. Gregoret,[‡] Payman Amiri,[∥] Ken Aldape,[∥] Johnny Railey,[∥] and James H. McKerrow*,[∥,⊥]
*Departments of Pharmaceutical Chemistry, Box 0446, Medicine, Box 0120, and Pathology, Box 0506, University of California, San Francisco, San Francisco, California 94143, and Department of Veterans Affairs Medical Center, San Francisco, California 94121*

*Received December 6, 1990; Revised Manuscript Received July 9, 1991*

ABSTRACT: Computer modeling of the three-dimensional structure of an enzyme, based upon its primary sequence alone, is a potentially powerful tool to elucidate the function of enzymes as well as design specific inhibitors. The cercarial (larval) protease from the blood fluke *Schistosoma mansoni* is a serine protease hypothesized to assist the schistosome parasite in invading the human circulatory system via the skin. A three-dimensional model of the protease was built, taking advantage of the similarity of the sequence of the cercarial enzyme to the trypsin-like class of serine proteases. A large hydrophobic S-1 binding pocket, suspected from previous kinetic studies, was located in the model and confirmed by new kinetic studies with both synthetic peptide substrates and inhibitors. Unexpected structural characteristics of the enzyme were also predicted by the model, including a large S-4 binding pocket, again confirmed by assays with synthetic peptides. The model was then used to design a peptide inhibitor with 4-fold increased solubility, and a series of synthetic inhibitors were tested against live cercariae invading human skin to confirm that predictions of the model were also applicable in a biologic assay.

The elucidation of three-dimensional structures of enzymes by X-ray crystallography, coupled with advances in computer graphics, has led to studies of inhibitor and drug design by computer modeling (Blundell et al., 1987a; Ripka et al., 1987; Roberts et al., 1988; Freudenreich et al., 1984; McQuade et al., 1990; Knight, 1990). Unfortunately, not all enzymes are readily crystallized because of problems of quantity, purity, or chemistry. Alternatively, a molecular model of the enzyme structure could be constructed on the basis of amino acid sequence data alone. The structure of a protein can be modeled successfully if the structure of proteins with homologous sequences are known (Blundell et al., 1987b; Greer et al., 1989). The trypsin family of serine proteases offers an ideal case for model building by homology: many homologous sequences and structures are available, a critical catalytic triad of residues (His, Asp, and Ser) are evenly distributed throughout the sequence, and sequence–structure correlates for specific noncatalytic residues exist to facilitate sequence alignment (James et al., 1978).

Schistosomiasis (bilharziasis) is a parasitic disease caused by schistosomes (blood flukes) affecting over 250 million people (Cline, 1989). Upon stimulation by the lipid present at the surface of skin, invasive, aquatic schistosome larvae, called cercariae, secrete a serine protease structurally related to the trypsin family of enzymes (Stirewalt, 1974; McKerrow et al., 1985a; Newport et al., 1988). The cercarial protease has been shown in vitro to cleave keratin, laminin, fibronectin, and type IV collagen as well as elastin (McKerrow et al., 1985a,b). It is postulated that the parasite uses this enzyme to penetrate the epidermis, basement membrane, and extracellular matrix of the dermis (McKerrow et al., 1989). The primary amino acid sequence of the protease has been predicted from the sequence of a cDNA clone (Newport et al., 1988), and purified enzyme is available for biochemical assays (McKerrow et al., 1985a). It is possible to assay skin invasion by living cercariae in vitro (Clegg & Smithers, 1972). One test of the hypothesis that the protease of cercariae is necessary for invasion would be if very specific, nontoxic inhibitors of the enzyme were identified and assayed for their effect on invasion of skin by schistosome larvae. Also, although inhibitors of this enzyme would act to prevent infection rather than to cure victims of the disease, the design of a topical ointment which could suppress skin invasion might be a very useful adjunct to other control measures (Cherfas, 1989).

In order to better understand the substrate specificity of the cercarial protease, and to test the validity of computer modeling of enzyme structure in predicting inhibitors for use in a biologic assay of enzyme function, we have built a three-dimensional model of the enzyme, taking advantage of its sequence similarity to the trypsin-like serine proteases (Greer, 1981, 1990). Using the model, we are able to rationalize existing data on the P-1 specificity and design inhibitors that exploit the P-1, P-3, and P-4 substrate side-chain binding subsites. These predictions are then tested with standard in vitro enzyme kinetics assays, using peptide-based substrate analogues and inhibitors, and also by measuring the effectiveness of the inhibitors at suppressing cercarial invasion of human skin (Clegg & Smithers, 1972), thus testing the hypothesis that the protease itself facilitates invasion of host skin.

## Experimental Procedures

*Modeling the Structure of Cercarial Protease.* Previous studies with mechanism-based inhibitors suggested that the cercarial protease was a serine protease (McKerrow et al., 1985a). The sequence of cercarial protease is similar to that of the trypsin-like class of serine proteases (Newport et al.,

[†] Supported by grants from the NIAID (AI20452 to J.H.M.), the NIGMS (GM39900 to F.E.C.), the Defense Advanced Research Projects Agency (N00014-86K-0757 to F.E.C.), and the Searle Scholars/Chicago Community Trust (F.E.C.) and by a fellowship from the Giannini Foundation (J.R.).

* Author to whom correspondence should be addressed at the Department of Pathology, Box 0506, HSW 501, University of California, San Francisco, San Francisco, CA 94143.

[‡] Department of Pharmaceutical Chemistry, University of California, San Francisco.

[§] Department of Medicine, University of California, San Francisco.

[∥] Department of Pathology, University of California, San Francisco.

[⊥] Department of Veterans Affairs Medical Center, San Francisco.

0006-2960/91/0430-11221$02.50/0

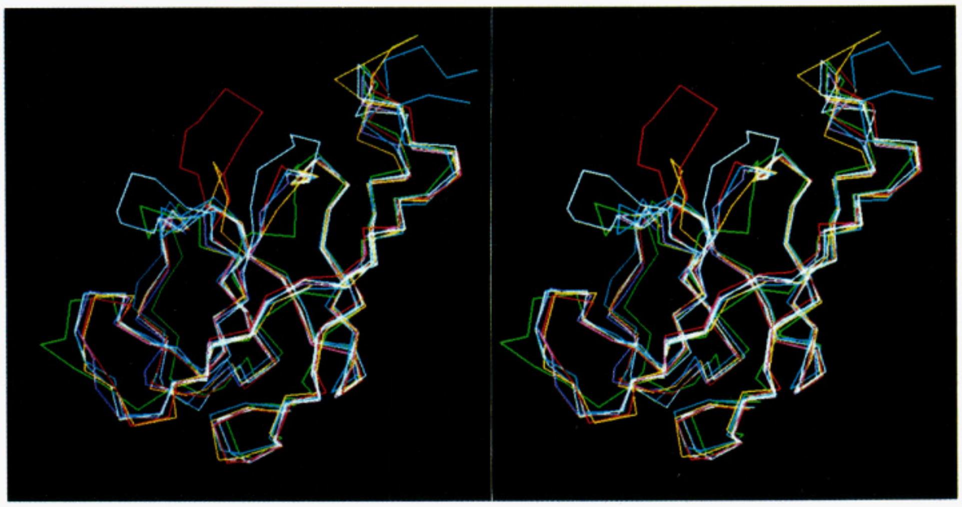


FIGURE 1: Structural superposition of the $C^\alpha$ backbone of the first domain of six crystallographically determined serine proteases. Catalytic residues His 57 and Asp 102 (chymotrypsin numbering scheme) are indicated. The lack of structural conservation in loop regions is evident.

Table I: Amino Acid Identity Matrix for Structural Alignment of Serine Proteases[a]

| | 3EST | 3PTN | 4CHA | 2PKA | 3RP2 | 1SGT | CERC |
|---|---|---|---|---|---|---|---|
| 3EST | | 36.7 | 38.9 | 31.8 | 31.5 | 30.2 | 19.7 |
| 3PTN | 85 | | 42.6 | 37.4 | 32.2 | 30.5 | 20.9 |
| 4CHA | 91 | 96 | | 32.2 | 30.1 | 29.7 | 19.4 |
| 2PKA | 75 | 85 | 74 | | 32.0 | 25.1 | 17.5 |
| 3RP2 | 73 | 72 | 68 | 73 | | 24.2 | 24.3 |
| 1SGT | 70 | 68 | 67 | 57 | 54 | | 20.9 |
| CERC | 47 | 48 | 45 | 41 | 56 | 48 | |

[a] The upper diagonal of the matrix shows the percentage of identical residues in identical positions. The lower diagonal of the matrix is the raw number of identities. Values for cercarial elastase (CERC) are based on a sequence alignment. Abbreviations: 3EST, porcine elastase; 3PTN, bovine trypsin; 4CHA, bovine α-chymotrypsin; 2PKA, porcine kallikrein A; 3RP2, rat mast cell protease II; 1SGT, *S. griseus* trypsin.

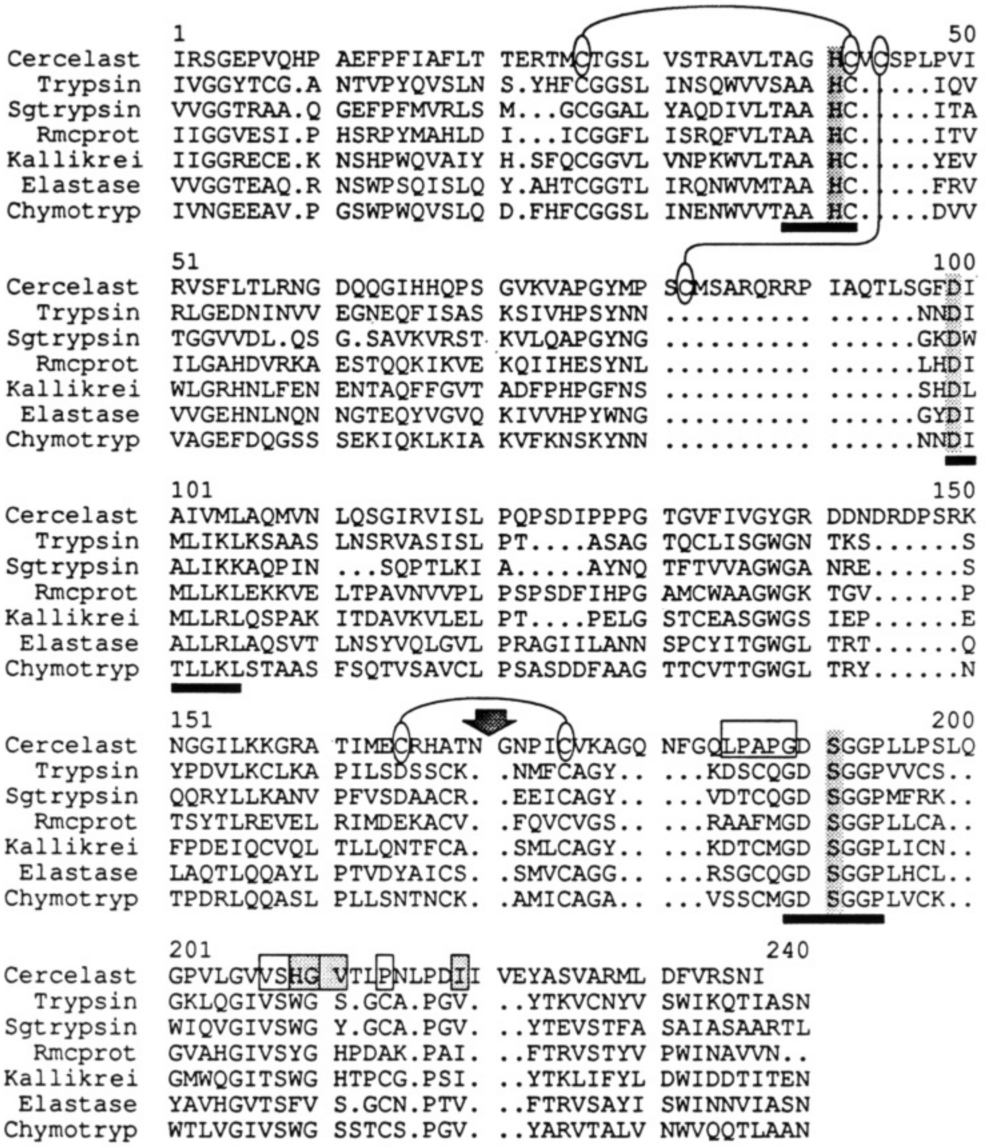
1988). Sequence identity is quite high in the region of the three residues of the catalytic triad (His-57, Asp-102, and Ser-195, chymotrypsin numbering scheme; His-41, Asp-99, and Ser-191, cercarial protease sequential numbering) but rather low overall (see Table I). Sequence identities smaller than 30% make accurate modeling difficult, particularly in unconserved regions (Blundell et al., 1987b). Consequently, we focused the most attention on modeling the active site and the substrate side-chain specificity pockets.

We used the mammalian serine proteases whose X-ray coordinates are available in the Brookhaven Protein Data Bank (PDB) (Abola et al., 1987; Bernstein et al., 1977) as a basis set for modeling the cercarial enzyme. The cercarial protease is equally similar in sequence identity to the bacterial serine proteases (*Streptomyces griseus* proteases A and B and α-lytic protease) and mammalian serine proteases. The chain length of the cercarial enzyme is comparable to that of the mammalian enzymes. Thus, a mammalian template requires fewer insertions to accommodate the cercarial sequence. At the time of modeling, the structures of six different mammalian or mammalian-like[1] proteases had been deposited in the data bank. These were bovine trypsin [3PTN (Walter et al., 1982)], porcine pancreatic elastase [3EST (Meyer et al., 1988)], bovine chymotrypsin [4CHA (Tsukada & Blow, 1984)], rat mast cell protease [3RP2 (Remington et al., 1988)], porcine kallikrein [2PKA (Bode et al., 1983)], and *S. griseus* trypsin [1SGT (Read & James, 1984)].

The backbones of the six known structures were superimposed manually with the assistance of computer graphics (UCSF MidasPlus: Ferrin et al., 1988; Jarvis et al., 1988). Each protease structure was divided into its two six-stranded Greek key β-barrel domains, and the first and second domains were superimposed separately (see Figure 1). By noting which amino acid residues were in topologically equivalent positions, a structurally based multiple sequence alignment was derived. The sequence of cercarial protease was aligned manually to the multiple alignment. There was some ambiguity as to where to place a fairly large insertion of 10–15 residues between the catalytic histidine and aspartate. Two different alignments appeared equally reasonable. To clarify this issue, the three-dimensional structures implied by the alternative alignments were constructed. The backbone of porcine elastase (3EST) was used as the template for the substitution of aligned amino acids. The alternatives were indistinguishable on the basis of burial of hydrophobic residues. Automatic alignment using the program ProfileGap from the University of Wisconsin UWGCG package placed the insertion closer to the catalytic aspartate. This alignment was one of the two manually generated alternatives. The final alignment is shown in Figure 2. The final model of cercarial protease was built by using the backbone of porcine pancreatic elastase. Elastase was chosen because it was the closest in length (240 residues) to cercarial protease (237 residues). Side chains were substituted in their statistically most frequently observed conformations (Ponder & Richards, 1987), except in the region of the substrate binding site. Here, side chains were modeled to match the conformations seen in the basis set structures, because active-site geometries are highly conserved. Although it was possible to identify conformationally sensible loops to join the β-strands, unalignable loops were not modeled for two reasons. First, with the exception of the loop following the catalytic histidine, most loops were far from the active site. Second, in cases of low sequence identity, loop conformations are not guaranteed to be conserved even if loop length is conserved (Read & James, 1984). Moreover, energy calculations do not provide a gold standard for validating or invalidating alternative loop conformations (Novotny et al., 1988).

Substrate binding was modeled by using the coordinates of α-lytic protease mutant Met 192 → Ala 192 with the bound boronic acid inhibitor Ala-Ala-Pro-Phe-BOH (Bone et al.,

[1] Mammalian-like refers to *S. griseus* trypsin, which is more closely related in sequence and structure to the mammalian proteases than to the other bacterial proteases in the data bank.



```
            1                                                  50
Cercelast   IRSGEPVQHP AEFPFIAFLT TERTMCTGSL VSTRAVLTAG HCVCSPLPVI
  Trypsin   IVGGYTCG.A NTVPYQVSLN S.YHFCGGSL INSQWVVSAA HC.....IQV
Sgtrypsin   VVGGTRAA.Q GEFPFMVRLS M...GCGGAL YAQDIVLTAA HC.....ITA
  Rmcprot   IIGGVESI.P HSRPYMAHLD I...ICGGFL ISRQFVLTAA HC.....ITV
Kallikrei   IIGGRECE.K NSHPWQVAIY H.SFQCGGVL VNPKWVLTAA HC.....YEV
 Elastase   VVGGTEAQ.R NSWPSQISLQ Y.AHTCGGTL IRQNWVMTAA HC.....FRV
Chymotryp   IVNGEEAV.P GSWPWQVSLQ D.FHFCGGSL INENWVVTAA HC.....DVV

            51                                                100
Cercelast   RVSFLTLRNG DQQGIHHQPS GVKVAPGYMP SCMSARQRRP IAQTLSGFDI
  Trypsin   RLGEDNINVV EGNEQFISAS KSIVHPSYNN .......... ......NNDI
Sgtrypsin   TGGVVDL.QS G.SAVKVRST KVLQAPGYNG .......... ......GKDW
  Rmcprot   ILGAHDVRKA ESTQQKIKVE KQIIHESYNL .......... ......LHDI
Kallikrei   WLGRHNLFEN ENTAQFFGVT ADFPHPGFNS .......... ......SHDL
 Elastase   VVGEHNLNQN NGTEQYVGVQ KIVVHPYWNG .......... ......GYDI
Chymotryp   VAGEFDQGSS SEKIQKLKIA KVFKNSKYNN .......... ......NNDI

            101                                               150
Cercelast   AIVMLAQMVN LQSGIRVISL PQPSDIPPPG TGVFIVGYGR DDNDRDPSRK
  Trypsin   MLIKLKSAAS LNSRVASISL PT....ASAG TQCLISGWGN TKS......S
Sgtrypsin   ALIKKAQPIN ...SQPTLKI A.....AYNQ TFTVVAGWGA NRE......S
  Rmcprot   MLLKLEKKVE LTPAVNVVPL PSPSDFIHPG AMCWAAGWGK TGV......P
Kallikrei   MLLRLQSPAK ITDAVKVLEL PT....PELG STCEASGWGS IEP......E
 Elastase   ALLRLAQSVT LNSYVQLGVL PRAGIILANN SPCYITGWGL TRT......Q
Chymotryp   TLLKLSTAAS FSQTVSAVCL PSASDDFAAG TTCVTTGWGL TRY......N

            151                                               200
Cercelast   NGGILKKGRA TIMECRHATN GNPICVKAGQ NFGQLPAPGD SGGPLLPSLQ
  Trypsin   YPDVLKCLKA PILSDSSCK. .NMFCAGY.. ...KDSCQGD SGGPVVCS..
Sgtrypsin   QQRYLLKANV PFVSDAACR. .EEICAGY.. ...VDTCQGD SGGPMFRK..
  Rmcprot   TSYTLREVEL RIMDEKACV. .FQVCVGS.. ...RAAFMGD SGGPLLCA..
Kallikrei   FPDEIQCVQL TLLQNTFCA. .SMLCAGY.. ...KDTCMGD SGGPLICN..
 Elastase   LAQTLQQAYL PTVDYAICS. .SMVCAGG.. ...RSGCQGD SGGPLHCL..
Chymotryp   TPDRLQQASL PLLSNTNCK. .AMICAGA.. ...VSSCMGD SGGPLVCK..

            201                                  240
Cercelast   GPVLGVVSHG VTLPNLPDII VEYASVARML DFVRSNI
  Trypsin   GKLQGIVSWG S.GCA.PGV. ..YTKVCNYV SWIKQTIASN
Sgtrypsin   WIQVGIVSWG Y.GCA.PGV. ..YTEVSTFA SAIASAARTL
  Rmcprot   GVAHGIVSYG HPDAK.PAI. ..FTRVSTYV PWINAVVN..
Kallikrei   GMWQGITSWG HTPCG.PSI. ..YTKLIFYL DWIDDTITEN
 Elastase   YAVHGVTSFV S.GCN.PTV. ..FTRVSAYI SWINNVIASN
Chymotryp   WTLVGIVSWG SSTCS.PGV. ..YARVTALV NWVQQTLAAN
```

FIGURE 2: Alignment of cercarial elastase with proteins of known structure. Abbreviations: Cercelast, cercarial elastase; Trypsin, bovine trypsin (3PTN); Sgtrypsin, *S. griseus* trypsin (1SGT); Rmcprot, rat mast cell protease (3RP2); Kallikrei, porcine kallikrein (2PKA); Elastase, porcine elastase (3EST); Chymotryp, bovine chymotrypsin (4CHA). Only structurally conserved amino acids are shown for the known structures. The catalytic residues are stippled in grey and the region surrounding these residues is underlined. Residues near the S-1 binding site of cercarial elastase are enclosed in clear boxes, and residues comprising the S-4 site are enclosed in stippled boxes. The location of a large deletion that affects S-4 specificity is indicated with an arrow. Cysteine residues postulated to be connected by disulfide bonds are also indicated with ovals and connecting lines.

1989; PDB entry 1P08). The backbone of the inhibitor was positioned in the binding cleft of cercarial protease by superimposing the $C^\alpha$ and $C^\beta$ atoms of the catalytic residues in $\alpha$-lytic protease and the porcine elastase derived cercarial protease model.

*Assay of Protease with Peptide Substrates.* The serine protease was isolated from the acetabular glands of cercariae as described previously (McKerrow et al., 1985a). All tetrapeptide thioester (Sbzl)[2] substrates with the exception of MeO-Suc-Ala-Ala-Pro-Lys-Sbzl (a gift from Dr. James C. Powers, School of Chemistry, Georgia Institute of Technology, Atlanta, GA) were purchased from Enzyme System Products (Dublin, CA). Stock solutions (25 mM) of all substrates were prepared in DMSO. The rates of hydrolysis of substrates were measured by adding 1–10 $\mu$L of substrate and 25 $\mu$L of 1 mM 4,4′-dithiodipyridine in DMSO to buffer (100 mM glycine/NaOH, 2 mM $CaCl_2$, pH 9.0) to give a total volume of 1.0 mL. A 10-$\mu$L enzyme sample was added to start the reaction. The increase in the absorbance at 325 nm was followed on a Gilford spectrophotometer. An $\epsilon$ value of 19 800 was used for the thiopyridine production.

Assays of pNA- and AMC-derivatized peptides were carried out as described in detail previously (McKerrow et al., 1985a). MeO-Suc-Ala-Ala-Pro-Phe-pNA was from Vega Biochemicals, Tucson, AZ. MeO-Suc-Ala-Ala-Pro-Leu-pNA was a gift of Dr. David Agard, Department of Biochemistry, UCSF. All other pNA substrates were a gift of Dr. Corey Largman, Veterans Administration Hospital, Martinez, CA. All AMC substrates were from Enzyme System Products, Dublin, CA.

For each peptide substrate, the substrate concentration range was varied over 10-fold. The kinetic constants were determined from the initial rates of product formation by using a least-squares analysis according to method of Lineweaver and Burk (1934). Five points were measured for each plot and corre-

[2] Abbreviations: Sbzl, benzyl thioester; DMSO, dimethyl sulfoxide; MeO-Suc, methoxysuccinyl; pNA, *p*-nitroanilide; AMC, 7-amino-4-methylcoumarin; CMK, chloromethyl ketone.

lation coefficients were all greater than 0.95.

*Testing of Synthetic Peptide Inhibitors versus Purified Protease.* All the chloromethyl ketone derivatized peptide inhibitors were from Enzyme System Products (Dublin, CA). Stock solutions of inhibitors (25 mM) were prepared in DMSO. The rate of irreversible inactivation of the protease was followed by withdrawing a 25-$\mu$L sample at four time intervals after mixing 125 $\mu$L of enzyme with 5 $\mu$L of inhibitor. Kitz and Wilson's methods of analysis were used for calculation of $K_i$, $k_3$, and $k_3/K_i$ (Kitz & Wilson, 1962) using MeO-Suc-Ala-Ala-Pro-Phe-Sbzl as substrate. Inactivation constants for the boronic acid derivatized peptides [MeO-Suc-Ala-Ala-Pro-(D or L)-boro-Phe-OH, a gift from Dr. Charles Kettner, E. I. du Pont de Nemours, Wilmington, DE] were calculated by using MeO-Suc-Ala-Ala-Pro-Phe-Sbzl as the substrate at 125 and 250 $\mu$M. Five different concentrations of each inhibitor were used per substrate concentration. The data were plotted according to Dixon, and $k_{inact}$ values were extrapolated from the graph (Dixon, 1953).

*Testing of Inhibitors for Their Effect on Cercarial Penetration of Human Skin.* A modification of the assay developed by Clegg and Smithers (1972) was used. Two Plexiglass chambers were manufactured which could be screwed together to hold a 25 × 25 mm section of skin at the interface of the two chambers. Human skin was obtained from autopsy (6–12 h following death) or from amputation specimens received at the Department of Surgical Pathology, UCSF. The chamber below the skin was filled with tissue culture medium (DME H16) prewarmed to 37 °C. After the skin was clamped, the upper chamber was filled with water containing 9000 cercariae of *Schistosoma mansoni* (Puerto Rican strain). Cercariae will follow a thermal gradient, and they are stimulated by the lipid on the surface of skin to invade (Stirewalt, 1974). After 1-h exposure to cercariae the skin was removed, fixed in 10% phosphate-buffered formalin for 24 h, and sectioned at 2-mm intervals. Following dehydration and routine paraffin embedding, 5-$\mu$m sections were cut by microtome and stained with hemotoxylin and eosin. The number of cercariae that had invaded the epidermis or dermis was counted in each section. The percent inhibition of invasion was calculated as

$$1.0 - \frac{\text{no. of cercariae invaded with inhibitor}}{\text{no. of cercariae invaded in control}} \times 100$$

All assays were done in triplicate, and the standard deviation of the mean was calculated.

## Results

*Model Evaluation.* The most crucial part of any model, predicted on the basis of its similarity to a structure of known three-dimensional structure, is the sequence alignment (Greer, 1981, 1990). If the alignment is incorrect, all subsequent modeling that depends on it, such as the positioning of side chains and loops, will be incorrect. In the case of the cercarial protease model, the spatial location of cysteine residues reinforces confidence in the sequence alignment. The cercarial protease has six cysteines. Although the disulfide connectivity has not been determined experimentally, it is possible to assign the likely pairings. One pair of cysteines (Cys 26–Cys 42) is conserved in all proteases. The remaining four cysteines are in unique positions in the sequence. Two of these cysteines are in the first domain and two in the second. Disulfide pairing usually occurs between cysteines local in sequence (Thornton, 1981), so it is expected that the four unique cysteines form disulfide pairs within their respective domains (Figure 2). In the three-dimensional model, the cysteines that are postulated to be paired are close to one another spatially. If this alignment was incorrect, the cysteines would be less likely to fall within a disulfide pairing distance.

Another feature of the model that supports the sequence alignment is the distribution of hydrophobic and hydrophilic residues. Trypsin-like serine proteases are composed primarily of antiparallel $\beta$-sheets arranged as two six-stranded Greek key $\beta$-barrels (Richardson, 1981). "Exterior strands", that is, $\beta$-strands that face the protein interior on one side and the solvent exterior on the other side, have an alternating periodicity of hydrophobic and hydrophilic residues. If the sequence alignment was out of phase by an odd number of residues, hydrophobic residues would be found on the exterior of the protein and hydrophilic residues on the interior. Buried residues in the model are composed almost exclusively of hydrophobic residues. There are also some hydrophobic residues on the exterior of the model. This is consistent with the typical distribution of side-chain functionalities on a protein surface. However, since our model does not contain loops, some of the exposed hydrophobic residues may actually be buried in the true structure. The cercarial protease model also does not contain any buried charged residues. These would be energetically unfavorable.

We did not model any features that relied on arbitrary decisions. For instance, we did not model side-chain conformations in the core of the protein accurately because our primary interest is inhibitor binding—a surface phenomenon. With low sequence identity, shifts in the protein backbone of cercarial protease relative to any of the mammalian serine proteases used to model the structure are likely. The average root-mean-square deviation for backbone atoms between structures with 20% sequence identity is 1.8 Å (Chothia & Lesk, 1986). Such large shifts can influence side-chain packing dramatically, making accurate modeling of the side-chain conformations difficult. Length-unconserved loops were also not modeled. These are the regions of the lowest sequence identity and presumably the lowest structural similarity. The three-dimensional structure of longer loops in particular is not well conserved when sequence identity is low (Greer, 1990).

*Testing of Synthetic Peptide Inhibitors and Substrates Predicted To Be Optimal for the Enzyme by the Computer Model. (A) The P-1 Site.* The structure of the $\alpha$-lytic protease mutant Met → Ala 192 with a tetrapeptide protease inhibitor has been determined crystallographically (Bone et al., 1987). The tetrapeptide inhibitor backbone mutant from this X-ray structure was used to position a peptide substrate in the active site of the cercarial protease model.

We first tested the validity of the model by studying the binding specificity of the S-1 cleft of the enzyme. Figure 3 shows that the S-1 cleft in the model is large and hydrophobic, suggesting that a large hydrophobic amino acid at the P-1 position would be bound optimally. We confirmed this by examining a set of synthetic peptide substrates in which both the P-1 amino acid and the leaving group were varied (Table II). Regardless of the leaving group used to detect substrate hydrolysis, a large hydrophobic amino acid at P-1 always gave the most optimal substrate, confirming our previous observations with a more limited group of substrates (McKerrow et al., 1985a). The predictions of S-1 binding site specificity were also borne out when examining a set of tetrapeptide inhibitors (Table III). Phenylalanine and leucine at P-1 gave optimal $k_3/K_i$.

There was a preference for substrates with the smaller thiobenzyl and $p$-nitroanilide leaving groups over the bulkier aminomethylcoumarin derivatives (Table II). The differences in $k_{cat}/K_m$ noted between the nitroanilide and thioester sub-

Table II: Kinetic Constants for Hydrolysis of Tetrapeptide Substrates by the Cercarial Protease[a]

| substrate | $K_m$ ($\mu$M) | $k_{cat}$ ($s^{-1}$) | $k_{cat}/K_m$ ($M^{-1}\ s^{-1}$) |
|---|---|---|---|
| P-4 P-3 P-2 P-1 | | | |
| MeO-Suc-Ala-Ala-Pro-Phe-Sbz1 | 96 | 19.4 | 202 100 |
| MeO-Suc-Ala-Ala-Pro-Leu-Sbz1 | 464 | 7.5 | 16 200 |
| MeO-Suc-Ala-Ala-Pro-Val-Sbz1 | very low activity[b] | | |
| MeO-Suc-Ala-Ala-Pro-Ala-Sbz1 | no activity | | |
| MeO-Suc-Ala-Ala-Pro-Lys-Sbz1 | no activity | | |
| MeO-Suc-Phe-Ala-Pro-Phe-Sbz1 | 244 | 3.48 | 14 000 |
| MeO-Suc-Trp-Ala-Pro-Phe-Sbz1 | no activity | | |
| MeO-Suc-Ala-Ala-Pro-Leu-pNA | 118 | 0.33 | 2 800 |
| MeO-Suc-Ala-Ala-Pro-Phe-pNA | 119 | 0.19 | 1 600 |
| MeO-Suc-Ala-Ala-Pro-Met-pNA | 300 | 0.05 | 185 |
| MeO-Suc-Ala-Ala-Pro-Nle-pNA | 300 | 0.02 | 56 |
| MeO-Suc-Ala-Ala-Pro-Val-pNA | very low activity[b] | | |
| MeO-Suc-Ala-Ala-Pro-Ile-pNA | very low activity[b] | | |
| MeO-Suc-Ala-Ala-Pro-Phe-AMC | low but detectable activity | | |
| MeO-Suc-Ala-Ala-Pro-Val-AMC | no activity | | |
| MeO-Suc-Ala-Ala-Pro-Ala-AMC | no activity | | |

[a] $r^2$ for plots were 0.95 → 0.99. P-1, P-2, etc. refer to the position of residues relative to the site of enzyme-catalyzed cleavage (I. Schecter). In this case the bond between the peptide and the Sbz1, pNA, or AMC leaving group (which allows spectrophotometric or spectrofluorometric measurement) is the one cleaved. The S-1 binding pocket of the enzyme would accommodate the P-1 side chain. MeO-Suc = methoxysuccinyl blocking group. [b] Absorbance change less than 0.001 Å in 40 min with 1.5 mM substrate.

Table III: Inhibition of Cercarial Protease by Chloromethyl Ketone Derivatized Peptides[a]

| inhibitors | $K_i$ ($\mu$M) | $k_3$ ($s^{-1} \times 10^3$) | $k_3/K_i$ ($M^{-1}\ s^{-1}$) |
|---|---|---|---|
| P-4 P-3 P-2 P-1 | | | |
| MeO-Suc-Ala-Ala-Pro-Leu-CMK | 12 | 18 | 1485 |
| MeO-Suc-Ala-Ala-Pro-Phe-CMK | 13 | 11 | 798 |
| MeO-Suc-Ala-Ala-Pro-Trp-CMK | 20 | 10 | 493 |
| MeO-Suc-Ala-Ala-Pro-Val-CMK | NI[b] | | 13[c] |
| MeO-Suc-Ala-Ala-Pro-Ala-CMK | NI | | 0.7[c] |
| MeO-Suc-Ala-Lys-Pro-Phe-CMK | 7 | 37 | 563 |
| MeO-Suc-Ala-Ala-Pro-Leu-CMK | 12 | 18 | 1485 |
| MeO-Suc-Trp-Ala-Pro-Leu-CMK | 2 | 8 | 3846 |
| MeO-Suc-Ala-Ala-Pro-Phe-CMK | 13 | 11 | 798 |
| MeO-Suc-Phe-Ala-Pro-Phe-CMK | 1 | 6 | 5483 |
| MeO-Suc-Trp-Ala-Pro-Phe-CMK | 12 | 6 | 521 |

[a] Abbreviations: CMK = chloromethyl ketone; MeO-Suc = methoxysuccinyl blocking group. [b] NI = no inhibition. [c] These values are $k_{obsd}/[I]$, which are equal to $k_3/K_i$ since $[I] \ll K_i$.

strates are probably due to differences in the rate-limiting step. The thioesters are much more reactive and the acylation rate is very fast. The rate-limiting step is probably deacylation. In contrast, acylation and deacylation rates are similar for the nitroanilide substrates, and in some cases acylation is rate limiting (Stein et al., 1987).

Table II shows that the increases noted in $k_{cat}/K_m$ for the best substrates often result more from differences in $k_{cat}$ than in $K_m$. This may reflect nonprotective binding of the substrates with respect to the "catalytic register" (Craik et al., 1985). $K_m$ is affected by all possible productive and nonproductive binding domains. But even when the substrate is bound well, the position with respect to the catalytic apparatus may not be optimal, resulting in incorrect orientation of the nucleophile relative to the scissile bond (Jencks, 1987).

We also explored the size limitations of the P-1 residue. For substrate hydrolysis, the optimal side-chain size appears to be phenylalanine (Figure 3). Leucine at P-1 is the best chloromethyl ketone inhibitor. Tryptophan defines the size limit of the P-1 pocket: the substrate analogue MeO-Suc-AAPW-Sbzl is a poor substrate, but the chloromethyl ketone inhibitor MeO-Suc-AAPW-CMK is still reasonable, with a 20 $\mu$M inhibition constant. According to the model, tryptophan can fit into the P-1 pocket, but distortion in $\chi_2$ away from the optimal value is required to ameliorate some steric conflicts. Residue 185 of cercarial protease is leucine. This position is equivalent to residue 189 in the chymotrypsin numbering scheme. Trypsin has an aspartate at this position which gives the enzyme specificity toward positively charged substrates. Chymotrypsin, which prefers large, hydrophobic residues, has a serine at position 189. The size limitation of the chymotrypsin S-1 pocket appears to be slightly larger than that of cercarial protease (Dorovska et al., 1972). This is sensible, given the size difference between leucine and serine.

$\beta$-Branching of the P-1 amino acid significantly reduced activity of both substrates and inhibitors (Table III). This observation provided us with the opportunity to use the model as a tool to identify specific residues that might sterically hinder $\beta$-branched amino acids in P-1. In the model structure, $\beta$-branching at P-1 would result in a steric conflict with residue Pro 188. Pro 188 occurs in a loop extension unique to the cercarial protease. Most other eukaryotic serine proteases have Cys at the position analogous to 187 (191 chymotrypsin numbering) that participates in a disulfide bridge. This covalent cross-link, which pulls the loop (residues 189–192) away from the P-1 binding pocket, is replaced by alanine in the cercarial protease. This may result in a constriction of the S-1 pocket.

(*B*) *The P-2 Site.* Proline in the P-2 site was not altered in the test peptides, since it was thought that the restricted geometry enhanced binding. For $\alpha$-lytic protease, substituting alanine for proline at this position reduces $k_{cat}/K_m$ by a factor of 10 (Bone et al., 1987).

(*C*) *The P-3 Site.* The P-3 site is solvent exposed. A hydrophilic residue at this position should improve solubility, while not affecting binding affinity (see invasion assay results below).

(*D*) *The P-4 Site.* While P-1 specificity of the cercarial protease was similar to that of chymotrypsin, an unexpected prediction of the model was that the cercarial enzyme would tolerate bulkier hydrophobic side chains at the P-4 subsite than chymotrypsin (Figure 4). As compared to chymotrypsin, the P-4 site is much more exposed in the cercarial protease—a loop that hangs over this site in chymotrypsin is missing. This loop is situated between residues 159 and 162 in the cercarial protease sequential numbering and contains residues 170 through 178 in the proteases of known three-dimensional structure. Given the lack of this loop, we speculated that a large side chain of the inhibitor or substrate could fit at the P-4 subsite. To test this prediction, substrates with large, hydrophobic amino acids (Phe, Trp) at this position were assayed. These large amino acids substantially abolished or diminished substrate hydrolysis (Table II). However, chloromethyl ketone inhibitors with tryptophan at P-4 and phenylalanine or leucine at P-1 worked well (Table III). We speculate that the interaction of large hydrophobic residues at P-4 with the residues lining the P-4 pocket distorts the geometry of the scissile bond relative to the active site without destroying binding affinity for the inhibitor. The lower $k_3/K_i$ for MeO-Suc-WAPF-CMK versus MeO-Suc-WAPL-CMK suggests that the conformational flexibility of the leucine side chain is useful in filling the P-1 pocket when the tetrapeptide CMK has been anchored to the binding cleft by the interaction of tryptophan with the P-4 pocket. This echoes the nonpro-

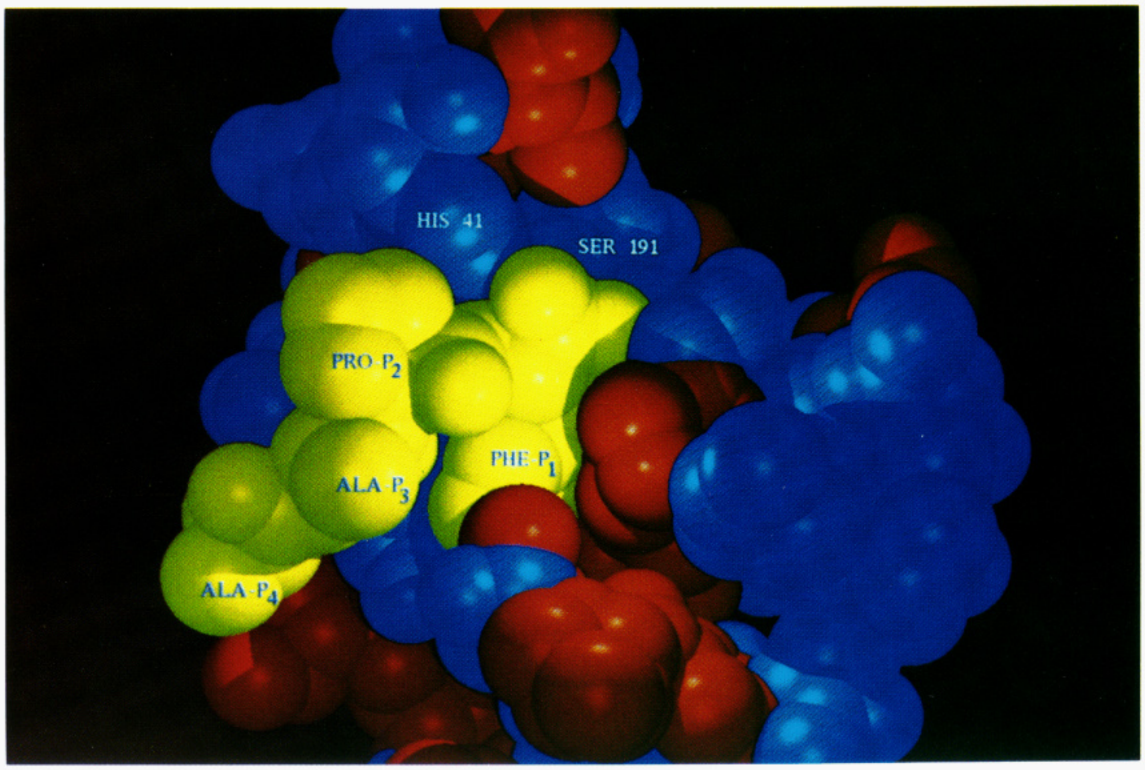


FIGURE 3: Model of the active site of the cercarial protease with the substrate Ala-Ala-Pro-Phe. Hydrophobic amino acids are colored red and hydrophilic amino acids are colored blue. Catalytic residues Ser 191 and His 41 are indicated.

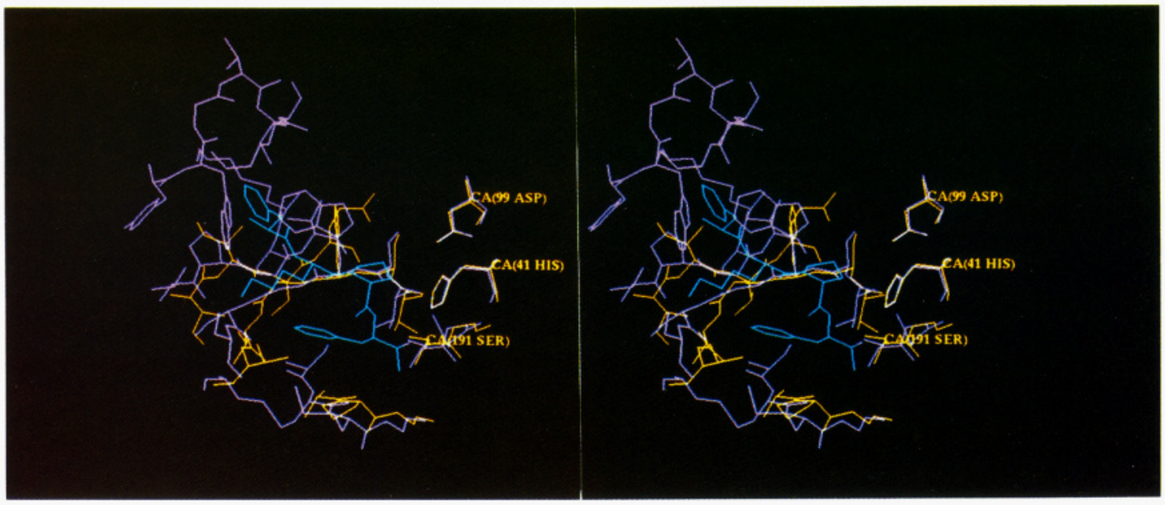


FIGURE 4: Comparison of cercarial protease (yellow) and chymotrypsin (magenta) with the substrate FKPF (cyan). Subsite S-4 in chymotrypsin is crowded with bulky tryptopan residues, while cercarial protease lacks the large loop which hangs over the S-4 site. The P-3 subsite could accommodate a large charged side chain as shown and not significantly affect inhibitor binding.

ductive binding of tryptophan to the P-4 pocket in the tetrapeptide substrate analogues.

*Effect of Peptide Inhibitors on Cercarial Penetration of Human Skin.* On the basis of observations from structural modeling of the cercarial protease, and the subsequent data from assays of synthetic peptide substrates and inhibitors, we tested a series of chloromethyl ketone and boronic acid derivatized tetrapeptides against live cercariae in an in vitro model of human skin penetration. When cercariae were introduced into a chamber containing human skin warmed to 37 °C, they were attracted to the skin surface and stimulated to invade by lipid on the surface of skin. Within 1 h cercariae could be observed entering the dermal extracellular matrix (Figure 5, top). In the concentration range of peptide inhibitors tested (20 $\mu$M to 1 mM), boronic acid derivatized peptides and chloromethyl ketone derivatized peptides had no effect on cercarial motility, movement toward skin, or viability. In fact, even if inhibited from invading, cercariae still swim to and attach to the surface of skin by mucous secretions in the presence of these inhibitors (Figure 5, bottom).

Tetrapeptide inhibitors with large hydrophobic P-1 side chains (Leu, Phe), predicted by the model to be favored by the enzyme, were effective in inhibiting cercarial penetration of skin at 50 $\mu$M (Figure 6). In contrast, a chloromethyl ketone derivatized tetrapeptide, differing only in having a small hydrophobic amino acid at P-1 (Ala), had no significant effect on cercarial invasion.

Unfortunately, the solubility of these hydrophobic chloromethyl ketone peptides is limited at 50 $\mu$M. We therefore examined the possibility of using the model to design an in-

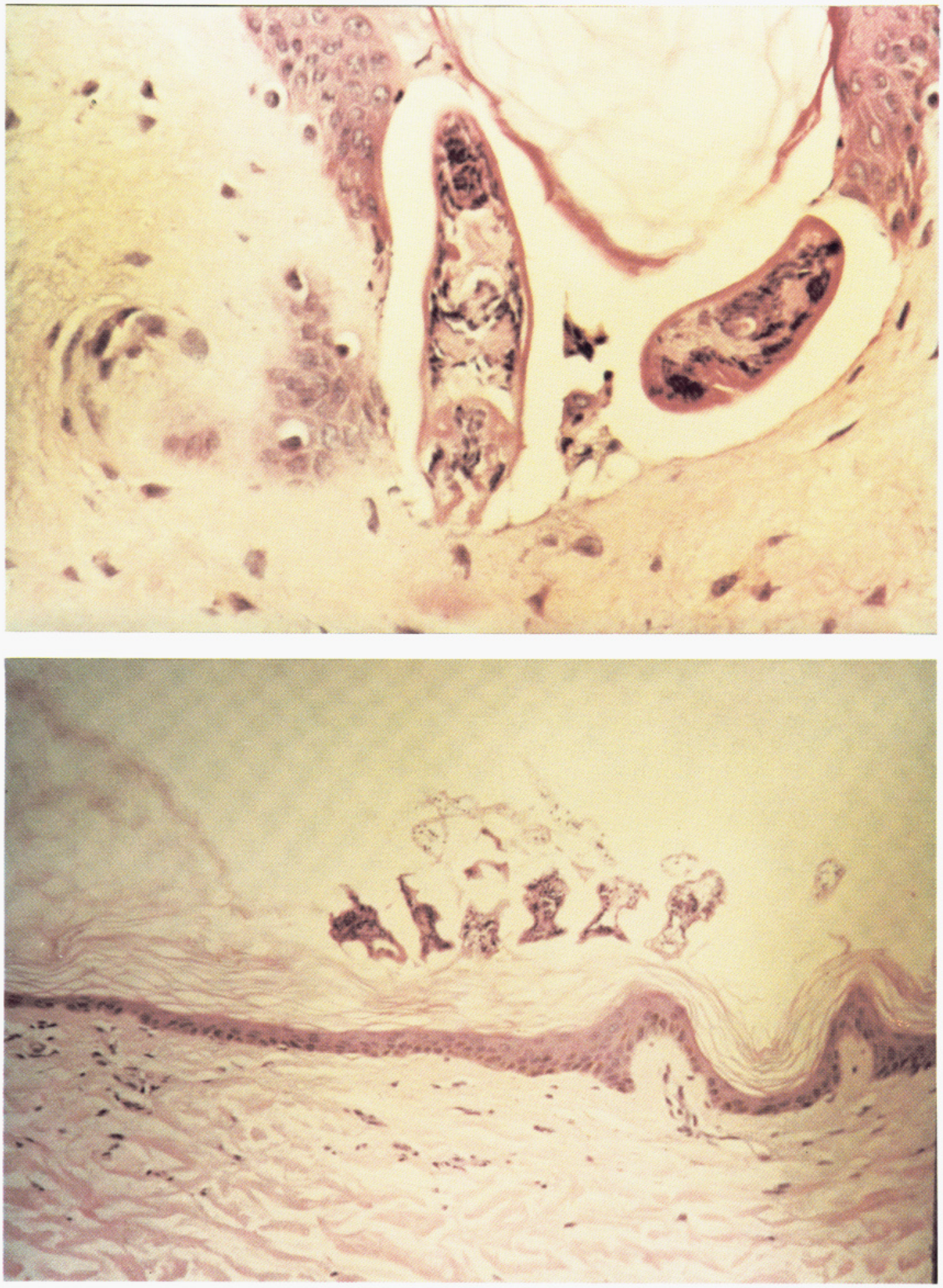


FIGURE 5: (Top) Cercaria invading the epidermis and dermis in the assay described under Experimental Procedures. Cross section of human skin at 250× magnification; hematoxylin and eosin strain. (Bottom) Cercariae bound to the surface of the skin but not invading in the presence of the 50 $\mu$M MeO-Suc-AAPL-CMK inhibitor.

hibitor with increased solubility. The P-3 side chain of the modeled inhibitor points away from the protein in the direction of the solvent, suggesting that any amino acid could be sterically accommodated at this position. A hydrophilic amino acid should increase inhibitor solubility. Table III shows that a synthetic peptide with lysine at P-3 was only slightly less effective at inhibiting the protease ($k_3/K_i$ of 563 $M^{-1}$ $s^{-1}$ versus 798 $M^{-1}$ $s^{-1}$ for the "parent" peptide), but its aqueous solubility increased from 50 to 200 $\mu$M. At this higher concentration, 85% of the cercariae were inhibited from invading skin.

Boronic acid derivatized peptides have greater solubility relative to the corresponding chloromethyl ketone peptides. We therefore also tested a boronic acid derivatized peptide analogous to the chloromethyl ketone derivatized Ala-Ala-Pro-Phe sequence. The inhibitory capacity (measured as $IC_{50}$ because the chloromethyl ketone inhibitor is irreversible while

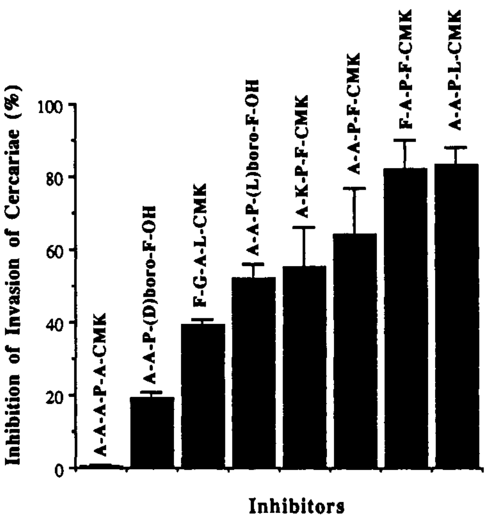


FIGURE 6: Comparison of effectiveness of peptide inhibitors in preventing skin invasion by cercariae at 50 $\mu$M.

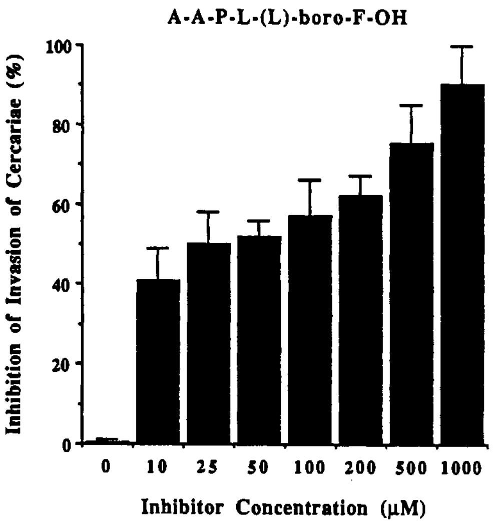


FIGURE 7: Inhibition of cercariae invading skin by MeO-Suc-Ala-Ala-Pro-L-boro-Phe-OH.

Table IV: Comparison of Chloromethyl Ketone Inhibitors to Boronic Acid Inhibitors versus Cercarial Proteinase

| inhibitor | $IC_{50}$ at 20 min (nM) |
|---|---|
| MeO-Suc-Ala-Ala-Pro-Leu-CMK | 36 |
| MeO-Suc-Ala-Ala-Pro-Phe-CMK | 88 |
| MeO-Suc-Ala-Ala-Pro-L-boro-Phe-OH | 136 |
| MeO-Suc-Ala-Ala-Pro-D-boro-Phe-OH | $>10^4$ |
| MeO-Suc-Ala-Ala-Pro-Ala-CMK | $>10^5$ |

the boronic acid inhibitor is not) of the two peptides against purified enzyme was similar (Table IV). The more soluble boronic acid inhibitor was then used to confirm inhibition of skin penetration at varying concentrations of inhibitor (Figure 7). Stereoisomers of the same peptide were also evaluated. At 50 $\mu$M the L stereoisomer inhibited 50% of cercariae from invading skin while the D stereoisomer inhibited less than 20% (Figure 6). This was consistent with the preference of the protease for L amino acids (Table III).

## Discussion

We have proposed a three-dimensional computer model of a parasite serine protease on the basis of the primary sequence of the enzyme and its homology with other serine proteases. The model structure is consistent with proteins of known structure with respect to packing volume, side-chain accessibility profile, and local conformation. From the perspective of enzyme function, the catalytic triad sits at the base of the binding cleft suitable for polypeptide substrates. While some errors in side-chain conformation are expected, the identifiable homology with a structurally well-characterized family of enzymes makes the cercarial protease model useful, especially in the region around the active site.

Using the model, we made specific predictions as to which peptide substrates or inhibitors would be optimal for this enzyme. These predictions were confirmed by examining a set of synthetic peptide substrates with different leaving groups and a set of chloromethyl ketone and boronic acid derivatized peptide inhibitors. While the binding specificity for large hydrophobic residues at P-1 was not unexpected, on the basis of data from a small group of synthetic peptides assayed previously (McKerrow, 1985a), it served as an initial test of the accuracy of the model. The model then made three unexpected and useful predictions. First, Pro 188 was identified as a likely candidate for blocking $\beta$-branched residues at P-1. We plan to test this prediction in future studies by site-directed mutagenesis. Second, a site for addition of a lysine to increase solubility for tetrapeptide inhibitors was predicted accurately. Finally, an unexpected hydrophobic binding site at P-4 was identified and confirmed by assays with synthetic substrates and inhibitors. None of these predictions would have been efficiently or intuitively made without the model.

The interactive interplay of modifying and testing synthetic inhibitors and modifying the molecular model helped to refine both the model and the inhibitors before they were used in a biological assay. For example, the position of a negatively charged side chain near the S-1 pocket was more accurately modeled after it was determined that lysine at P-1 was unfavorable as a substrate (Table II). Inhibitors found to be optimal for the enzyme, as predicted by the computer model, and confirmed by assays with purified enzyme, were also optimal at inhibiting invasion of skin by cercariae.

The success of applying algorithms for prediction of three-dimensional protein structure from primary sequence provides further evidence that this is an approach of great potential for design of enzyme inhibitors, not only as potential pharmacologic agents but also for analysis of the function of biologically interesting enzymes. Serine proteases play important roles not only in infectious diseases but also in hemostasis, embryonic development, and the immune response (Neurath, 1986). It is now theoretically possible to amplify and isolate genes coding for any eukaryotic serine or cysteine protease using oligonucleotide primers based upon conserved structural motifs and the polymerase chain reaction (Eakin et al., 1990; Sakanari et al., 1989). Sequence data on the majority of the coding regions of these enzymes can thus be obtained even when organisms or cells producing the enzymes are in relatively short supply. The primary sequence predicted from these amplified gene fragments has been shown to be quite accurate (Sakanari et al., 1989) and, as demonstrated here, could be used to model three-dimensional structures of the enzymes. The amplified gene fragments can also be used as homologous probes to isolate and sequence full-length genes or cDNAs. Coupled with expression systems that can produce active recombinant proteases (Graf et al., 1987), new inhibitors can be predicted, synthesized, and tested without the need for labor-intensive protein purification or abundant natural sources of the enzyme.

## Acknowledgments

We thank David Agard, James Powers, Charles Craik, and

David Rasnick for their helpful advice on enzyme kinetics. We also thank Charles Kettner, David Agard, and James Powers for providing synthetic peptide derivatives and Ramona Soto for preparing the manuscript.

## References

Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F., & Weng, J. (1987) in *Crystallographic Databases—Information Content, Software Systems, Scientific Applications* (Allen, F. H., Bergeroff, G., & Sievers, R., Eds.) pp 107–132, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977) *J. Mol. Biol. 112*, 535–542.

Blundell, T. L., Cooper, J., Foundling, S. I., Jones, D. M., Atrash, B., & Szelke, M. (1987a) *Biochemistry 26*, 5585–5590.

Blundell, T., Sibanda, B. L., Sternberg, M. J., & Thornton, J. M. (1987b) *Nature 326*, 347–352.

Bode, W., Chen, Z., Bartels, K., Kutzbach, C., Schmidt-Kastner, G., & Bartunik, H. (1983) *J. Mol. Biol. 164*, 237–282.

Bone, R., Shenvi, A. B., Kettner, C. A., & Agard, D. A. (1987) *Biochemistry 26*, 7609–7614.

Bone, R., Silen, J., & Agard, D. (1989) *Nature 339*, 191–195.

Cherfas, J. (1989) *Science 246*, 1242–1243.

Chothia, C., & Lesk, A. (1986) *EMBO J. 5*, 823–826.

Clegg, J. A., & Smithers, S. R. (1972) *Int. J. Parasitol. 2*, 79–98.

Cline, B. L. (1989) in *Tropical Medicine and Parasitology* (Goldsmith, R., & Heyneman, D., Eds.) pp 434–458, Appleton and Lange, San Mateo, CA.

Craik, C. S., Largman, C., Fletcher, T., Roczniak, S., Barr, P. J., Fletterick, R., & Rutter, W. J. (1985) *Science 228*, 291–297.

Dixon, M. (1953) *Biochem. J. 55*, 170–171.

Dorovska, V., Varfolomeyev, S., Kazanskaya, N., Klyosov, A., & Martinek, K. (1972) *FEBS Lett. 23*, 122–124.

Eakin, A. E., Bouvier, J., Sakanari, J. A., Craik, C. S., & McKerrow, J. H. (1990) *Mol. Biochem. Parasitol. 39*, 1–8.

Ferrin, T. E., Huang, C. C., Jarvis, L. E., & Langridge, R. (1988) *J. Mol. Graphics 6*, 13–37.

Freudenreich, S. C., Samana, J.-P., & Biellmann, J.-F. (1984) *J. Am. Chem. Soc. 106*, 3344–3353.

Graf, L., Craik, C. S., Patthy, A., Roczniak, S., Fletterick, R., & Rutter, W. J. (1987) *Biochemistry 26*, 2616–2623.

Greer, J. (1981) *J. Mol. Biol. 153*, 1027–1042.

Greer, J. (1990) *Proteins: Struct., Funct., Genet. 7*, 317–334.

Greer, J., Mollison, K. W., Carter, G. W., & Zuiderweg, E. R. (1989) *Prog. Clin. Biol. Res. 289*, 385–397.

James, M. N. G., Delbaere, L. T. J., & Brayer, G. D. (1978) *Can. J. Biochem. 56*, 396–402.

Jarvis, L., Huang, C., Ferrin, T., & Langridge, R. (1988) *J. Mol. Graphics 6*, 2–27.

Jencks, W. P. (1987) in *Catalysis in Chemistry and Enzymology*, Dover ed., pp 291–296, Dover Publications, Inc., New York.

Kitz, R., & Wilson, I. B. (1962) *J. Biol. Chem. 237*, 3245.

Knight, P. (1990) *Bio/Technology 8*, 105–107.

Lineweaver, H., & Burk, D. (1934) *J. Am. Chem. Soc. 56*, 658.

McKerrow, J. H., Pino-Heiss, S., Lindquist, R., & Werb, Z. (1985a) *J. Biol. Chem. 260*, 3703–3707.

McKerrow, J. H., Jones, P., Sage, H., & Pino-Heiss, S. (1985b) *Biochem. J. 231*, 47–51.

McKerrow, J. H., Sakanari, J. A., Brown, M., Brindley, P., Railey, J. F., Weiss, N., & Resnick, S. D. (1989) in *Models in Dermatology* (Maibach, H., & Lowe, N. J., Eds.) Vol. 4, pp 276–284, Karger, Basel, Switzerland.

McQuade, T. J., Tomasselli, A. G., Liu, L., Karacostas, V., Moss, B., Sawyer, T. K., Heinrikson, R. L., & Tarpley, W. G. (1990) *Science 247*, 454–456.

Meyer, E., Cole, G., Radahakrishnan, R., & Epp, O. (1988) *Acta Crystallogr., Sect. B 44*, 26–38.

Neurath, H. (1986) *J. Cell. Biochem. 32*, 35–49.

Newport, G. R., McKerrow, J. H., Hedstrom, R., Petitt, M., McGarrigle, L., Barr, P. J., & Agabian, N. (1988) *J. Biol. Chem. 263*, 13179–13184.

Novotny, J., Rashin, A. A., & Bruccoleri, R. E. (1988) *Proteins 4*, 19–30.

Ponder, J. W., & Richards, F. M. (1987) *J. Mol. Biol. 193*, 775–791.

Read, R. J., & James, M. N. G. (1984) *Biochemistry 23*, 6570–6575.

Remington, S. J., Woodbury, R. G., Reynolds, R. A., Matthews, B. W., & Neurath, H. (1988) *Biochemistry 27*, 8097–8105.

Richardson, J. S. (1981) *Adv. Protein Chem. 34*, 167–339.

Ripka, W. C., Sipio, W. J., & Blaney, J. M. (1987) *Lect. Heterocycl. Chem. 9*, 95–104.

Roberts, N. A., Martin, J. A., Kinchington, D., Broadhurst, A. V., Sham, H. L., Bolis, G., Stein, H. H., Fesik, S. W., Marcotte, P. A., Plattner, J. J., Rempel, C. A., & Greer, J. (1988) *J. Med. Chem. 31*, 284–295.

Sakanari, J. A., Staunton, C. E., Eakin, A. E., Craik, C. S., & McKerrow, J. H. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 4863–4867.

Stein, R. L., Strimpler, A. M., Hori, H., & Powers, J. C. (1987) *Biochemistry 26*, 1301–1305.

Stirewalt, M. A. (1974) *Adv. Parasitol. 12*, 115–180.

Thornton, J. M. (1981) *J. Mol. Biol. 151*, 261–287.

Tsukada, H., & Blow, D. M. (1985) *J. Mol. Biol. 184*, 703–711.

Walter, J., Steigemann, W., Singh, T. P., Bartunik, H., Bode, W., & Huber, R. (1982) *Acta Crystallogr., Sect. B 38*, 1462–1472.