

Modeling of protease I collagenolytic enzyme from the fiddler crab *Uca pugilator*

B. Arnoux^{a,*}, A. Lecroisey^b and A. Ducruix^a

^aInstitut de Chimie des Substances Naturelles, C.N.R.S., F-91198 Gif-sur-Yvette Cedex, France

^bUnité de Chimie des Protéines, Institut Pasteur, 28 rue du Docteur Roux, F-75724 Paris Cedex, France

Received 11 July 1989

Accepted 30 October 1989

Key words: Protease; Modeling; Chymotrypsin family; Collagenolytic enzyme

SUMMARY

Collagenolytic protease I from the fiddler crab *Uca pugilator* belongs to the serine proteases of the trypsin family. A graphic molecular model was built using information from sequences and X-ray structures of four homologous proteins which were superimposed to define structurally conserved regions. Protease I sequence was aligned, with sequences of the model proteins, without permitting any deletion or insertion in these regions. Elastase α -carbon chain was selected as a template molecule. For the structurally variable regions, fragments of the four homologous proteins which were 'closest' in sequence were selected. Intra-molecular steric hindrance, that resulted from the substitution of the residues of the templates by protease I residues, was corrected by adjustment of the side-chain conformational angles. The model was then optimized by energy minimization. The primary specificity pocket in the model of collagenolytic protease I predicts a substrate preference for both P1 hydrophobic and positively charged residues which is in agreement with the biochemical observations. As soybean trypsin inhibitor (STI) is known to inhibit collagenolytic protease I, a tentative model of the complex was constructed and possibilities of interaction examined.

INTRODUCTION

Collagenases are enzymes which cleave the native collagen in its helical part under physiological conditions of pH, temperature and ionic strength [1]. Most collagenases belong to the metalloenzyme family whereas protease I from the fiddler crab *Uca pugilator* is a member of the collagenolytic enzymes related to the trypsin family. This group, which consists of collagenases with digestive rather than morphogenic functions, includes among others a second protease from the same arthropod, protease II [2] and a protease from fly larvae, the collagenase from *Hypoderma lineatum* [3].

Crab protease I (226 amino acids, m.w. 23 505) degrades collagen producing multiple cleavages

*To whom correspondence should be addressed.

in the triple helix [4]. However, the major early cleavages occur at a 3/4 1/4 locus, resulting in fragments resembling the 3/4 and 1/4 fragments produced by the action of mammalian and *H. lineatum* collagenases [5]. In contrast to mammalian collagenases, protease I [6] and II [2] and *H. lineatum* collagenase exhibit a good general proteolytic activity. Protease I displays a broader peptide bond specificity than either trypsin or chymotrypsin on non-collageneous substrates, cleaving on the carbonyl terminal side of residues with both positively and negatively charged side chains as well as hydrophobic side chains. The enzyme is very efficiently inhibited by chymostatin and soybean trypsin inhibitor (STI) whereas chloromethyl ketones which inhibit trypsin and chymotrypsin are without effect on its activity.

Sequences of both crab protease I [7] and *H. lineatum* collagenase [8] are published. The *H. lineatum* enzyme was crystallized [9] and its 3D structure is under refinement in our laboratory. The ability of both protease I and *H. lineatum* collagenase to specifically cleave the collagen molecule must impart conformational homologies in their substrate binding site as compared with pancreatic proteases, and structural studies would show if these enzymes effectively display more features in common. We therefore undertook modeling of the 3D structure of protease I. It was carried out by means of computer and graphical methods and based on the hypothesis that homologous serine proteases of the trypsin family exhibit similar tertiary structures [10] where β -sheets are essentially conserved, whereas loops have to be redesigned.

METHODS

The coordinates of the structures of selected serine proteases were taken from the Brookhaven Protein Data Bank [11]. Modeling was achieved by means of several programs. MANOSK [12] was used for most of the work (building the α -carbon chain and substituting the appropriate sequence). FRODO [13] was used for side-chain adjustments and CHARMM [14] for refinement of the model. Calculations were performed on a microVaxII computer system linked to a PS390 Evans & Sutherland graphic display. CHARMM was run on a MATRA MD570.

MODELING OF THE PROTEASE I STRUCTURE

(1) Sequence alignment

The first step in modeling was to align sequences of protease I with other serine proteases of known 3D structure. In spite of a low percentage of homology between their sequences, serine proteases show striking similarities in their 3D structures which are mainly folded in β -sheets with a protruding C-terminal α -helix. In a classical paper, Greer [15] proposed the definition of 'structurally conserved regions' (SCR). This work was based on the X-ray structures of trypsin, chymotrypsin and elastase in which conserved regions include approximately 60% of the structures. We redefined the SCR by adding the structure of kallikrein [16]. The four structures were superimposed by least square refinement on C α -carbons starting with the three atoms of the active site: His⁵⁷, Asp¹⁰², Ser¹⁹⁵. Structures were compared two at a time. The percentage of sequence identity between protease I and each model sequence is 32%. Pairs of C α -atoms were used in the next refinement only if they were not more than 1.2 Å apart. Successive iterations led to SCR common to the four proteins. On the base that no deletion or insertion were admitted in the SCR, the se-

TABLE 1
ALIGNMENT OF PROTEASE I WITH ELASTASE [17], CHYMOTRYPSIN [17], TRYPSIN [17] AND KALLIK-
REIN [18]

	20		30		40		50		60
ELA	VVG	GTEA	QRNSW	PSQISLQ	YRSGSSWA	HTCGGT	LIRQNWVMTAAHC		VDRE
CHT	IVN	GEEA	VPGSW	PWQVSLQ	DKT---GF	HFCGGS	LINENWVVTAACH		GVT-
TRP	IVG	GYTC	GANTV	PYQVSLN	S-----GY	HFCGGS	LINSQWVVSAAHC		YKS-
KAL	IIG	GREC	EKN	SH	PWQVAIY	HYS----S	FQCGGV	LVNPKWVLTAACH	KND-
UCA	IVG	GVEA	VPNSW	PHQAALF	IDD----M	YFCGGS	LISPEWILTAACH		MDGA

	70		80		90		100		
ELA	LT	FRVVVGEHN	LNQNNGT	EQYVGVQKIVVHPYW	NTDDVAA--		GYDIALLR		
CHT	TS	DVVVAGEFD	QGS	SSEK	IQKLKIAKVFKN	SKY	NSLTI----	NNDITLLK	
TRP	-G	IQVRLGGDN	INV	VEGN	QQFISASKSIVHPSY		NSNTL----	NNDIMLIK	
KAL	-N	YEVGWL	RHN	LFEN	ENT	AQFFGV	TADFPHPGF	NLSADGKDY	SHDLMLLR
UCA	GF	VDVVLGAHN	IRE	DEAT	QVTIQST	DFTVHENY	NSFVI----		SNDIAVVR

	110		120		130		140		150
ELA	LAQSVTL	NS	YVQLGVLP	RAGTILANNS	PCYITGWGLT	RTN-GQL--		AQT	
CHT	LSTAASF	SQ	TVSAVCLP	SASDDFAAGT	TCVTTGWGLT	TY--ANT--		PDR	
TRP	LKSAASL	NS	RVASISLP	T--SCASAGT	QCLISGWGNT	KSSGTSY--		PDV	
KAL	LQSPAKI	TD	AVKVLELP	T--QEPELGS	TCQASGWGSI	EPGPDDFEF		PDE	
UCA	LPVPVTL	TA	AIATVGLP	S--TDVGVGT	VVTPTGWGLP	SDSALGI--		SDV	

	160		170		180		190		200	
ELA	LQQA	YLP	TVDYAICSSSYWG	STVKNS	MVCAG	GDG-V	RSGCQGD	SGGPLHC		
CHT	LQQA	SLPL	LSNTNCKK--	YWG	KINDA	MICAG	A--SG	VSSCMGD	SGGPLVC	
TRP	LKCL	KAPIL	SDSSCKS--	AYPGQ	ITSN	MFCAG	YLQGG	KDSCQGD	SGGPVVC	
KAL	IQCV	QLTLL	QNTFCAD--	AHPDK	VTES	MLCAG	YLP	GG	KDTCMGD	SGGPLIC
UCA	MRQV	DVPIM	NSADCD	A--VY-G	IVTDG	NICID	STG-G	KGTCGD	SGGPLNY	

	210		220		230		240
ELA	LVNGQYAV	HGVTSFV	SRLG	C	NVTRK	PTVFTRVSAYISWINNV	IASN
CHT	KKNGAWTL	VGIVSWG	SS-T	C	STS-T	PGVYARVTALVNWVQQT	LAAN
TRP	S-----GKL	QGIVSWG	SG--	C	AQANK	PGVYTKVCNYVSWIKQT	IASN
KAL	N-----GMW	QGITSWG	HTP-	C	GSANK	PSIYTKLIFYLDWIDDT	ITENP
UCA	D-----GLT	YGITSFG	AAAG	C	EAG-Y	PDAFTRVTYFLDWIQTQ	TGITP

Sequences in boxes represent the structurally conserved regions (SCR). Residues are numbered according to the chymotrypsinogen numbering system.

quence of *U. pugilator* protease I was then aligned with the others (Table 1). From this alignment, it was decided to choose elastase as a template and so elastase SCR were the base for building the secondary structure of protease I.

(2) Loop building

In building loops between β -sheets defined by SCR, the positions of the fragments were adjusted, relative to the elastase skeleton by the use of splicers (2 residues) at each extremity (Fig. 1). Residues included in splicers on each side of the loop were fused with respect to stereochemical evidences. Two different situations were encountered:

- The loop in protease I had the same length as one or more of the four reference molecules.
- If only one reference molecule was concerned, the loop was conserved as it was. That was the

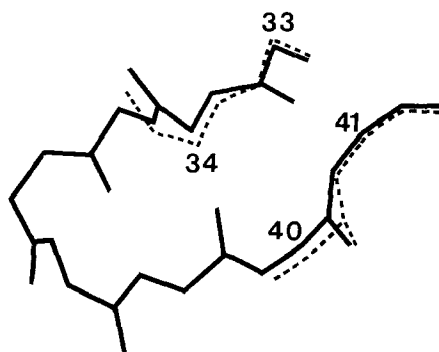


Fig. 1. Construction of the loop 35–39. α -chains of residues 33,34 and 40,41 from elastase and kallikrein were superimposed. Residues 33 and 40 from elastase were respectively linked up to residues 34 and 39 from kallikrein.

case in loops 34–39, 59–64, 145–151, 185–188, 221–225. (ii) If two or more reference molecules had the same length, the chain presenting the maximum homology by Garnier et al. [19] and Dayhoff et al. [20] methods was chosen.

TABLE 2

HYBRID MOLECULE: \square REGIONS CONSTRUCTED FROM ELASTASE SCR; — REGIONS CONSTRUCTED FROM DIFFERENT TEMPLATE LOOPS

ELA 16 — 22	23 — CHT 29	ELA 30 — 33	34 — KAL 39	ELA 40 — 58
59 — ELA 64	ELA 65 — 72	73 — ELA 79	ELA 80 — 91	92 — CHT 101
ELA 102 — 114	115 — ELA 116	ELA 117 — 123	124 — KAL 135	ELA 136 — 144
145 — TRP 151	ELA 152 — 158	159 — TRP 180	ELA 181 — 184	185 — ELA 188
ELA 189 — 199	200 — TRP 209	ELA 210 — 216	217 — ELA 219	ELA 220
221 — CHT 225	ELA 226 — 240	241 — KAL 246		

(b) In the case of loop 159–179 of protease I, the sequence was shorter than in any of the four reference molecules. The corresponding trypsin loop was chosen because it implied deletion of only one residue and best sequence homology with protease I. Proline¹⁷³ was deleted as suggested by sequence alignment and structural aspects. The FRODO program was used to delete the residue and reconstruct the peptide bond.

Table 2 shows parts of the sequence from protease I which was constructed from elastase SCR and parts which were built from different loops. Figure 2 shows the 3D skeleton of the hybrid molecule corresponding to Table 2. It is to be noted that all insertions or deletions are located on the surface of the protease I structure.

(3) *Substituting the protease I sequence*

At this stage, the skeleton (main chain) belongs to protease I but the side chains were still those of the initial structures. The subroutine SUBSTITUTION of MANOSK was used to substitute all the side chains according to the protease I sequence as given in Table 1. Whenever possible, the torsion angles χ_1, χ_2 from templates were kept to maintain the side-chain orientation.

All intramolecular distances were calculated to point out short van der Waals contacts. With the help of the FRODO program, side chains in contact were moved away, using the information given by the orientation of the side chains of the four reference molecules and energetically imposed restriction on χ_1 given by the most commonly observed conformations.

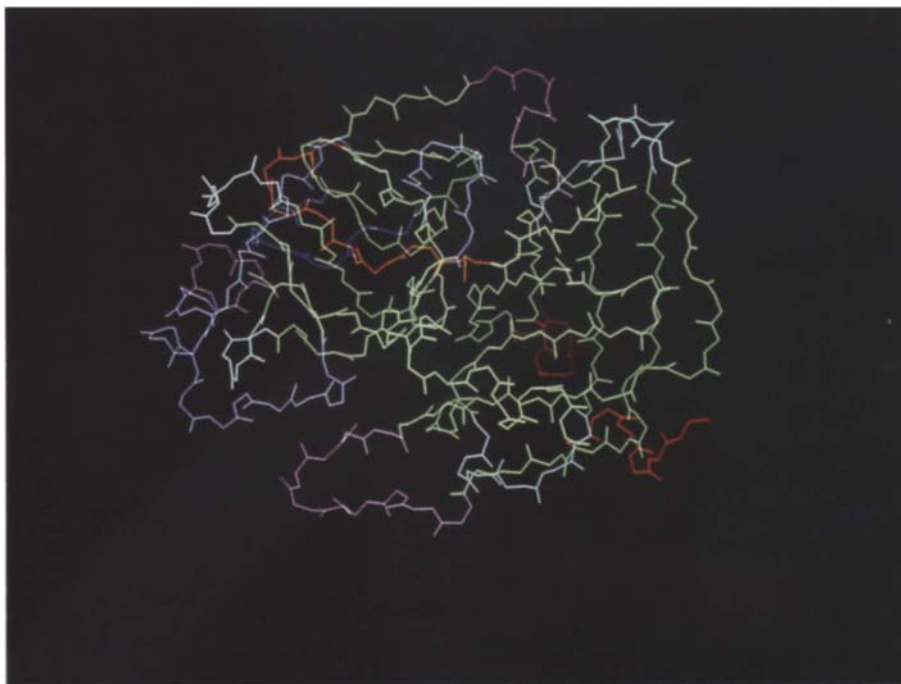


Fig. 2. Constructed hybrid molecule. Green: regions constructed from elastase SCR; light blue: regions from elastase loops; red: regions from kallikrein loops; fuchsia: regions from chymotrypsin loops; violet: regions from trypsin loops.

TABLE 3
RESULTS OF THE ENERGY MINIMIZATION

	Total energy	Total van der Waals energy	Van der Waals repul- sive energy	Bonds energy	Angles energy	Dihedrals energy	rms deriv.
Starting (kcal/mol)	- 5 176.29	1 599.66	5 024.38	959.78	898.53	360.91	77.90
Final (kcal/mol)	-16 933.53	-1 185.55	463.69	206.15	816.39	524.15	1.46
Number of atoms	1 967						
rms shift C α (Å)	0.79						
rms shift all atoms (Å)	0.98						

Most of the collisions could be avoided but short contacts remained. To regularize the model coordinates, the CHARMM program was used. Constraints were introduced on disulfide bonds, and on the active site (His⁵⁷, Asp¹⁰², Ser¹⁹⁵, Ser²¹⁴). The final energy values (van der Waals and total), obtained after 200 cycles of minimization, are given in Table 3. The values on van der Waals energy clearly indicate an important reduction in short contacts.

RESULTS AND DISCUSSION

A view of the refined model of protease I is presented in Fig. 3. The overall shape of protease I computed using spline functions [21] is represented in Fig. 4. The protein is very acidic (isoelectric point = 3.0) due to a low content of basic residues (4 Arg, 1 Lys, 4 His) as compared with acidic residues (20 Asp and 6 Glu). As expected, the majority of the charged side chains is located at the surface of the molecule pointing into the solvent. Many of them form or could potentially form ion-pair interactions, especially those buried or only partially exposed to the solvent.

The configuration of the binding pocket of protease I partially explains its specificity toward polypeptide and synthetic low-molecular-weight substrates. In the binding pocket, position 216 is occupied by a Gly, as in chymotrypsin and trypsin. This allows the presence of a bulky side chain at the P1 position of the substrate whereas in elastase, a valine obstructs the access of the S1 crevice. In chymotrypsin, residues Ser¹⁸⁹ and Gly²²⁶ allow for large hydrophobic residues at P1. Trypsin differs by having an Asp in position 189 which determines its specificity for positively charged side chains. Protease I combines both situations by having a Gly in 189 and an Asp in 226. The carboxylate-negative charge of Asp²²⁶ is nearly at the same location as the one of Asp¹⁸⁹ in trypsin. The enzyme indeed displays a broad specificity, cleaving on the carboxyl-terminal side of hydrophobic residues (Tyr, Phe, Leu, Ile) and glutamyl residues, as α -chymotrypsin does and positively charged residues (Lys, Arg) as trypsin does. The fact that no cleavage was observed after a tryptophan corresponds to a steric hindrance due to the presence of Asp²²⁶ which protrudes into the S1 crevice. As in chymotrypsin, preference for hydrophobic residues in P2 may be influenced by the presence of an Ile at position 99.

Serine proteases are known to bind polypeptidic substrates through different regions among which fragment 214–216. In protease I and elastase, this fragment is followed by a small loop (217–219) which is longer, compared to chymotrypsin (plus one) or trypsin (plus two residues).

The loop is located at the top of the binding pocket entrance (Fig. 3A) and is composed, in the case of protease I, of small hydrophobic residues (Ala-Ala-Ala-Gly). This sequence may represent additional subsites which may explain the broad primary specificity of this enzyme, as extended regions of interaction with substrate are known to favor a large specificity.

In order to evaluate the correctness of our model of protease I, we attempted to build a complex between protease I and STI which is known to give full inhibition [22], as opposed to other serine protease inhibitors such as bovine trypsin inhibitor (BPTI) or ovomucoid inhibitor (OMI). A partially refined X-ray structure of the complex trypsin/STI has been published [23]. The coordinates of the trypsin moiety were used to superimpose protease I with trypsin. The best fit was performed by using pairs of atoms belonging to the catalytic triad and fragments 40–42, 189–195, 214–215, which are implied in most of the polar interactions between trypsin and STI (Table 4A). Then the trypsin molecule was removed. A minor reorientation was applied to the side chain of Arg^{63'} (STI). No further attempt was made to refine the model of the complex. In this conformation,

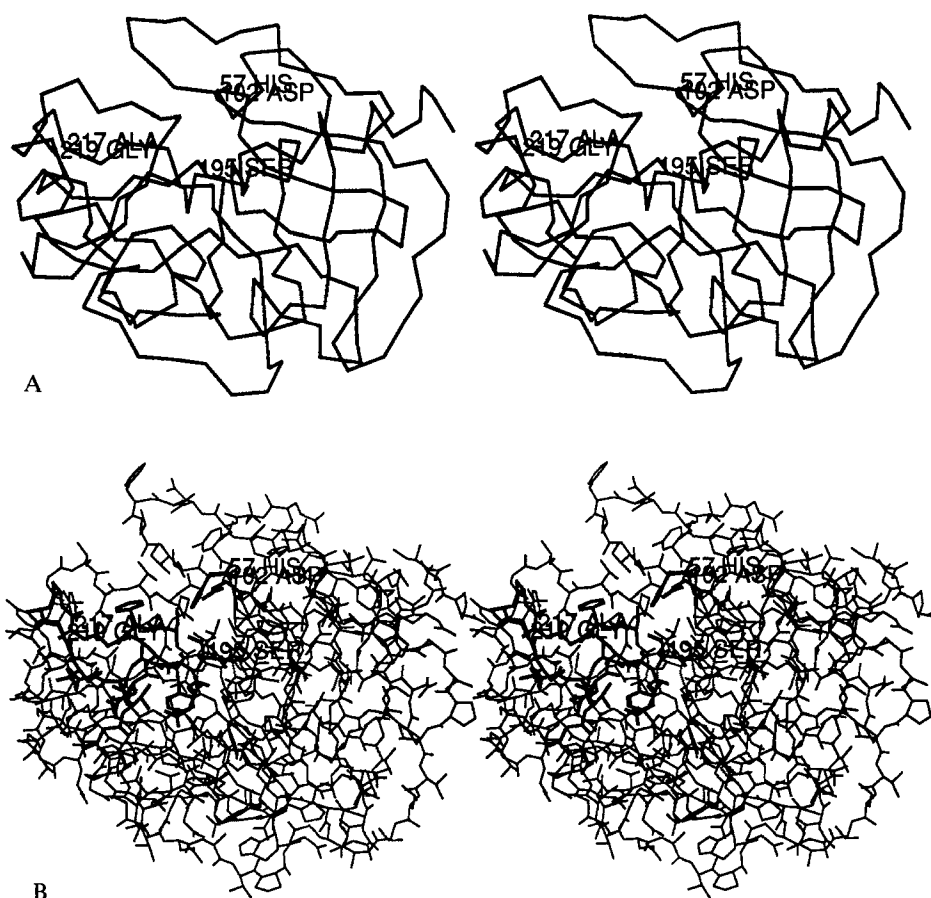


Fig. 3. Stereoview of the refined model of protease I. (A) C α -chain; (B) residues 189–194, 214–220 and 225–228 are indicated in bold.

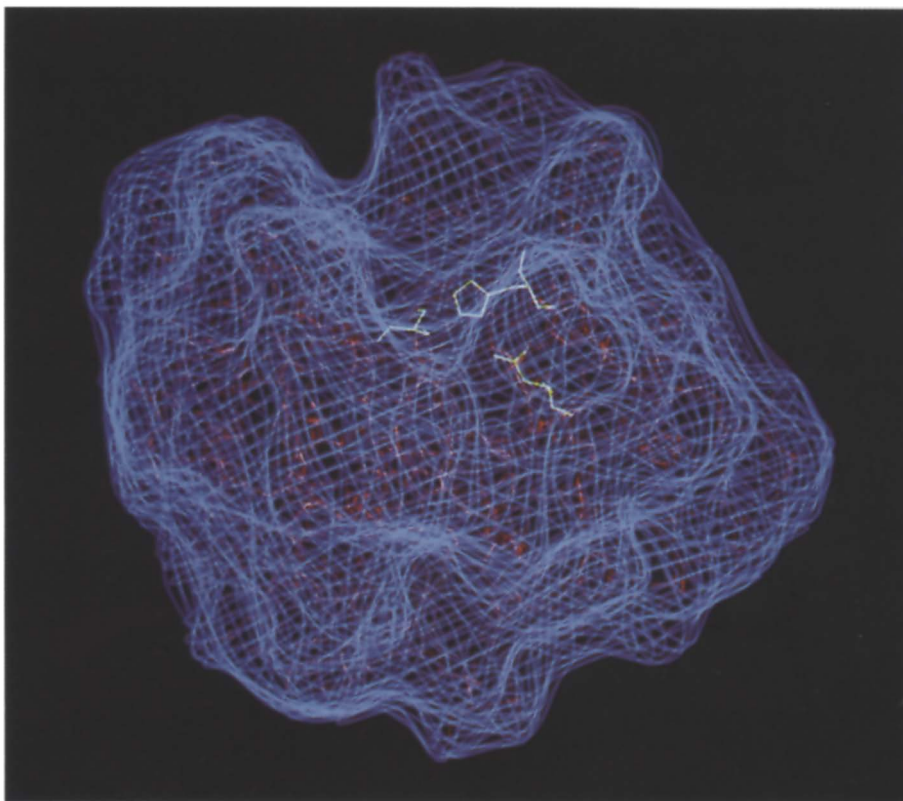


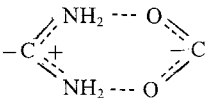
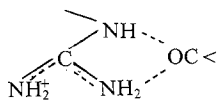
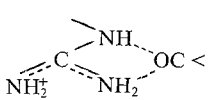
Fig. 4. Envelope of protease I computed using spline function [21]. Cleft of the active site is located on the top with catalytic triad in white.

Arg⁶³ (STI) is implied in a possible salt bridge with Asp²²⁶ of protease I as expected from the binding pocket conformation. Other possible interactions are given in Table 4B.

Comparison with the 3D structure of *H. lineatum* collagenase and other serine proteases will be published after refinement of the former. These studies may lead to a better understanding of the common features of the collagenolytic serine enzymes and of their specific action on collagen.

TABLE 4

(A) SELECTED INTERACTIONS BETWEEN STI AND TRYPSIN [22]; (B) POSSIBLE INTERACTIONS BETWEEN STI AND PROTEASE I

(A) STI-trypsin complex				(B) STI-protease I complex			
Involved residues		Involved groups		Involved residues		Involved groups	
STI	Trypsin	STI	Trypsin	STI	Protease I	STI	Protease I
Ser ⁶¹	Gly ²¹⁶	>CO	--- NH<	Ser ⁶¹	Gly ²¹⁶	>CO	--- NH<
Tyr ⁶²	Asn ⁹⁷	---OH	--- OCNH ₂ ---				
Arg ⁶³	Asp ¹⁸⁹			Arg ⁶³	Pro ²²⁵	---NH ₂ --- OC<	
		Gly ¹⁹³	>CO --- HN<			Asp ²²⁶	=NH ₂ ⁺ --- -OC-
		Ser ¹⁹⁵	>CO --- HN<			Gly ¹⁹³	>CO --- HN<
		Ser ²¹⁴	>NH --- OC<			Ser ¹⁹⁵	>CO --- HN<
Arg ⁶⁵	His ⁴⁰			Arg ⁶⁵	Tyr ⁴⁰		
		His ⁷¹	>NH --- OC<			His ⁷¹	>NH --- OC<

ACKNOWLEDGEMENTS

We thank D. Housset and J.L. Risler for constructive advice.

REFERENCES

- 1 Harper, E., Ann. Rev. Biochem., 49 (1980) 1063.
- 2 Grant, G.A., Sachettini, J.C. and Welgus, H.G., Biochemistry, 22 (1983) 354.
- 3 Lecroisey, A., Boulard, C. and Keil, B., Eur. J. Biochem., 101 (1979) 385.
- 4 Welgus, H.G., Grant, G.A., Jeffrey, J.J. and Eisen, A.Z., Biochemistry, 21 (1982) 5183.
- 5 Lecroisey, A. and Keil, B., Eur. J. Biochem., 151 (1985) 123.
- 6 Grant, A.G. and Eisen, A.Z., Biochemistry, 19 (1980) 6089.
- 7 Grant, G.A., Henderson, K.O., Eisen, A.Z. and Bradshaw, R.A., Biochemistry, 19 (1980) 4653.
- 8 Lecroisey, A., Gilles, A.M., de Wolf, A. and Keil, B., J. Biol. Chem., 262 (1987) 7546.
- 9 Ducruix, A., Arnoux, B., Pascard, C., Lecroisey, A. and Keil, B., J. Mol. Biol., 151 (1981) 327.
- 10 Steitz, T.A. and Shulmann, R.G., Annu. Rev. Biophys. Bioeng., 11 (1982) 419.
- 11 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
- 12 Cherfils, J., Vaney, M.C., Morize, I., Surcouf, E., Colloc'h, N. and Moron, J.P., J. Mol. Graphics, 6 (1988) 155.
- 13 Jones, T.A., J. Appl. Cryst., 11 (1978) 268.

- 14 Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M., *J. Comput. Chem.*, 4 (1983) 187.
- 15 Greer, J., *J. Mol. Biol.*, 153 (1981) 1027.
- 16 Bode, W., Chen, Z., Bartels, K., Kutzbach, C., Schmidt-Kastner, G. and Bartunik, H., *J. Mol. Biol.*, 164 (1983) 237.
- 17 Dayhoff, M.O. (1972) *Atlas of Protein Sequence and Structure*, Vol. 5, National Biomedical Research Foundation, Washington, DC.
- 18 Tschesche, H., Ehret, W., Godec, G., Hirschauer, C., Kutzbach, C., Schmidt-Kastner, G. and Fiedler, F., In Sicuteri, F., Back, N. and Haberland, G.L. (Eds.) *Kinins N: Pharmacodynamic and Biological Roles*, Plenum Press, New York, 1976, pp. 123–133.
- 19 Garnier, J., Osguthorpe, D.J. and Robson, B., *J. Mol. Biol.*, 120 (1978) 97.
- 20 Dayhoff, M.O., Barker, W.C. and Hunt, L.T., *Methods Enzymol.*, 91 (1983) 524.
- 21 Colloc'h, N. and Morion, J.P., In *Proceedings of the Molecular Graphics Society*, San Francisco, August, 1988.
- 22 Grant, G.A., Eisen, A.Z. and Bradshaw, R.A., *Methods Enzymol.*, 80 (1981) 722.
- 23 Sweet, R.M., Wright, H.T., Janin, J., Chothia, C.H. and Blow, D.M., *Biochemistry*, 13 (1974) 4212.