THE UNIVERSITY OF ALBERTA

X-ray Crystallographic Studies of Two Serine Proteases:α-lytic Protease and Tonin

by

Masao Fujinaga

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy

Department of Biochemistry

EDMONTON, ALBERTA

FALL 1986

ISBN 0-315-32473-2

Department of Biochemistry
University of Alberta
Edmonton, Alberta
T6G 2H7
June 6, 1986.

Academic Press Inc. (London) Limited
24-28 Oval Road
London NW1 7DX
England

Dear Sir:

I would like to have the permission to use the material in

"Refined Structure of Alpha-lytic Protease at 1.7 Angstrom Resolution : Analysis of Hydrogen Bonding and Solvent Structure"
by M. Fujinaga, L. T. J. Delbaere, G. D. Brayer, and M. N. G. James,
published in *Journal of Molecular Biology* 184: 479-502 (1985),

in my Ph.D. thesis, entitled
X-ray Crystallographic Studies of Two Serine Proteases: Alpha-lytic Protease and Tonin.

Thank you for your cooperation.

Yours truly,

Masao Fujinaga

PERMISSION GRANTED
provided:

Permission of author(s) obtained;
The material to be used has appeared in our publication without credit or acknowledgement to another source;
Proper credit is given to our publication.

Date 16 6 86 By J. Walls

Rights and Permissions
ACADEMIC PRESS INC. (LONDON) LTD.
London, England

Co-author Permission Form

Permission is hereby granted to Masao Fujinaga to use material from

"Refined Structure of Alpha-lytic Protease at 1.7 Angstrom Resolution : Analysis of Hydrogen Bonding and Solvent Structure"

by M. Fujinaga, L. T. J. Delbaere, G. D. Brayer, and M. N. G. James,

published in *Journal of Molecular Biology* 184: 479-502 (1985).

in his thesis, entitled

X-ray Crystallographic Studies of Two Serine Proteases: Alpha-lytic Protease and Tonin.

Signed L. T. J. Delbaere

Name L. T. J. DELBAERE

Date 8 July '86

Co-author Permission Form

Permission is hereby granted to Masao Fujinaga to use material from

"Refined Structure of Alpha-lytic Protease at 1.7 Angstrom Resolution : Analysis of Hydrogen Bonding and Solvent Structure"

by M. Fujinaga, L. T. J. Delbaere, G. D. Brayer, and M. N. G. James,

published in *Journal of Molecular Biology* **184**: 479-502 (1985),

in his thesis, entitled

X-ray Crystallographic Studies of Two Serine Proteases: Alpha-lytic Protease and Tonin.

Signed Michael James

Name M N G JAMES

Date Aug. 11 /86

Co-author Permission Form

Permission is hereby granted to Masao Fujinaga to use material from

"Refined Structure of Alpha-lytic Protease at 1.7 Angstrom Resolution : Analysis of Hydrogen Bonding and Solvent Structure"

by M. Fujinaga, L. T. J. Delbaere, G. D. Brayer, and M. N. G. James,

published in *Journal of Molecular Biology* 184: 479-502 (1985),

in his thesis, entitled

X-ray Crystallographic Studies of Two Serine Proteases: Alpha-lytic Protease and Tonin.

Signed Gary D. Brayer

Name Gary D. Brayer

Date July 17/86

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR Masao Fujinaga

TITLE OF THESIS X-ray Crystallographic Studies of Two Serine Proteases:α-lytic Protease and Tonin

DEGREE FOR WHICH THESIS WAS PRESENTED Doctor of Philosophy

YEAR THIS DEGREE GRANTED FALL 1986

Permission is hereby granted to THE UNIVERSITY OF ALBERTA LIBRARY to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(SIGNED) Masao Fujinaga

PERMANENT ADDRESS:

11315 - 36 Ave
Edmonton Alberta
T6J 0C5

DATED Aug. 19 1986

Et j'aimerai le bruit du vent dans le blé...

Le Petit Prince

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled X-ray Crystallographic Studies of Two Serine Proteases:α-lytic Protease and Tonin submitted by Masao Fujinaga in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Michael James
Supervisor

D. Richard Morgan

Michael G. Rossmann
External Examiner

Date Aug. 14, 1986

Abstract

The crystal and molecular structures of two serine proteases, α-lytic protease and rat submaxillary gland tonin, have been determined using x-ray diffraction methods. The crystal structure of α-lytic protease had been solved previously at 2.8Å resolution using multi-isomorphous replacement methods; the refinement of its structure at high resolution is described. The structure of tonin was solved using molecular replacement methods, and has subsequently been refined. In addition to these experimental studies, the use of a new translation function is discussed.

The structure of α-lytic protease, a serine protease produced by the bacterium *Lysobacter enzymogenes*, has been refined at 1.7Å resolution. The conventional R-factor is 0.131 for the 14996 reflections between 8 and 1.7Å resolution with $I \geq 2\sigma(I)$. The model consists of 1391 non-hydrogen protein atoms, 2 sulfate ions and 156 water molecules. The overall root-mean-square error in the coordinates is estimated to be about 0.14Å. The refined structure was compared with homologous enzymes, α-chymotrypsin, and *Streptomyces griseus* protease A and B. A new sequence numbering for the bacterial enzymes was derived based on the alignment of these structures. The comparison showed that the greatest structural homology is around the active site residues Asp102, His57 and Ser195 and that basic folding pathways are maintained despite pronounced chemical changes in the hydrophobic cores. A detailed analysis of the

hydrogen bonding in the structure was carried out and the parameters describing the interactions were tabulated. The analysis revealed the presence of close intra-residue interactions. Charged groups on the protein are either paired with a counter-ion or are exposed to the solvent. The positions of the bound solvents were carefully determined and the resulting structure indicated no significant second shell of solvent molecules.

The 3-dimensional structure of tonin has been determined and refined at 1.8Å resolution. Tonin is a mammalian serine protease that is capable of generating the vasoconstrictive agent, angiotensin II, directly from its precursor protein, angiotensinogen, a process that normally requires two enzymes, renin and angiotensin converting enzyme. Tonin is an enzyme that is closely related to kallikrein. The structure solution was by the molecular replacement method using as the search model, the structure of bovine trypsin. The refined model of tonin consists of 227 amino acid residues out of the 235 in the complete molecule, 149 water molecules, and one Zn^{2+} ion. The R-factor is 0.196 for the 14997 measured data between 8-1.8Å resolution with $I \geq \sigma(I)$. It is estimated that the overall r.m.s. error in the coordinates is about 0.3Å. The structure of tonin that has been determined is not in its active conformation, but one that has been perturbed by the binding of a Zn^{2+} ion in the active site. Zn^{2+} was included in the buffer to aid the crystallization. Nevertheless the structure of tonin that is described is for the most part similar to its native form as indicated by the close tertiary structural homology with kallikrein. The differences in the structures of the two enzymes are concentrated in several loop regions and are probably responsible for the differences in their reactivities and specificities.

One of the key steps in the structure solution of tonin was the development of a new translation function. It is one that uses a linear correlation coefficient between observed and calculated structure factor amplitudes to determine the correct position of an oriented molecule in the crystal cell. The method has been implemented in a program called BRUTE. In the course of solving other structures the program has incorporated useful features such as the ability to refine the orientation of the model and the inclusion of a fixed set of atoms. A 'standard' procedure for solving the translation problem has evolved from experience and examples of difficult structures solved using this program are given. These examples illustrate the type of structures that may be solved using molecular replacement and indicate the limitations of the method.

Acknowledgement

I chose to work in this laboratory because of the people in it and not so much because of the science, and I have never regretted making that decision. They have given me warmth and friendship and I have grown much in their presence. The atmosphere of the department has also been good and I would like to thank Laura Frost in particular, for being a big sister to me and for the words of encouragement when things were bad.

On the scientific side, Mike James has been a perfect supervisor for me, continually trying to get some enthusiasm into me while being patient with my stubbornness. Randy Read has been a close collaborator on many projects and we learned many things together. The work on BRUTE described in Chapter IV was done with him. All the crystals that I used were grown by Koto Hayakawa. Without her it would have taken much longer to complete my thesis, since I seem to lack the ability and patience for growing good crystals. Other technical assistance has been provided by Marion Weber who helped me with paper electrophoresis, Mike Nattriss who ran some amino acid analyses, and Bernie Lemire who helped me do enzyme kinetics on the spectrophotometer.

This research has been supported by the Medical Research Council of Canada, through a grant to the MRC Group in Protein Structure and Function. I have been supported by MRC Studentships and have been given Research Allowances from the Alberta Heritage Foundation for Medical Research.

Table of Contents

List of Tables

List of Figures

List of Symbols and Abbreviations

a, b, c	unit cell axes
B	thermal motion parameter $= 8\pi^2\overline{U^2}$ where $\overline{U^2}$ is the mean square displacement of the atomic position
d_{min}	minimum interplanar spacing of diffraction data
e	electron
$\|E\|$	normalized structure factor amplitude
$\|E_c\|$	normalized calculated structure factor amplitude
$\|E_o\|$	normalized observed structure factor amplitude
f	atomic scattering factor
F_c	calculated structure factor
$\|F_o\|$	observed structure factor amplitude
I	intensity
k_{cat}	first-order rate constant for the chemical conversion of the enzyme-substrate complex to the enzyme product complex
MIR	Multiple Isomorphous Replacement
NMR	Nuclear Magnetic Resonance
OMTKY3	domain 3 of turkey ovomucoid inhibitor
PTI	Pancreatic Trypsin Inhibitor
R	standard crystallographic residual
r.m.s.	root-mean-square
SGPA	*Streptomyces griseus* Protease A
SGPB	*Streptomyces griseus* Protease B

SGT	*Streptomyces griseus* trypsin
α_C	calculated structure factor phase
$\sigma(x)$	standard deviation in x
σ_A	phase probability parameter
$<x>$	expected value of x
$\overline{x}$	mean value of x
z	number of molecules in the unit cell

I. Introduction to Serine Proteases:

Proteases are enzymes that catalyze the hydrolysis of peptide bonds. Serine proteases are characterized by a reactive serine residue. They are irreversibly inhibited by organophosphates such as diisopropylfluorophosphate and these reagents are often used to test whether a newly discovered protease belongs to the serine protease family. They are endopeptidases, that is they catalyze the hydrolysis of a peptide in the middle of the chain. The point of cleavage is determined by the sequence of amino acids about the scissile bond. In addition these enzymes can act as esterases, catalyzing the hydrolysis of ester linkages. Serine proteases occur in various organisms, from bacteria to mammals and are involved in a wide variety of functions ranging from indiscriminate degradation of proteins to delicate regulation of physiological processes.

One of the most well studied serine protease is chymotrypsin. It is a digestive enzyme secreted by the pancreas into the intestine as an inactive zymogen form, chymotrypsinogen. When it reaches the intestines a series of enzymatic cleavages transforms it into the active enzyme consisting of 241 amino acid residues making up three polypeptide chains. It has a primary specificity for large aromatic residues, cleaving on the C-terminal side of Trp, Phe, and Tyr (Hess, 1971).

In addition to the reactive serine, the pH dependence of the activity implicated a histidine residue in the enzymatic reaction (Bender & Killheffer, 1973). The ability of an active site directed specific alkylating agent, tosylphenylalanylchloromethylketone (TPCK), to inhibit α-chymotrypsin also supported the presence of a histidine (Schoellmann & Shaw, 1963; Powers, 1977). When the three-dimensional structure of chymotrypsin was determined (Matthews *et al*., 1967; Sigler *et al*., 1968), it showed that the histidine residue at position 57 in the sequence' was indeed hydrogen bonded to the reactive Ser195, confirming the earlier chemical studies. In addition, a buried aspartate residue at position 102 was found hydrogen bonded to the histidine (Blow *et al*., 1969). These three residues, Ser, His, and Asp, form what is known as the catalytic triad of serine proteases. Blow *et al*. (1969) proposed the 'charge relay mechanism' in which the nucleophilicity of the serine hydroxyl group is increased by the aspartate acting through the histidine. Even though the mechanism in the original form does not seem to be true, the interactions among these residues is still believed to be important in the catalytic mechanism (James *et al*., 1980; Steitz & Shulman, 1982).

As mentioned above, serine proteases have been isolated from many different sources. So far there are two distinct families of serine proteases. Within each family, the enzymes exhibit varying degrees of sequence homology but

'The sequence numbering used throughout this thesis for the chymotrypsin family of enzymes corresponds to the one obtained by the alignment of the sequence with that of chymotrypsinogen (Hartley, 1964).

share the same basic folding pattern. Despite the fact that the enzymes in the two families are folded in a completely different manner, they have the same catalytic groups located roughly in the same relative positions on the molecules. These two families are thought to be the product of convergent evolution. That is, each evolved independently to perform a common function and resulted in common structural features.

The first family includes the pancreatic enzymes chymotrypsin and trypsin. Members of this family are characterized by the sequence Gly-Asp-Ser-Gly-Gly around the active serine residue. Other members of this family perform a diversity of physiological functions. Elastase is another digestive enzyme from the pancreas (Hartley & Shotton, 1971) with an ability to digest elastin. Acrosomal protease is involved in the penetration of the zona pellucida of the ovum by the sperm (Stambaugh & Buckley, 1969). There are also many serine proteases that regulate physiological processes. The cleavage of fibrinogen to form fibrin by thrombin is the last step in the regulatory process leading to the formation of blood clots (Magnusson, 1971). Kallikreins are serine proteases that produce kinin from kininogen resulting in lowering of the blood pressure (Schacter, 1980). There are also enzymes of this family from prokaryotic sources. *Streptomyces griseus* secretes a mixture of enzymes collectively called pronase (Hiramatsu & Ouchi, 1963). Three of the enzymes have been characterized as being chymotrypsin-like serine proteases. These are called *Streptomyces griseus* protease A and B (SGPA and SGPB, respectively), and *Streptomyces griseus* trypsin (SGT) (Johnson & Smillie, 1974; Jurášek *et al*., 1974). SGPA and SGPB are very similar to each other and have chymotrypsin like activity. From an evolutionary standpoint, SGT is of particular importance because of its close resemblance to bovine trypsin (Hartley, 1970, 1979; Hewett-Emmett *et al*., 1981; Read *et al*., 1984). α-Lytic protease is another bacterial enzyme, obtained from *Lysobacter enzymogenes*. (Whitaker *et al*., 1965). It has been studied using NMR methods because the active site histidine is unique in the protein (review by Steitz & Shulman, 1982).

The second family of serine proteases is typified by the subtilisins. The enzymes in this family contain the sequence Thr-Ser-Met around their active site serine. So far only prokaryotic enzymes are known to belong to this family. It includes subtilisin Carlsberg from *Bacillus subtilis* (Smith *et al*., 1966; McPhalen *et al*., 1985a), subtilisin Novo (also known as subtilisin BPN') from *Bacillus amyloliquefaciens* (Wright *et al*., 1969; Drenth *et al*., 1972; McPhalen *et al*., 1985b), and proteinase K from the fungus *Tritirachium album* Limber (Pähler *et al*., 1984). The subtilisins have a chymotryptic specificity, cleaving the polypeptide chain on the carbonyl side of large aromatic amino acids. The discussion in this thesis will be mainly restricted to the chymotrypsin family of enzymes but the behavior of the subtilisins is similar.

The various members of the serine protease family exhibit not only differences in their specificities for substrates but also in the relative importance of their binding sites. Schechter and Berger (1967) have introduced a notation which is useful in discussing the interactions of a substrate with a protease. The amino acid residues of the substrate are labeled P_1', P_2, P_3, etc. toward the N-terminus from the scissile bond and P_1', P_2', etc. toward the C-terminus. The complementary regions on the enzyme are denoted as S_1, S_2, etc. and S_1', S_2', etc. (Fig. I.1). The binding sites on the enzyme can consist of more than one residue.

There have been extensive kinetic studies done to elucidate the nature of the substrate-enzyme interactions in several serine proteases (review by Fruton, 1975). In most of the enzymes studied the binding region extends for about

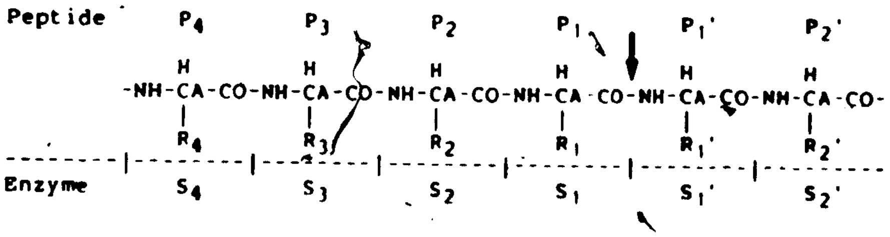

Figure I.1. Interactions Between a Protease and a Substrate. The notation introduced by Schecter & Berger (1967) is useful in describing protease-substrate interactions. The peptide substrate is labeled P_n and P_n' from the scissile bond (denoted by an arrow) and the sites on the enzymes that are complementary to these residues are labeled S_n and S_n', respectively.

six to seven sites of which about four are on the acyl side of the scissile bond. Chymotrypsin and trypsin are highly specific for the P_1 residue with the former recognizing Trp, Tyr, and Phe and the latter, Arg and Lys. For these enzymes the other sites are less important. Increasing the substrate length to cover all of the binding sites increases the hydrolysis rate about 4000 fold for chymotrypsin (Bauer, 1976; Bauer *et al.*, 1976) and about 300 fold for trypsin (Izumiya & Uchio, 1959; Yamamoto & Izumiya, 1967). In contrast SGPA, SGPB, α-lytic protease, and elastase all rely on interactions outside of the S_1 site. For these enzymes there is about 10^7 fold increase in the rate of hydrolysis of longer substrates (Thompson & Blout, 1970, 1973; Bauer, 1976; Bauer *et al.*, 1976; Bauer, 1978). The increase is mainly from an increase in k_{cat}. These results have been interpreted to mean that for these enzymes the S_1 pocket is not specific enough so that the interactions at the other sites are required for the optimal positioning of the substrate on the enzyme.

The primary specificities of SGPA and SGPB are similar to that of chymotrypsin except that Trp is not as favourable. α-Lytic protease and elastase are less specific, their binding sites recognize only small aliphatic groups in the P_1 position. Subtilisin also has a chymotrypsin-like P_1 specificity but in addition it has a site for binding a large hydrophobic residue at the P_4 position as deduced from structural studies (Robertus *et al.*, 1972).

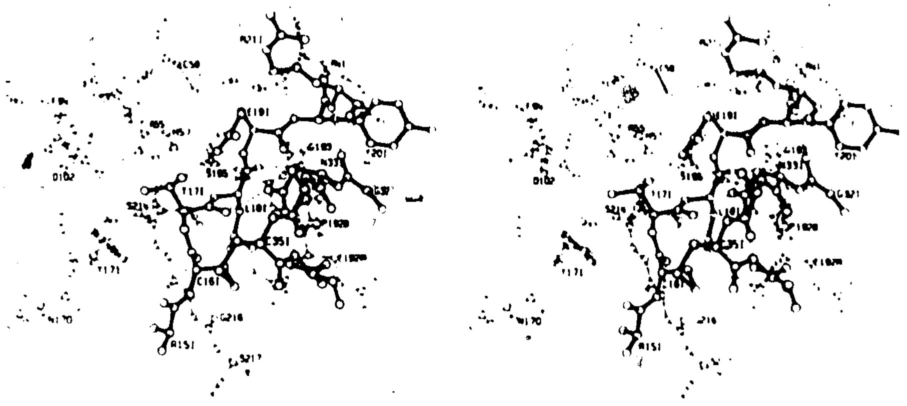

Figure I.2. Interactions Between a Protease (SGPB) and an Inhibitor (OMTKY3). The structure of the complex of SGPB and OMTKY3 has been refined at 1.8Å resolution (Fujinaga *et al.*, 1982). Shown are the residues in the active site region of SGPB (open bonds) and residues of the inhibitor (filled bonds) that are in the binding cleft of the enzyme. The side chain of Lys34I has been omitted for clarity.

Crystal structures of serine proteases and the complexes with inhibitors have helped to understand the kinetic results in structural terms (Figure I.2). There is a well defined pocket for the P_1 residue. For chymotrypsin, SGPA, and SGPB this is a deep depression that can accept a large aromatic side chain. In trypsin, the bottom of the pocket contains an aspartate residue which can form an ion pair with an Arg or a Lys side chain in the P_1 position of the substrate. There is only a shallow pocket in elastase and α-lytic protease, consistent with their specificity for small aliphatic residues.

Another important interaction for the P_1 residue is the binding of the carbonyl oxygen in the so called oxyanion hole (Henderson, 1970). This oxygen atom receives two hydrogen bonds from the main chain amide groups of Gly 193 and Ser 195. These interactions are thought to stabilize the oxyanion that forms in the transition state of the acylation reaction of the substrate with the active site serine. The N-terminal segment of the substrate is bound to the enzyme by β-sheet type interactions, making a pair of anti-parallel main chain hydrogen bonds at residue 216. There is also a long and poorly oriented hydrogen bond from the NH of the P_1 residue to the carbonyl group of Ser214 on the enzyme. The segment from 214 to 217 also forms one side of the S_1 pocket. The binding sites for the side chains of the N-terminal segment are not as well defined as the primary binding pocket. On the C-terminal side the only structural information comes from the complexes with the protein inhibitors such as the pancreatic trypsin inhibitor (PTI) (Huber *et al.*, 1974; Chen & Bode, 1983), ovomucoid inhibitor (OMTKY3) (Fujinaga *et al.*, 1982), and the pancreatic secretory trypsin inhibitor (PSTI) (Bolognesi *et al.*, 1982). These structures show binding sites for P_1' to P_3' with a possibility of main chain hydrogen bonds to be formed from P_2'.

The pathway for the reaction catalyzed by serine proteases has been worked out mainly by kinetic methods and has been reviewed by Bender and Killheffer (1973). The enzyme will catalyze the hydrolysis of an amide or an ester bond and in fact the process can be thought of as an acyl

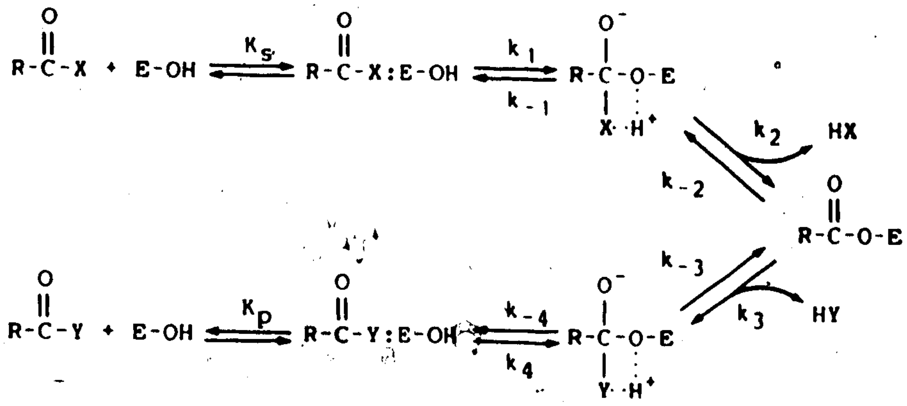

Figure I.3. The Reaction Pathway for Serine Proteases. The scheme drawn after Kraut (1977) shows the symmetrical nature of the acylation and deacylation steps. E represents the enzyme and X is the leaving group which is replaced by a nucleophile, Y.

transfer reaction. The symmetrical nature of the acylation and the deacylation steps are shown in Figure I.3 after the one given by Kraut (1977).

The first step in the reaction is the formation of the Michaelis complex. This is a state proposed for all enzymic reactions to explain saturation kinetics and it represents the productive binding of the substrate to the enzyme. The reaction proceeds with the nucleophilic attack of the carbonyl carbon atom of the scissile bond by the active site serine hydroxyl group. A tetrahedral transition state is formed which then breaks down to the acyl intermediate, releasing the C-terminal part of the substrate. The tetrahedral state is presumed to exist in order to form the covalent acyl-enzyme intermediate. According to the transition state stabilization theory (Pauling, 1946, 1948; Wolfenden, 1972) the enzyme should be designed to stabilize the tetrahedral species along the reaction pathway. The existence of an intermediate was established by the observation of a burst kinetic behavior using ester substrates for which the breakdown of the intermediate is rate limiting (Hartley & Kilby, 1954). The intermediate was shown to be a covalent acyl-enzyme by X-ray studies under conditions in which the intermediates are stable (Henderson, 1970; Alber *et al.*, 1976). It should be noted that this conclusion is based on difference maps calculated from unrefined structures and are thus not totally reliable. For amide substrate the formation of the acyl-enzyme is rate limiting and its existence could only be established by indirect means (Fastrez & Fersht, 1973). An intermediate before the acyl-enzyme had been observed by stopped-flow methods (Hunkapiller *et al.*, 1976; Petkov, 1978; Fink & Meehan, 1979; Compton & Fink, 1980). However, these results were interpreted as being artifacts when a subsequent study showed no evidence for a long lived tetrahedral intermediate (Markley *et al.*, 1981). The second half of the reaction, deacylation, is the reverse of the first half. It becomes a hydrolysis reaction when the attacking nucleophile is H_2O.

The reaction pathway should not be confused with the catalytic mechanism or how the enzyme can accelerate the reaction. In order for catalysis to occur the activation energy barrier to reach the transition state must be reduced. This may be achieved by at least two ways. First the reaction pathway can be changed from that of the uncatalyzed reaction so that smaller activation barriers are encountered, and second the transition states can be stabilized so as to reduce the activation energies. Serine proteases use various mechanisms to effect catalysis.

The binding sites on the enzyme position the substrate optimally. This is seen in kinetic studies using different lengths of polypeptide substrates. The increase in the hydrolysis rate for longer chains is mainly due to the increase in k_{cat} (Bauer *et al.*, 1981). The additional interactions that a longer peptide makes with the enzyme either improve its position with respect to the nucleophilic serine or place it in the optimal configuration in the transition state.

The reaction with the serine, to form an acyl-enzyme which is subsequently rapidly hydrolyzed, can be seen as nucleophilic catalysis. Before the attack by water takes place, the substrate is changed , by a nucleophile, into a more reactive species. This process is aided by general-base catalysis by the Asp-His pair and an electrostatic catalysis by the hydrogen bonds to the oxyanion hole.

The original proposal of a 'charge-relay' mechanism (Blow *et al.*, 1969) in which the negative charge on Asp 102 is transfered to Ser 195 by the double proton transfer from His 57 to Asp 102 and Ser 195 to His 57 is no longer acceptable. It has been shown that the histidine and aspartate residues have normal pK_a's and that the proton on the ND atom[1] of the histidine which is hydrogen bonded to Asp 102 stays there as the pH is lowered (Bachovchin *et al.*, 1981; Kossiakoff & Spencer, 1980,1981). The fate of the hydroxyl proton of Ser 195 as the reaction proceeds is not clear. In the native enzyme structures, the hydrogen bond between the NE of the histidine and the serine OG is weak making proton transfers difficult. On the other hand in the structures of the enzyme complexed with protein inhibitors such as PTI and OMTKY3 , there is a strong hydrogen bond between the serine and the histidine (Huber & Bode, 1978; Fujinaga *et al.*, 1982). There are conflicting NMR data about the charge on His 57 in the complex, but it is possible that it is protonated having received a proton from Ser 195 (Markley, 1979; Steitz & Shulman, 1982). That would leave the serine residue as an alkoxide ion which would be a very nucleophilic species. In these complexes the Ser195 OG is about 2.6Å from the carbonyl carbon atom of the scissile bond in the inhibitor (Huber & Bode, 1978; Fujinaga *et al.*, 1982). This distance is too long for a covalent interaction and too short for a van der Waals interaction. The chemical nature of this close approach of the oxygen atom to the carbon atom is not understood so that the possibility of a stable alkoxide ion cannot be ruled out. In addition to the deprotonation of the serine, the histidine is believed to be

[1] The atom nomenclature used in this thesis is according to the IUPAC-IUB recommendations of 1969 (1970).

important in protonating the leaving group on the substrate. The importance of the histidine residue is indicated by the pH dependence of the enzyme activity and inhibition by chemical modification of the histidine (Bender & Killheffer, 1973).

The role of Asp 102 is not clear. Its importance is implied by the fact that this residue is conserved in all serine proteases. It seems unlikely that it functions only to position the histidine residue. Computer simulations (Nakagawa & Umeyama, 1984; Weiner *et al*., 1986) indicate that the charge on Asp 102 helps to stabilize the charge developed on the tetrahedral transition state.

The catalytic effect of the oxyanion hole is much more straightforward. The two hydrogen bonds are formed from the amide nitrogen atoms of Gly 193 and Ser 195 to the carbonyl oxygen atom of the scissile bond on the substrate. This interaction would stabilize the negative charge developed on the oxygen atom in the transition state. There is an additional hydrogen bond which is thought to stabilize the transition state (Robertus *et al*., 1972). This is the interaction between the amide nitrogen atom of the P_1 residue and the carbonyl oxygen atom of Ser 214, which is long in the complexes with protein inhibitors (Chen & Bode, 1983; Read *et al*., 1983) but may become short in the tetrahedral state as indicated by the structure of SGPA complex with an aldehyde substrate (James *et al*., 1980). The complex is formed as a hemiacetal with a covalent link between Ser195 and the aldehyde and mimics a tetrahedral transition state.

Recently, more emphasis has been placed on understanding the differences between individual serine proteases. This is aided by the large number of serine protease structures that are now available and the advent of site directed mutagenesis. By understanding how various structural features give rise to characteristics such as specificity, reactivity, and stability, it will be possible to design new enzymes.

Bibliography

Alber, T., Petsko, G. A., & Tsernoglou, D. (1976) *Nature (London) 263*, 297-300.

Bachovchin, W. W., Kaiser, R., Richards, J., & Roberts, J. D. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78* 7323-7326.

Bailey, K., Bettelheim, F. R., Lorand, L., & Middlebrook, W. R. (1951) *Nature (London) 167* 233-234.

Bauer, C.-A. (1976) *Biochim. Biophys. Acta 438* 495-502.

Bauer, C.-A. (1978) *Biochemistry 17* 375-380.

Bauer, C.-A., Brayer, G. D., Sielecki, A. R., & James, M. N. G. (1981) *Eur. J. Biochem. 120* 289-294.

Bauer, C.-A., Thompson, R. C., & Blout, E. R. (1976) *Biochemistry 15* 1291-1295.

Bender, M. L. & Killheffer, J. V. (1973) *CRC Crit. Rev. Biochem. 1* 149-199.

Blow, D. M., Birktoft, J. J., & Hartley, B. S. (1969) *Nature (London) 221* 337-340.

Bolognesi, M., Gatti, G., Menegatti, E., Guarneri, M., Marquart, M., Papamokos, E., & Huber, R. (1982) *J. Mol. Biol. 162* 839-868.

Chen, Z. & Bode, W. (1983) *J. Mol. Biol. 164* 283-311.

Compton, P. & Fink, A. L. (1980) *Biochem. Biophys. Res. Commun. 93* 427-431.

Drenth, J., Hol, W. G. J., Jansonius, J. N., & Koekoek, R. (1972) *Eur. J. Biochem. 26* 177-181.

Fastrez, J. & Fersht, A. R. (1973) *Biochemistry 12* 2025-2041.

Fink, A. L. & Meehan, P. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76* 1566-1569.

Fruton, J. S. (1975) in *Proteases and Biological Control, Cold Spring Harbor Conferences on Cell Proliferation* Vol. 2 (Reich, E., Rifkin, D. B., & Shaw, E., eds.) pp33-50, Cold Spring Harbor Laboratory.

Fujinaga, M., Read, R. J., Sielecki, A., Ardelt, W., Laskowski, M. Jr., & James, M. N. G. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79* 4868-4872.

Hartley, B. S. (1964) *Nature (London) 201* 1284-1287.

Hartley, B. S. (1970) *Phil. Trans. Roy. Soc. Lond. B 257* 77-87.

Hartley, B. S. (1979) *Proc. R. Soc. Lond. B 205* 443-452.

Hartley, B. S. & Kilby, B. A. (1954) *Biochem. J. 56* 288-297.

Hartley, B. S. & Shotton, D. M. (1971) in *The Enzymes* Vol. 3, Third edition, (Boyer, P. D., ed.) pp323-373, Academic Press, New York.

Henderson, R. (1970) *J. Mol. Biol. 54* 341-354.

Hess, G. P. (1971) in *The Enzymes* Vol. 3, Third edition, (Boyer, P. D., ed.) pp213-248, Academic Press, New York.

Hewett-Emmett, D., Czelusniak, J., & Goodman, M. (1981) *Annals New York Acad. Sci. 370* 511-527.

Hiramatsu, A. & Ouchi, T. (1963) *J. Biochem. (Tokyo) 54* 462-464.

Huber, R. & Bode, W. (1978) *Acc. Chem. Res. 11* 114-122.

Huber, R., Kukla, D., Bode, W., Schwager, P., Bartels, K., Deisenhofer, J., & Steigemann, W. (1974) *J. Mol. Biol. 89* 73-101.

Hunkapiller, M. W., Forgac, M. D., & Richards, J. H. (1976) *Biochemistry 15* 5581-5588.

IUPAC-IUB Commission on Biochemical Nomenclature 1969 (1970) *Biochemistry 9* 3471-3479.

Izumiya, N. & Uchio, H. (1959) *J. Biochem. (Tokyo) 46* 645-652.

James, M. N. G., Sielecki, A. R., Brayer, G. D., Delbaere, L. T. J., & Bauer, C.-A. (1980) *J. Mol. Biol. 144* 43-88.

Johnson, P. & Smillie, L. B. (1974) *FEBS Letters 47* 1-6.

Jurášek, L., Carpenter, M. R., Smillie, L. B., Gertler, A., Levy, S., & Ericsson, L. H. (1974) *Biochem. Biophys. Res. Commun. 61* 1095-1100.

Kossiakoff, A. A. & Spencer, S. A. (1980) *Nature (London) 288* 414-416.

Kossiakoff, A. A. & Spencer, S. A. (1981) *Biochemistry 20* 6462-6474.

Kraut, J. (1977) *Ann. Rev. Biochem. 46* 331-358.

Magnusson, S. (1971) in *The Enzymes* Vol. 3, Third edition, (Boyer, P. D., ed.) pp277-321, Academic Press, New York.

Markley, J. L. (1979) in *Biological Applications of Magnetic Resonance* (Shulman, R. G. ed.) pp397-461, Academic Press, New York.

Markley, J. L., Travers, F., & Balny, C. (1981) *Eur. J. Biochem. 120* 477-485.

Matthews, B. W., Sigler, P. B., Henderson, R., & Blow, D. M. (1967) *Nature (London) 214* 652-656.

McPhalen, C. A., Schnebli, H. P., & James, M. N. G. (1985a) *FEBS Letters 188* 55-58.

McPhalen, C. A., Svendsen, I., Jonassen, I., & James, M. N. G. (1985b) *Proc. Natl. Acad. Sci. U.S.A. 82* 7242-7246.

Nakagawa, S. & Umeyama, H. (1984) *J. Mol. Biol. 179* 103-123.

Pähler, A., Banerjee, A., Dattagupta, J. K., Fujiwara, T., Lindner, K., Pal, G. P., Suck, D., Weber, G., & Saenger, W. (1984) *EMBO J. 3* 1311-1314.

Pauling, L. (1946) *Chem. Eng. News 24* 1375-1377.

Pauling, L. (1948) *Am. Sci. 36* 51-58.

Petkov, D. D. (1978) *Biochim. Biophys. Acta 523* 538-541.

Powers, J. C. (1977) in *Methods in Enzymology* Vol. 46 (Jakoby, W. B. & Wilcheck, M., eds.) pp197-208, Academic Press, New York.

Read, R. J., Brayer, G. D., Jurášek, L., & James, M. N. G. (1984) *Biochemistry 23* 6570-6575.

Read, R. J., Fujinaga, M., Sielecki, A. R., & James, M. N. G. (1983) *Biochemistry 22* 4420-4433..

Robertus, J. D., Kraut, J., Alden, R. A., & Birktoft, J. J. (1972) *Biochemistry 11* 4293-4303.

Schacter, M. (1980) *Pharm. Rev. 31* 1-17.

Schechter, I. & Berger, A. (1967) *Biochem. Biophys. Res. Commun. 27* 157-162.

Schoellmann, G. & Shaw, E. (1963) *Biochemistry 2* 252-255.

Sigler, P. B., Blow, D. M., Matthews, B. W., & Henderson, R. (1968) *J. Mol. Biol. 35* 143-164.

Smith, E. L., Markland, F. S., Kasper, C. B., DeLange, R. J., Landon, M., & Evans, W. H. (1966) *J. Biol. Chem. 241* 5974-5976.

Stambaugh, R. & Buckley, J. (1969) *J. Reprod. Fert. 19* 423-432.

Steitz, T. A. & Shulman, R. G. (1982) *Ann. Rev. Biophys. Bioeng. 11* 419-444.

Thompson, R. C. & Blout, E. R. (1970) *Proc. Natl. Acad. Sci. U.S.A. 67* 1734-1740.

Thompson, R. C. & Blout, E. R. (1973) *Biochemistry 12* 57-65.

Weiner, S. J., Weibel, G. L., & Kollman, P. A. (1986) *Proc. Natl. Acad. Sci. U.S.A. 83* 649-653.

Whitaker, D. R., Roy, C., Tsai, C. S., & Jurášek, L. (1965), *Can. J. Biochem. 43* 1961-1970.

Wolfenden, R. (1972) *Acc. Chem. Res. 5* 10-18.

Wright, C. S., Alden, R. A., & Kraut, J. (1969) *Nature (London) 221* 235-242.

Yamamoto, T. & Izumiya, N. (1967) *Arch. Biochem. Biophys. 120* 497-502.

II. α-Lytic Protease[1]

α-Lytic protease (EC3.4.21.-) is a bacterial serine protease isolated from the extracellular filtrate produced by the soil microorganism *Lysobacter enzymogenes* (Christensen & Cook, 1978, ATCC 29487, formerly known as *Myxobacter* 495). The enzyme has an elastase-like primary specificity, cleaving oligopeptide substrates on the carbonyl side of those amino acids with short aliphatic side chains (Whitaker *et al.*, 1965; Kaplan & Whitaker, 1969; Kaplan *et al.*, 1970). Kinetic studies with synthetic peptides have shown that the substrate binding interactions extend over at least six subsites (Bauer *et al.*, 1981).

Various NMR studies, designed to probe the protonation state of the catalytic triad (Asp102, His57, Ser195), have been done with α-lytic protease (see review by Steitz & Shulman, 1982). The enzyme is an ideal subject for such studies because the active site histidine is unique in the protein.

Comparison of the three dimensional structure of α-lytic protease at 2.8Å resolution (Brayer *et al.*, 1979) with that of elastase (Sawyer *et al.*, 1978) showed that in spite of relatively weak sequence homology (James *et al.*, 1978) the enzymes display extensive tertiary structural homology. Large insertions and deletions in the sequence of α-lytic protease relative to the pancreatic enzymes made the

[1] A version of this chapter has been published [Fujinaga, M., Delbaere, L. T. J., Brayer, G. D., & James, M. N. G. (1985) *J. Mol. Biol.* *184* 479-502].

tertiary structural predictions from comparative model-building exceedingly difficult (McLachlan & Shotton, 1971; Delbaere *et al.*, 1979). α-Lytic protease is much more homologous to the bacterial serine proteases from *Streptomyces griseus*, SGPA and SGPB (James *et al.*, 1978). The sequence identity with these enzymes is 40% and 43% respectively and the structural equivalences are approximately 80%. In this chapter, the sequences of the three bacterial enzymes, α-lytic protease, SGPA and SGPB and that of α-chymotrypsin have been aligned. This alignment is based on the refined high resolution structures of all 4 molecules and differs in detail in several places from the previous alignment based only on unrefined α-carbon atom coordinates (James *et al.*, 1978). The amino acid numbering scheme used herein for α-lytic protease is based on this new alignment. The structures of α-lytic protease and SGPA have been compared and the differences in the hydrophobic cores of the enzymes were analyzed. In addition, the refinement of the structure allows for a detailed examination of the hydrogen bonding and the solvent structure. The refined parameters of α-lytic protease have been deposited with the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977).

A. Experimental Procedures

Crystallization and Data Collection

A single crystal grown from 1.3M Li_2SO_4, pH 7.2 (Brayer *et al.*, 1979) was used to collect a complete 1.8Å resolution data set on which much of the refinement was based. This data set was replaced at a later stage in the refinement by a 1.7Å resolution intensity data set collected from a crystal grown from 1.8M Li_2SO_4, pH 6.1. This crystal, grown in slightly different conditions from the first crystal, was the only one suitable for high resolution data collection. The details of the crystal parameter and data collection conditions are summarized in Table II.1.

Refinement

The 2.8Å resolution multi-isomorphous replacement structure for α-lytic protease (Brayer *et al.*, 1979) was refined using the restrained-parameter least-squares program of Hendrickson & Konnert (1980). The progress of the refinement was slow as seen in the plot of the R-factor[1] versus refinement cycle (Figure II.1). About every ten cycles of refinement, the current model was displayed with a superimposed electron density map that had been computed with coefficients $2|F_O|-|F_C|$, α_C (α_C = calculated phase), on the MMS-X graphics system (Barry *et al.*, 1976) using the

[1] $R = \Sigma||F_O|-|F_C||/\Sigma|F_C|$, where $|F_O|$ and $|F_C|$ are the observed and calculated structure factor amplitudes, respectively.

Table II.1

Crystal Data and Data Processing Information

	Crystal 1	Crystal 2
Growth condition	1.3 M Li_2SO_4 (pH7.2)	1.8 M Li_2SO_4 (pH6.1)
Space group	$P3_221$	$P3_221$
a	66.28(3) Å	66.21(3) Å
b	66.28(3) Å	66.21(3) Å
c	80.08(4) Å	80.01(4) Å
Diffractometer	Picker FACS-1	Enraf-Nonius CAD4
Incident beam	Ni-filtered CuKα,40kV,26mA	Ni-filtered CuKα,40kV,26mA
Diffracted beam	65 cm crystal-counter He-filled beam path	60 cm crystal-counter He-filled beam path
Scan type	ω scan, continuous	ω scan, continuous
Scan width	0.55° at 2°/min	0.8° at 0.8°/min
Background measurement	Two 4s fixed position counts taken at 0.42° offset from ω=0° in ω direction	0.2° in ω on either side of the peak scan
Background correction	Bicubic spline fit of the measured background as a function of 2θ and φ	Same
Absorption correction	North *et al.* (1[illegible])	Same
Decay correction	Hendrickson (1976)	Same
Geometric correction	Lorentz and polarization	Same
Min. d-spacing	1.8Å	1.7Å
Total no. of reflections	19,304	22,828
Observed reflections	14,246 [≥2σ(I)]	15,241 [≥2σ(I)]

$\sigma^2(I)$	$I+c^2I^2+(t_I/t_{Bk})^2(\Sigma Bk+c^2\Sigma Bk^2)$
	I=total intensity
	Bk=background counts
	t_I=time taken for intensity measurement
	t_{Bk}=time taken for background measurement
	c=instrument instability=0.01	

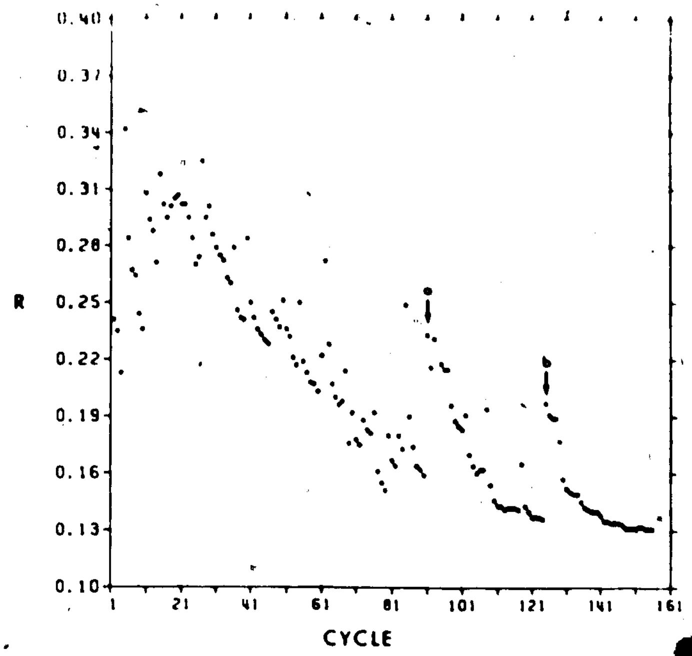

Figure II.1. Progress of the Refinement. A plot of the R-factor at each cycle of refinement. Point a indicates the first cycle using the new data from crystal 2 (Table II.1). All solvent molecules in the model were deleted. Point b indicates the first cycle of the solvent position redetermination. Solvent molecules were deleted at this point as well. Data of different resolutions were used to calculate the R-factor at different points of refinement. For example the low values of R at the beginning of the refinement is due to the low resolution data used.

program M3 written by Colin Broughton of this laboratory (Sielecki *et al.*, 1982). Manual adjustments were made to the model as indicated by the electron density. Table II.2 gives some of the details of the refinement progress at various stages. At cycle 91, the data set was replaced by a new one (crystal 2, Table II.1) with the hope that better data may speed up the refinement. However, no significant improvement in the rate of refinement was observed.

Table II.2

Summary of the Refinement of α-lytic Protease

Cycle	2	34	66	91[1]	124	156
R	0.241	0.241	0.196	0.233	0.136	0.131
r.m.s. Δ bond distance (Å)	0.023	0.034	0.029	0.032	0.019	0.019
r.m.s. Δ angle distance (Å)	0.089	0.056	0.038	0.040	0.031	0.030
$\|F_o\|-\|F_c\|$	124.8	70.1	54.5	86.8	33.8	32.5
No. of parameters	4190	5670	5960	4175	6370	6385
No. of solvents	0	21	75	0	153	156
Resolution range	6-3.8Å	6-1.8Å	6-1.8Å	6-2.8Å	8-1.7Å	8-1.7Å
No. of data	1538 (≥3σ)	11,716 (≥3σ)	13,638 (≥2σ)	4200 (≥2σ)	14,996 (≥2σ)	14,996 (≥2σ)
$\bar{B}$ ($Å^2$)	5.7	8.6	10.6	13.7	13.6	14.3

[1] New data using crystal 2 (see Table II.1).

Solvent molecules were included in the model starting at cycle 33. Difference electron density maps computed with coefficients $|F_o|-|F_c|$, and phases α_c, were used to locate possible solvent positions. Peaks in the electron density that were in positions with hydrogen bonding potential were interpreted as water molecules and were allowed to refine as neutral oxygen atoms with individual temperature factors, B($Å^2$) and occupancies. The indicated shifts on B's and occupancies were applied in alterate cycles. Solvent molecules that refined to very high temperature factors or very low occupancy values were eliminated. When the data set was changed at cycle 91 all the solvents were deleted and a new difference map was calculated. The map showed that many of the solvents were in the same positions in the two crystals.

In addition to these water molecules, two sulfate ions have been included in the model. These were refined with just individual B factors.

Redetermination of Solvent Positions

At cycle 124 of the refinement, it was felt that there were problems with the procedure used to select solvent positions. The requirement that a potential solvent atom be in a hydrogen bonding position relative to polar protein atoms biases the resulting solvent structure. To overcome this problem, the solvent positions were redetermined using a new procedure as follows:

(1) Remove all solvents and allow the protein atoms to refine for a few cycles (typically 4).

(2) Calculate a difference map and select as solvent positions all peaks with electron densities higher than a relatively high cutoff level. A high cutoff is necessary because of the larger errors in the phases at this point.

(3) Refine with the new solvents for a few cycles (again, typically 4).

(4) Calculate a difference map and use a lower cutoff level to pick out new solvents.

(5) Repeat steps 3 and 4.

It is important, at each iteration, to keep the cutoff level high enough so that noise is not misinterpreted as a solvent atom. In this context, additional criteria that the total peak size be a certain value and that the peak be relatively convex by visual inspection were imposed. In the last difference map a cutoff level of 0.22 e/Å^3 and a minimum peak size of 0.21 e/peak were used. The error of the map was estimated to be 0.05 e/Å^3 (Blundell & Johnson, 1976).

Estimation of Errors

The determination of errors in atomic coordinates in a structure refined by a restrained least-squares procedure is not straightforward. The reader is referred to Read *et al.* (1983) for a full discussion.

Thus coordinate errors were estimated using three different methods. The method of Luzzati (1952) estimates the error from the variation of the R-factor with resolution. From such a plot (not shown), we estimate the overall root-mean-square (r.m.s.) error for the refined α-lytic protease structure as 0.14Å. However, from this method only an overall error estimate is obtained.

An estimate of the errors in individual atomic positions may be calculated from the formula given by Cruickshank (1949, 1954, 1967). The equation for a trigonal space group is similar to that for a monoclinic space group and is given by

$$\sigma_{zi} = \frac{c[\sum_{hkl} l^2(|F_o|-|F_c|)^2]^{1/2}}{2\pi \sum_{hkl} (m/2)l^2 f_{oi}\exp[-B_i\sin^2\theta/\lambda^2]}$$

where c is the axial length, f_{oi} is the atomic scattering factor, and m = 2 or 1 depending on whether the reflection is centric or not. The radial error is given by $\sigma_{ri} = 3\ \sigma_{zi}$. The r.m.s. error for all the atoms in the protein calculated using the Cruickshank formula is 0.10Å.

The third way to estimate positional errors is to compare the coordinates of the refined structure with those obtained by refining without restraints (Chambers & Stroud, 1979). Four cycles of unrestrained refinement converged with an R-factor of 0.117. The r.m.s. difference between the coordinates of the restrained and unrestrained structures as a function of the B factor for each atom type is shown in Figure II.2. The very large deviations of nitrogen atoms with large B-factors (Figure II.2b) correspond to disordered arginine side chains. The overall r.m.s. difference is 0.12Å for all the atoms of the protein. The errors predicted by the Cruickshank formula are superimposed on these plots and show reasonable agreement between the two methods. These plots can be used to obtain an estimate of the error as a function of B. The variation of B along the polypeptide chain is given in Figure II.3.

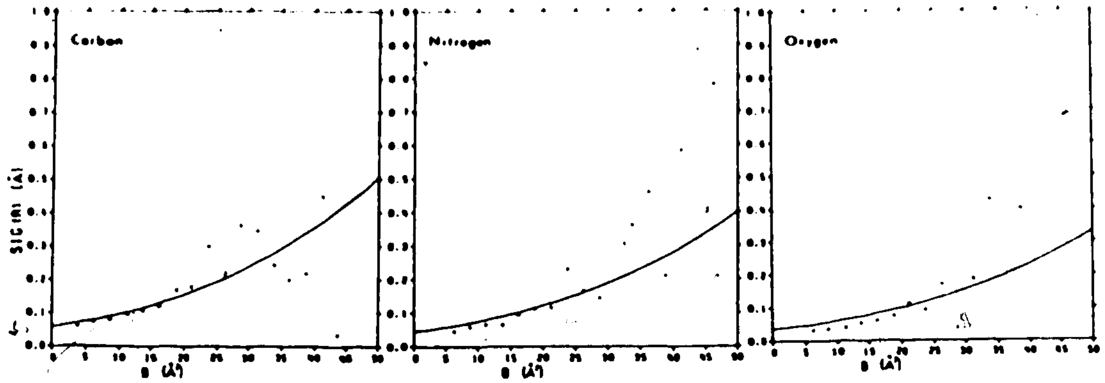

Figure II.2. Coordinate Errors. σ_r as a function of the isotropic temperature factor (B). The points in the plots represent the r.m.s. difference between restrained and unrestrained structures. The curve in each plot is the radial error [SIG(R)] calculated from the Cruickshank formula for the corresponding atom type.

Structure Equivalence and Sequence Alignment

The refined structures of SGPA at 1.5Å resolution (Sielecki *et al*., 1979; Sielecki & James, 1981), SGPB at 1.7Å resolution (Sawyer *et al*., unpublished), α-chymotrypsin at 1.8Å resolution (the α-chymotrypsin structure is that from the crystals of its complex with the third domain of the turkey ovomucoid inhibitor, Read, Fujinaga, Sielecki, Ardelt, Laskowski & James, unpublished results) and α-lytic protease were used. We used the program of W. Bennett that is based on the methodology of Rossmann and Argos (1975) to determine the structural equivalence of these four enzymes. Those pairs of α-carbon atoms that differed in position by less than 1.9Å were considered structurally equivalent in the two enzymes being compared. For the comparison of the three bacterial enzymes with bovine pancreatic

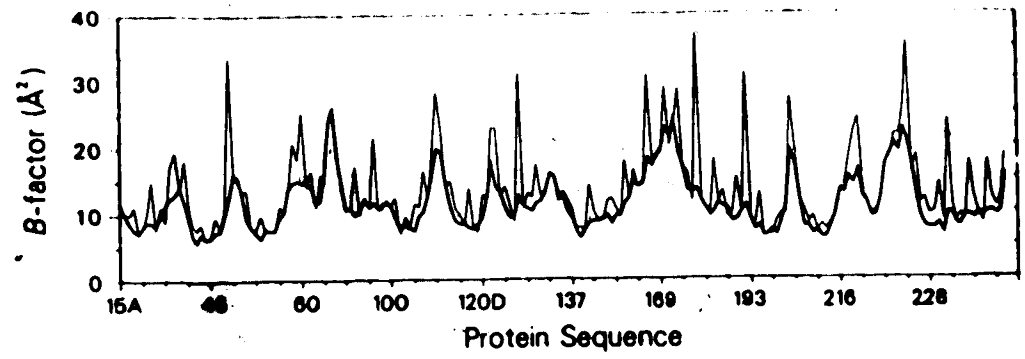

Figure II.3. The Variation in the B-factor Along the Polypeptide Chain. The thick line shows the mean B (Å^2) of the main chain atoms while the thin line shows the corresponding values for the side chain atoms.

α-chymotrypsin we relaxed this criterion to 3.0Å.

B. Results and Discussion

Overall Structure

The refinement procedure has resulted in a structure with good geometry and a low R-factor. The final model consists of 1391 protein atoms, 2 sulfate ions and 156 water molecules. The R-factor is 0.131 for the 14996 reflections between 8-1.7Å resolution that satisfy the criterion $I \geq 2\sigma(\bar{I})$.

The quality of the structure can be judged from the data in Table II.3 that summarizes the deviation from ideal geometry. Another indication of the quality can be seen in a ϕ-ψ plot (Ramakrishnan & Ramachandran, 1965) shown in Figure II.4. There are only five non-glycyl residues

Table II.3

r.m.s Deviations from Ideal Geometry at the End of Refinement

Distance restraints (Å)	
Bond distance	0.019(0.015)[1]
Angle distance	0.030(0.020)
Planar 1-4 distance	0.033(0.020)
Plane restraint (Å)	0.018(0.015)
Chiral-volume restraint (Å^3)	0.188(0.100)
Non-bonded contact restraint (Å)	
Single torsion contact	0.209(0.250)
Multiple torsion contact	0.128(0.250)
Possible hydrogen bond	0.132(0.250)
Peptide torsion angle restraint (°)	3.4(2.0)

[1] The values of σ, in parentheses, are the input estimated standard deviations that determine the relative weights for the corresponding restraints (see Hendrickson & Konnert, 1980).

(Ala39, Asn60, Pro95, Thr54, Pro120) whose ϕ,ψ values fall outside the acceptable regions. The first three are located in turns, while the remaining residues, Thr54 and Pro120, are located very close to the boundaries of the acceptable regions on the ϕ-ψ plot. Proline 95 with ϕ,ψ values (-81°,-155°) is in a cis conformation. The corresponding prolyl residues in SGPA and SGPB are also in cis-conformations with very similar ϕ-ψ angles (Brayer *et al*., 1978).

The general folding of the polypeptide chain is as described in the 2.8Å resolution study (Brayer *et al*., 1979). The major changes to the structure as a result of the refinement are due to side chain reorientation and flipping of four peptide bonds. The r.m.s. difference in position between the 2.8Å resolution structure and the refined structure is 0.88Å for the main chain atoms and

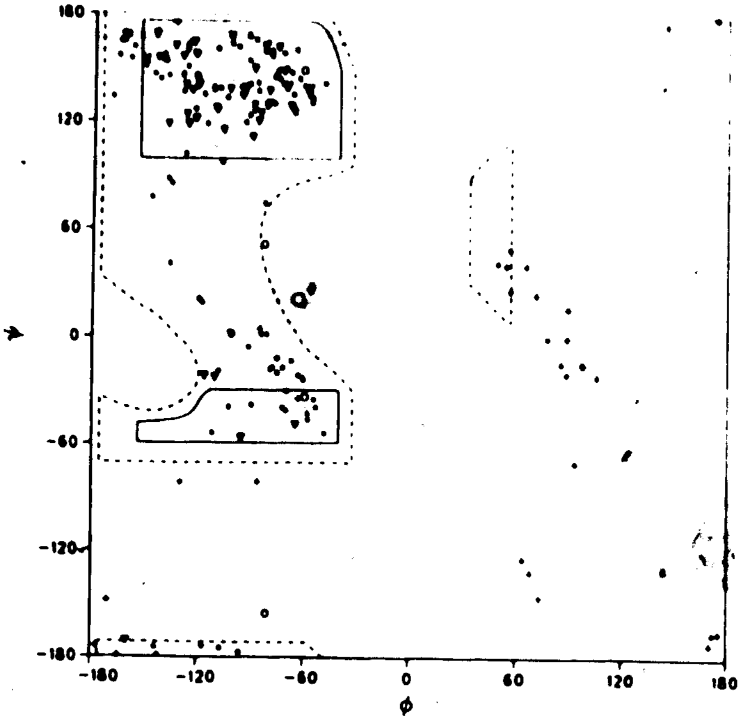

Figure II.4. ϕ-ψ Plot of the Refined Structure of α-lytic Protease. The symbols correspond to the following residue types: (o) proline, (∇) β-branched amino acids, (+) glycine, (•) others. The solid lines enclose areas that are the fully allowed conformational regions and the broken line shows the areas of acceptable van der Waals contacts for $\tau(C^{\alpha})$ of 115° (Ramakrishnan & Ramachandran, 1965).

1.60Å for the side chain atoms. These values include average shifts in x, y, z of 0.54, 0.06, 0.17Å for the main chain atoms and 0.49, 0.02, 0.12Å for the side chain atoms. Figure II.5(a) shows the complete molecule while Figure II.5(b) shows the main chain atoms with the hydrogen bonds in dashed lines.

The redetermination of solvent positions at cycle 124 resulted in 156 water molecules and 2 sulfate ions. Of the 153 water molecules in the model before the redetermination, 24 did not reappear. The water molecules were ordered

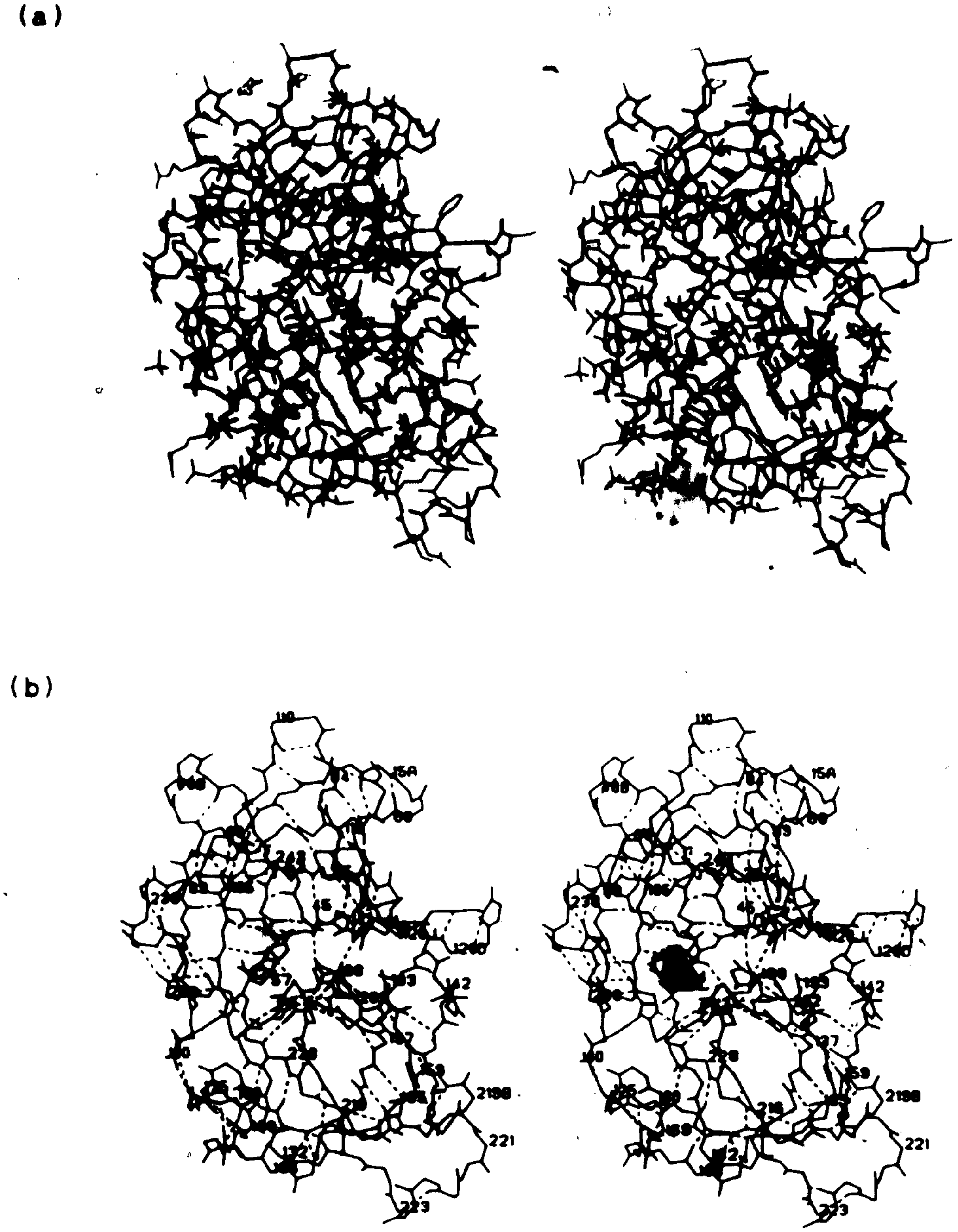

Figure II.5. Stereoscopic View of α-lytic Protease. (a) The complete molecule of α-lytic protease. The peptide backbone is shown in thick lines while the side chains are shown in thin lines.
(b) The same view as (a) but without the side chain atoms. The dashed lines indicate the hydrogen bonds between main chain atoms. The residue number is given every five residues (see Table II.4).

according to a quality factor (James & Sielecki, 1983) that is defined as occupancy^2/B and indicates roughly the reliability of each molecule. They were numbered sequentially in decreasing order of their quality factor. The 23 incorrect water molecules were among those with the lowest quality factors. It would seem that our refinement procedure is not always able to discriminate noise from weak features.

Comparison with Homologous Enzymes

The amino-acid sequences of α-lytic protease, SGPA, SGPB and α-chymotrypsin have been aligned (Table II.4) and the amino acid residues are numbered according to this new alignment. This alignment was constructed from the results of the pairwise structural comparisons of the four proteins using the refined atomic coordinates. The fitting of each structure, one to the other, was done on the basis of a rigid body least-squares refinement procedure with α-carbon atom coordinates only. The results are summarized in Table II.5. It is clear from Tables II.4 and II.5 that the three bacterial enzymes are much more similar to one another than they are to α-chymotrypsin although the structural equivalences to the latter are considerable. There are approximately 110 α-carbon atoms of the bacterial proteases that are structurally equivalent to corresponding α-carbon atoms in the mammalian pancreatic proteases.

Table II.4

The Sequence Alignment of Homologous Serine Proteases

	A	B												
	15	15	16	17	18	19	20	21	22	23	24	25	26	27
LP	A	N	I	V	G	G	-	-	-	-	-	-	-	-
PA	-	-	I	A	G	G	-	-	-	-	-	-	-	-
PB	-	-	I	S	G	G	-	-	-	-	-	-	-	-
CH	-	-	I	V	N	G	E	E	A	V	P	G	S	W

	28	29	30	31	32	33	34	35	36	37	38	39	40	41
LP	-	-	-	I	E	Y	S	I	N	-	N	A	S	L
PA	-	-	-	E	A	I	T	T	-	-	G	G	S	R
PB	-	-	-	D	A	I	Y	S	-	-	S	T	G	R
CH	P	W	Q	V	S	L	Q	D	K	T	G	F	H	F

				A					A	B	C			
	42	43	44	44	45	46	47	48	48	48	48	49	50	51
LP	C	S	V	G	F	S	V	T	R	G	A	T	K	G
PA	C	S	L	G	F	N	V	S	V	N	G	V	A	H
PB	C	S	L	G	F	N	V	R	S	G	S	T	Y	Y
CH	C	G	G	-	S	L	I	N	-	-	-	E	N	W

									A	B				
	52	53	54	55	56	57	58	59	59	59	60	61	62	63
LP	F	V	T	A	G	H	C	G	T	V	N	A	T	-
PA	A	L	T	A	G	H	C	T	N	I	S	A	S	-
PB	F	L	T	A	G	H	C	T	D	G	A	T	T	-
CH	V	V	T	A	A	H	C	G	-	-	V	T	T	S

	64	65	66	67	68	69	70	71	72	73	74	75	76	77
LP	A	R	I	G	-	-	-	-	-	-	-	-	-	-
PA	W	S	-	-	-	-	-	-	-	-	-	-	-	-
PB	W	W	A	N	S	A	-	-	-	-	-	-	-	-
CH	D	V	V	V	A	G	E	F	D	Q	G	S	S	S

(Table II.4 continued)

	78	79	80	81	82	83	84	85	86	87	88	A 88	89	90
LP	-	-	-	G	A	V	V	G	-	T	F	A	A	R
PA	-	-	-	-	-	-	I	G	-	T	R	T	G	T
PB	-	-	R	T	T	V	L	G	-	T	T	S	G	S
CH	E	K	I	Q	K	L	K	I	A	K	V	-	F	K

	91	92	93	94	95	96	97	98	99	100	101	102	103	104
LP	V	-	-	F	P	-	-	-	-	G	N	D	R	A
PA	S	-	-	F	P	-	-	-	-	N	N	D	Y	G
PB	S	-	-	F	P	-	-	-	-	N	N	D	Y	G
CH	N	S	K	Y	N	S	L	T	I	N	N	D	I	T

	105	106	107	108	109	110	111	112	113	A 113	114	115	116	117
LP	W	V	S	L	T	S	A	Q	T	-	L	-	-	-
PA	I	I	R	H	S	N	P	A	A	-	A	-	-	-
PB	I	V	R	Y	T	N	T	T	I	P	K	-	-	-
CH	L	L	K	L	S	T	A	A	S	-	F	S	Q	T

	118	119	120	A 120	B 120	C 120	D 120	E 120	F 120	G 120	H 120	I 120	J 120	K 120
LP	-	L	P	R	V	A	N	G	-	S	S	F	V	T
PA	-	D	G	R	V	Y	L	Y	N	G	S	Y	Q	D
PB	-	D	G	T	V	G	-	-	-	-	-	G	Q	D
CH	V	S	A	-	-	-	-	-	-	-	-	-	-	-

	121	122	123	124	125	126	127	128	129	130	131	132	133	134
LP	V	R	G	S	T	-	-	-	E	A	A	V	G	A
PA	I	T	T	A	G	-	-	-	N	A	F	V	G	Q
PB	I	T	S	A	A	-	-	-	N	A	T	V	G	M
CH	V	C	L	P	S	A	S	D	D	F	A	A	G	T

(Table II.4 continued)

	135	136	137	138	139	140	141	142	143	144	145	146	147	148
LP	A	V	C	R	S	G	R	T	T	-	-	-	-	-
PA	A	V	Q	R	S	G	S	T	T	-	-	-	-	-
PB	A	V	T	R	R	G	S	T	T	-	-	-	-	-
CH	T	C	V	T	T	G	W	G	L	T	R	Y	T	N

	149	150	151	152	153	154	155	156	157	158	159	160	161	162
LP	-	-	-	-	-	-	-	G	Y	Q	C	G	T	I
PA	-	-	-	-	-	-	-	G	L	R	S	G	S	V
PB	-	-	-	-	-	-	-	G	T	H	S	G	S	V
CH	A	N	T	P	D	R	L	Q	Q	A	S	L	P	L

	163	164	165	166	167	168	169	170	171	172	173	174	175	176
LP	T	A	K	N	V	T	A	N	Y	-	A	E	G	A
PA	T	G	L	N	A	T	V	N	Y	G	S	S	G	I
PB	T	A	L	N	A	T	V	N	Y	G	G	G	D	V
CH	L	S	N	T	N	C	K	K	Y	W	G	T	K	I

	177	178	179	180	181	182	183	184	185	186	187	188	189	190
LP	V	R	G	L	T	Q	G	N	A	-	-	-	C	M
PA	V	Y	G	M	I	Q	T	N	V	-	-	-	C	A
PB	V	Y	G	M	I	R	T	N	V	-	-	-	C	A
CH	K	D	A	M	I	C	A	G	A	S	G	V	S	S

	191	192	193	194	195	196	197	198	199	200	201	A 201	202	203
LP	G	R	G	D	S	G	G	S	W	I	T	S	A	-
PA	Q	P	G	D	S	G	G	S	L	F	A	-	G	-
PB	E	P	G	D	S	G	G	P	L	Y	S	-	G	-
CH	C	M	G	D	S	G	G	P	L	V	C	-	K	K

(Table II.4 continued)

	204	205	206	207	208	209	210	211	212	213	214	215	216	217
LP	-	-	-	G	Q	A	Q	G	V	M	S	G	G	N
PA	-	-	-	S	T	A	L	G	L	T	S	G	G	S
PB	-	-	-	T	R	A	I	G	L	T	S	G	G	S
CH	N	G	A	W	T	L	V	G	I	V	S	W	G	S

			A	B	C	D		A			A	B	C	
	218	219	219	219	219	219	220	220	221	222	222	222	222	223
LP	V	Q	S	N	G	N	N	C	G	I	P	A	S	Q
PA	-	-	-	-	-	G	N	C	R	T	G	-	-	-
PB	-	-	-	-	-	G	N	C	S	S	G	-	-	-
CH	S	T	-	-	-	-	C	-	S	T	-	-	-	S

	224	225	226	227	228	229	230	231	232	233	234	235	236	237
LP	R	S	S	L	F	E	R	L	Q	P	I	L	S	Q
PA	G	T	T	F	Y	Q	P	V	T	E	A	L	S	A
PB	G	T	T	F	F	Q	P	V	T	E	A	L	V	A
CH	T	P	G	V	Y	A	R	V	T	A	L	V	N	W

	238	239	240	241	242	243	244	245
LP	Y	G	L	S	L	V	T	G
PA	Y	G	A	T	V	L	-	-
PB	Y	G	V	S	V	Y	-	-
CH	V	Q	Q	T	L	A	A	N

The sequence alignment of α-lytic protease (LP), SGPA (PA), SGPB (PB), and α-chymotrypsin (CH) (the 1-letter code of the amino acids is used). This alignment was derived from the structural equivalence observed from the refined crystal structures of all four enzymes, based on the algorithm of Rossmann & Argos (1975). The continuous and dotted lines under the sequence numbers refer to the structurally equivalent residues common to the three bacterial serine proteases; continuous lines correspond to residues that have α-carbon atoms superimposable simultaneously to within 1.0Å, the dotted lines correspond to additional residues that are superimposable when the limit is relaxed to within 1.9Å. The detailed summary of the pairwise fitting is given in Table II.5. The continuous and dotted lines drawn under the amino

(Table II.4 continued)

acid sequence of α-chymotrypsin refer to the 112 structurally equivalent residues of α-lytic protease and α-chymotrypsin. The 94 residues with a continuous underline are those for which the α-carbon atoms are superimposed simultaneously to within 1.9Å; r.m.s. deviation 1.04Å. There are 18 additional α-carbon atoms superimposable when the limit is relaxed to 3.0Å. Table II.5 summarises this sequence identity in the structurally equivalent regions; it ranges from 21% (α-lytic protease versus α-chymotrypsin) to 70% (SGPA versus SGPB).

In addition to the least-squares superposition, we have compared the structures in pairs on the MMS-X interactive graphics system (Sielecki *et al*., 1982). In Table II.4, the so derived regions of structural equivalence are depicted in two ways. For the comparison of α-lytic protease with α-chymotrypsin, the 112 amino acid residues that are underlined (solid or dotted lines at the bottom of the rows) are considered to be structurally equivalent. The 94 residues with a solid underline are those for which the α-carbon atoms can be superimposed simultaneously within 1.9Å; the r.m.s. deviation in their positions is 1.04Å. The dotted underlines indicate the additional 18 residues that can be superposed within 3.0Å (Table II.4).

The regions of structural equivalence among the three bacterial serine proteases are indicated in Table II.4 by solid and dotted lines on top of the rows, below the sequence numbers. The criterion for deducing these regions was more stringent than that used for the comparison of α-lytic protease and α-chymotrypsin. Solid lines correspond to those regions that are superposed simultaneously to

Table II.5

Summary of the Structural Pairwise Fitting of the Bacterial Serine Proteases with α-chymotrypsin

	SGPA (181)	SGPB (185)	α-lytic protease (198)	α-chymotrypsin (230)
SGPA	-	162(0.55)	159(0.76)	91(0.98) [116(1.47)]
SGPB	114(70)	-	150(0.74)	88(0.96) [110(1.36)]
α-lytic protease	63(40)	64(43)	-	94(1.04) [112(1.34)]
α-chymotrypsin	27(23)	28(25)	24(21)	-

The number of amino acids in each protein are given below the name. (The A chain of α-chymotrypsin has been omitted from this comparison). The upper triangular portion of the matrix has the number of structurally equivalent α-carbon atom pairs that fit within 1.9Å of each other in the two enzymes compared. The r.m.s. deviation in Å for those α-carbon atom pairs is given in parentheses. Comparisons of the bacterial serine proteases with α-chymotrypsin also give in square brackets the number of α-carbon atom pairs that fit within 3.0Å. The lower triangular matrix gives the number of identical amino acid residues in the structurally equivalent portions of the pairs of enzymes. The number in parentheses is the percentage of the residues comprising the structurally equivalent segments that are identical.

within 1.0Å; the r.m.s. deviations range from 0.40Å to 0.52Å. The dotted lines correspond to those additional residues that are superposed simultaneously when the limit is relaxed to 1.9Å.

The strands of polypeptide chain that exhibit the greatest structural equivalence are those strands either containing the residues of the catalytic triad, Asp102, His57 and Ser195, or are the strands spatially close to them. There are ten such strands forming two antiparallel β-pleated sheets that are the junction between the two domains of the enzyme (see Figs. II.5(a) and (b) for the α-lytic protease structure). Within the structurally equivalent regions of α-lytic protease and α-chymotrypsin, the amino-acid sequence identity is relatively low; only 24 of the 112 structurally equivalent residues are identical and of these 11 are on the strands carrying His57 and Ser195 (Table II.4).

There are five major regions of the two enzymes with no structural homology. The N-termini, although exhibiting remarkable sequence identity, have completely different conformations. The Ca^{2+} binding loop of the trypsin family (residues 70 to 80 in the pancreatic serine proteases) is missing in α-lytic protease, as is the autolysis loop, 144-155. The part of the methionine loop from residue 164 to 176 has an α-helical conformation in α-chymotrypsin, but is in an extended β-sheet conformation in the bacterial enzymes. α-lytic protease has a four residue insertion at position 219 in the primary specificity pocket that, in part, limits the size of P_1 residues in potential substrates. These regions of the α-lytic molecule have been discussed in more detail (Brayer *et al.*, 1979).

The similarity of structure and sequence among the bacterial proteases can be exemplified by comparing SGPA and α-lytic protease. Within the same definition of structural equivalence as used above, there are 159 structurally homologous pairs of amino-acid residues (α-carbon atom positions) between SGPA (refined coordinates at 1.5Å resolution, Sielecki & James, 1981) and α-lytic protease (Table II.4). This represents 88% of the residues in SGPA (total of 181 amino acids). The r.m.s. deviation of these superposed α-carbon atom pairs is 0.76Å. Of the 159 structurally equivalent pairs of residues in SGPA and α-lytic protease, 63 (40%) are chemically identical.

This highly homologous folding of the polypeptide chain results in very similar active site regions of the two enzymes. Fitting 84 atoms from 15 amino-acid residues that comprise the active site region of α-lytic protease to corresponding ones of SGPA via a least-squares procedure results in an r.m.s. deviation of 0.28Å; for the corresponding atom pairs of SGPA and SGPB this r.m.s. deviation is 0.15Å.

With such remarkable structural similarity, it is not surprising that the chemical sequence identity is maximal in the neighborhood of the catalytic triads (Table II.4). The active sites of these enzymes are found at the junction of the two domains of the molecules. Since the polypeptide chains are structurally equivalent in these domains, it is also likely that the equivalent amino-acids of the hydrophobic core be identical. However the hydrophobic cores (for the N-terminal and C-terminal domains) have markedly different amino acid compositions in SGPA and α-lytic protease. Of the 159 equivalent pairs of residues in SGPA and α-lytic protease, the 72 in the N-terminal domain have an r.m.s. deviation of 0.87Å whereas the 87 in the C-terminal domain have an r.m.s. deviation of 0.64Å.

The N-terminal domains of the two bacterial enzymes are compared in Figure II.6. For clarity, some amino-acid side chains that do not contribute to the hydrophobic core have been omitted from Figure II.6(a). It can be seen that in the vicinity of His57, Asp102, Cys42 and Cys58 both main chain and amino acid side chains have extremely similar conformations. However, the hydrophobic interiors of the two enzymes have vastly different chemical character; that of SGPA is predominantly aliphatic, whereas there are more aromatic residues making up the interior of α-lytic protease. Also, it can be seen that as one moves farther from the active site, and more towards the periphery of the enzyme, the structural homology decreases and the polypeptide chains adopt less equivalent paths.

Two adjacent sections through the superimposed N-terminal domains of SGPA and α-lytic protease are represented in parts (b) and (c) of Figure II.6. Insertions in the amino acid sequence for α-lytic protease at 15A and B and at 66 to 83 relative to SGPA are probably the cause of major differences in the polypeptide chain conformations near these sites [Figure II.6(b)]. Residues Ile16, Ile31, Ile66, Val84 and Leu108 in α-lytic protease form an elliptically shaped region of hydrophobic side chains in this portion of the core. The roughly equivalent volume of SGPA is occupied by

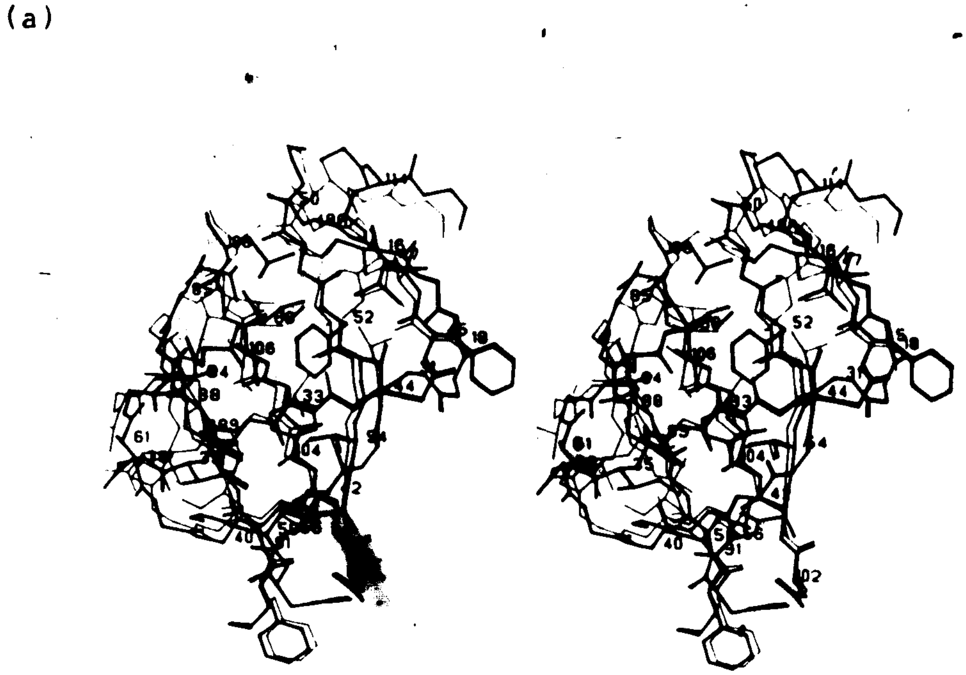

Figure II.6. N-terminal Hydrophobic Cores of α-lytic Protease and SGPA. α-lytic protease is shown in thick lines and SGPA in thin lines. Numbers refer to residues of α-lytic protease and correspond to the alignment in Table II.4. Mainly, only those side chains contributing to the hydrophobic core are represented in (a). The active site residues His57 and Asp102 are at the bottom of the figure in this orientation. (b) A section through the N-terminal cores of α-lytic protease and SGPA; this view is approximately orthogonal to that in (a) from the top. Major differences in polypeptide chain conformations for the two enzymes are evident in this comparison, e.g. from residues 64-85 where there is a five residue insertion in α-lytic protease (Table II.4). Residues of α-lytic protease contributing to the core are: Ile16, Ile31, Phe52, Ile66, Val84, Leu108 and Leu114. (c) The neighboring section of the cores to that in (b). The position of Phe52 is common to both views (b) and (c) for reference. The large volume difference between Phe and Ala at position 52 is only partially compensated by other changes in the region of these residues. The polypeptide chains in this section of both molecules is much more structurally equivalent than that shown in (b).

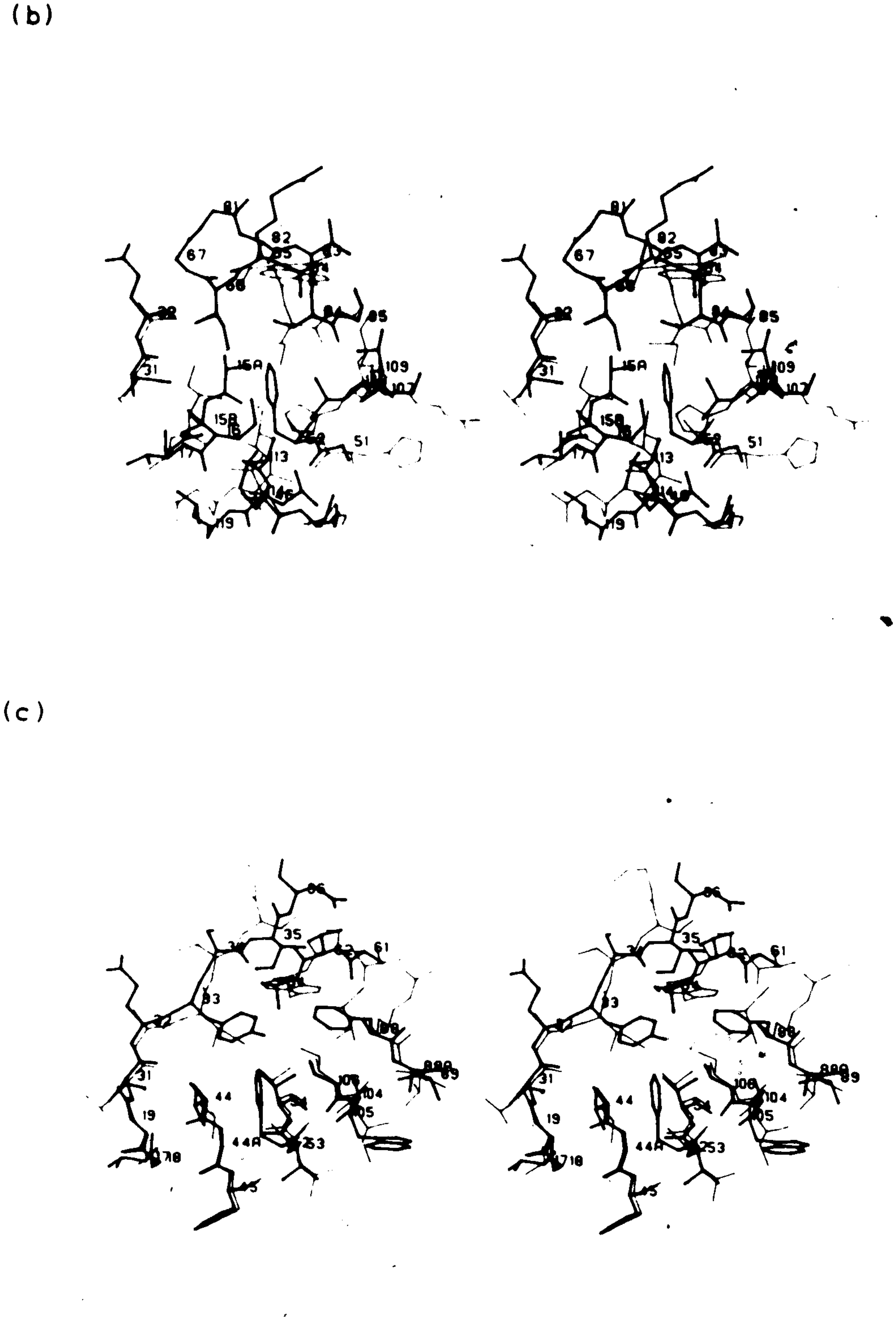

Figure II.6 (continued)

four side chains Ile16, Ile84, His 108 and Asn46, ~~but~~ these residues are not coincident with the above five side chains of α-lytic protease [Figure II.6(b)].

The side chain of Phe52 [seen in both Figs. II.6(b) and II.6(c)] is centrally located in the N-terminal hydrophobic core of α-lytic protease. The side chain of this residue is surrounded by the side chains of Ile16, Tyr33, Val44, Thr54, Ile66, Val84, Val106 and Leu108. The structurally equivalent residue 52 in SGPA is an alanine. The large volume change associated with this amino-acid substitution is accompanied by compensatory sequence changes in residues 33, 44, 84, 106 and 108 and some main chain rearrangements. In spite of these changes, there are two small cavities in SGPA, situated in a volume equivalent to that occupied by the phenyl ring of Phe52 in α-lytic protease. These cavities are about 80-90% the volume of a water molecule and are surrounded by apolar atoms. Similar small cavities have been observed in other protein structures (Lee & Richards, 1971).

The C-terminal domains of α-lytic protease and SGPA have a greater degree of structural equivalence than do their N-terminal domains (Figure II.7). The major conformational differences are in the region of Cys220A (Brayer *et al*., 1979) where α-lytic protease has two insertions relative to SGPA, one of 5 residues, 218-219C, and one of 3 residues at 222B-223 (Table II.4). The extended β-loop, Ala130 to Thr163, in both enzymes is structurally

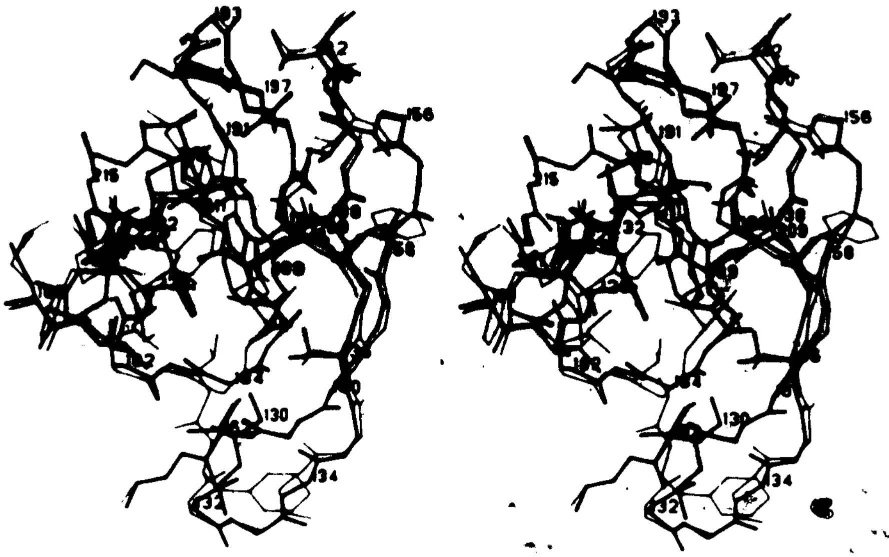

Figure II.7. C-terminal Hydrophobic Cores of α-lytic Protease and SGPA. α-lytic protease is shown in thick lines and SGPA in thin lines. Numbers refer to residues of α-lytic protease and correspond to the alignment in Table II.4. Ser195 is located towards the top of the diagram. The Trp199 to Leu199 difference is centrally located. Compensatory changes involve residues 136, 138, 181, 183, 185, 226 and 228. The conformations of the polypeptide chains in these C-terminal domains are much more similar than in the N-terminal domains of the two enzymes. For clarity some of the side chains not contributing to the hydrophobic core have been omitted.

equivalent. Where there are amino acid differences in this loop, the side chains point to the surface. A high degree of structural equivalence among the bacterial serine proteases is required for this loop as it is the location of the structurally and functionally important residue Arg138. The side chain of Arg138 makes a buried salt bridge to the carboxylate of Asp194 (James *et al*., 1978).

Trp199 is centrally located in the C-terminal hydrophobic core of α-lytic protease. The structurally equivalent residue in SGPA is a leucine (Figure II.7). The relatively large volume difference between the side chains of these two residues is associated with compensatory changes in both sequence and conformation at positions 136, 138, 181, 183, 185, 226 and 228. Two other sequence differences, Ala190 for Met190 and Thr213 for Met213 (Table II.4 and Figure II.7) are in part responsible for the difference in substrate specificity between SGPA and α-lytic protease.

This structural comparison reveals that in spite of pronounced differences in the nature of the amino acid side chains that constitute the hydrophobic interior of these related proteins, the basic folding pattern of the supporting polypeptide chain is preserved.

Hydrogen Bonds

Commonly, a hydrogen bond is defined as a bond between D-H and A (D = donor atom, A = acceptor atom) specifically involving the hydrogen atom (e.g. Pimentel and McClellan, 1960). It has been shown that much of the behavior of hydrogen bonds can be understood in terms of van der Waals and electrostatic interactions (van Duijneveldt-van de Rijdt and van Duijneveldt, 1971; Hagler *et al*., 1974). However, when analyzing a structure, one must resort to either geometrical or energetic criteria with cut-off values to classify an interaction as being a hydrogen bond. In addition, in almost all protein structures, the position of the hydrogen atoms cannot be determined experimentally due to the limited resolution of data that can be obtained. Due to these difficulties, any listing of hydrogen bonds will have a certain degree of uncertainty.

In our analysis we have used strictly geometrical criteria for determining the presence of a hydrogen bond. For nitrogen atoms with planar configurations, hydrogen atoms can be located reasonably well, assuming standard geometry (program of S. Phillips). For cases where the hydrogen atom position is ambiguous, as in hydroxyl groups and quaternary amino groups, the optimal position in the plane of the donor and acceptor groups was assumed. No attempt was made to position the hydrogen atoms on solvent molecules. An interaction was considered to be a possible hydrogen bond if the H···A distance was less than 2.4Å for oxygen acceptors and less than 2.5Å for nitrogen acceptors. For hydrogen bonds between water molecules, and water molecules to acceptors on the protein, a cutoff of 3.4Å for the interatomic distance was used. It should be noted that the use of such geometrical criteria and somewhat arbitrary cutoff values will result in some favourable interactions being rejected and possibly some unfavourable interactions being accepted as hydrogen bonds.

A listing of the possible hydrogen bonds between main chain atoms is given in Table II.6. The table is divided according to the type of secondary structural unit. It can

Table II.6

Hydrogen Bonds Between Main-chain Atoms

N-H	C=O	d(NO) (Å)	d(OH) (Å)	a(NHO) (°)	a(CON) (°)	a(COH) (°)
β-Sheets						
Tyr 33	Cys 42	3.0	2.1	161	153	154
Ser 34	Arg 65	2.9	2.0	158	152	145
Ile 35	Ser 40	2.8	1.9	153	154	153
Asn 36	Thr 62	2.8	1.8	169	155	151
Cys 42	Tyr 33	3.0	2.0	172	171	169
Val 44	Ile 31	2.8	1.8	159	156	149
Val 47	Gly 51	2.9	1.9	169	134	136
Thr 48	Ser241	2.8	1.9	158	156	149
Arg 48A	Thr 49	2.8	1.8	167	155	153
Gly 51	Val 47	2.9	2.1	140	152	164
Phe 52	Val106	3.1	2.2	154	157	148
Val 53	Phe 45	2.9	2.0	159	155	154
Thr 54	Ala104	3.0	2.1	162	141	144
Arg 65	Ser 34	3.0	2.1	157	143	137
Ile 66	Ala 82	2.8	1.9	154	155	156
Val 84	Ala 64	2.9	1.9	176	154	154
Thr 87	Ser107	3.0	2.1	155	157	154
Ala 88A	Trp105	2.8	1.8	161	154	148
Val 91	Arg103	2.9	1.9	163	160	158
Arg103	Val 91	2.9	1.9	166	158	153
Ala104	Thr 54	3.1	2.3	130	137	149
Val106	Phe 52	2.9	1.9	168	165	161
Ser107	Thr 87	2.8	2.0	144	161	172
Leu108	Lys 50	2.8	1.8	173	139	138
Thr109	Val 84	2.9	2.0	163	154	156
Val120B	Val120J	2.8	1.9	154	156	148
Asn120D	Ser120H	2.9	2.0	163	127	133
Val120J	Val120B	2.8	1.8	162	156	150
Val121	Pro120	2.7	1.7	176	134	135
Val136	Gly160	2.8	1.8	167	141	141
Cys137	Ile200	2.8	1.9	161	156	149
Arg138	Gln158	2.7	1.8	154	167	158
Ser139	Ser198	2.9	2.0	154	158	151
Gln158	Arg138	3.0	2.0	160	161	155
Gly160	Val136	2.8	1.8	162	177	170
Ile162	Ala134	2.8	1.9	162	168	162
Ala164	Gln182	3.3	2.4	158	134	138
Val167	Leu180	2.7	1.7	175	145	146
Ala169	Val177	2.8	1.8	165	155	151
Tyr171	Gly175	3.1	2.3	139	126	128
Val177	Ala169	2.8	1.8	165	165	162
Leu180	Val167	2.8	1.9	149	143	150
Thr181	Phe228	2.9	1.9	172	161	163
Gln182	Ala164	2.7	1.7	172	152	150
Gly183	Ser226	2.8	1.9	168	163	159

Table II.6 continued

N-H	C=O	d(NO) (Å)	d(OH) (Å)	a(NHO) (°)	a(CON) (°)	a(COH) (°)
Ile200	Cys137	2.8	1.8	165	154	149
Val212	Glu229	2.8	1.9	154	154	148
Gly215	Leu227	3.1	2.1	163	140	136
Asn217	Ser225	3.0	2.1	150	163	154
Ser226	Gly183	3.0	2.1	154	164	165
Leu227	Gly215	2.6	1.6	174	158	159
Phe228	Thr181	2.9	1.9	168	160	156
Glu229	Val212	2.8	1.9	164	152	157
Leu231	Gln210	3.0	2.1	160	121	123
Ser241	Thr 48	2.8	1.8	165	161	156
Val243	Ser 46	3.1	2.1	167	137	141
Bends						
Ala 39	Ile 35	2.9	2.0	153	130	123
Thr 49	Arg 48A	3.0	2.1	150	124	128
Cys 58	Ala 55	3.1	2.2	152	114	105
Ala 61	Thr 59A	2.9	2.0	154	144	141
Ala 82	Ile 66	3.1	2.2	152	119	118
Gln112	Thr109	3.2	2.2	166	123	118
Ser120H	Asn120D	2.8	2.0	148	135	135
Ala134	Ala131	3.0	2.0	160	145	142
Gly175	Tyr171	3.0	2.2	143	116	105
Asp194	Gly191	2.9	2.0	157	150	145
Gly197	Asp194	3.0	2.1	149	144	136
Gly207	Thr201	2.9	2.0	162	120	116
Gly219C	Gln219	2.9	1.9	154	122	116
Gln223	Pro222A	2.8	1.9	152	127	118
Arg224	Ala222B	3.1	2.1	166	115	111
Ile234	Leu231	3.0	2.3	121	122	107
Asn101	Pro 95	3.1	2.2	139	82	95
α-helices						
Leu235	Leu231	3.1	2.1	167	168	164
Ser236	Gln232	3.1	2.2	166	162	158
Gln237	Pro233	2.9	2.0	152	143	134
Tyr238	Ile234	3.0	2.1	160	154	150
Gly239	Leu235	2.9	2.3	113	145	129
Others						
Ile 16	Thr113	2.7	1.8	176	165	165
Gly 18	Arg120A	2.8	1.8	166	145	150
Gly 19	Val 44	2.8	1.8	170	142	138
Ser 43	Ser195	2.9	2.0	157	163	163
Gly 44A	Val 53	2.9	2.1	127	164	159
Asn 60	Phe 88	2.8	1.8	166	143	141
Ala 64	Gly 85	2.9	1.9	173	142	142
Phe 88	Ala 61	2.8	1.9	159	159	151
Gly100	Ala176	2.9	2.0	151	148	142
Trp105	Ala 89	3.1	2.2	157	158	155

Table II.6 continued

N-H	C=O	d(NO)' (Å)	d(OH) (Å)	a(NHO) (°)	a(CON) (°)	a(COH) (°)
Thr113	Ala 15A	3.0	2.0	176	162	164
Leu119	Ile 16	3.2	2.4	137	156	154
Arg120A	Leu119	2.9	2.1	144	84	92
Ala120C	Gly 18	2.8	1.8	177	153	152
Arg122	Gly207	2.7	1.7	169	138	136
Gly133	Ile162	2.8	1.9	154	144	136
Gly156	Gly140	3.0	2.3	128	146	136
Thr163	Gln182	2.8	1.8	163	147	141
Arg178	Gly100	2.9	1.9	158	153	152
Asn184	Thr161	3.0	2.1	143	161	149
Arg192	Gly219C	2.9	1.9	156	170	162
Gly196	Met213	2.8	1.8	173	133	131
Trp199	Gly211	2.9	2.1	136	147	157
Thr201	Gln208	3.1	2.1	165	139	135
Ala209	Gly123	2.8	1.8	160	147	153
Gln210	Trp199	2.8	1.8	166	143	143
Met213	Gly197	2.8	1.9	162	149	143
Ser214	Leu227	2.8	1.8	155	151	152
Val218	Gly216	3.0	2.2	135	91	104
Gln219	Asn219D	2.8	1.8	165	133	129
Asn220	Met190	3.0	2.1	149	169	160
Leu240	Leu235	2.8	1.9	149	147	142
Gly245	Val121	2.8	1.9	158	170	163

'd,distance; a,angle

be seen that the protein is mainly made up of β-sheets with only a short α-helix from Leu231 to Gly239 and a very short 3_{10} helix from Pro223 to Arg224. The turns were classified according to the criteria given by Venkatachalam (1968) and are summarized in Table II.7. A type III (3_{10}) turn occurs at the N-terminus of the α-helix between residue Leu231 and Ile234. The carbonyl group of Leu231 also accepts a hydrogen bond from Leu235 NH thus forming the start of an α-helix [Figure II.5(b)]. It is more common for a 3_{10} conformation to occur at the C-terminal end of a helix (Richardson, 1981). There is a single reverse open turn (Ramachandran &

Table II.7

ϕ-ψ Angles in Turns

		ϕ_2(°)	ψ_2(°)	ϕ_3(°)	ψ_3(°)
Standard type I		-60	-30	-90	0
Residues	201-207	-63	-21	-85	1
	219-219C	-60	-24	-81	1
	222B-224[1]	-61	-23	-102	1
Standard type I'		60	30	90	0
Residues	66-82	54	39	85	-15
Standard type II		-60	120	80	0
Residues	59A-61	-58	139	77	-1
	131-134	-58	135	98	-16
	191-194[2]	-56	133	106	-22
	194-197[2]	-49	142	88	-21
Standard type II'		60	-120	-80	0
Residues	48A-49	68	-132	-75	-12
	120D-120H	64	-124	-92	-6
Standard type III		-60	-30	-60	-30
Residues	55-58	-63	-35	-67	-13
	109-112	-61	-23	-72	-17
	171-175	-53	-39	-75	-20
	222A-223[1]	-54	-35	-61	-23
	231-234	-44	-54	-59	-34
Standard type III'		60	30	60	30
Residues	35-39	49	41	56	25
Reverse open turn					
Residues	94-101	-81	-155	93	-71

[1] Helical (3_{10})
[2] Far from standard values

Mitra, 1976) from Phe94 to Asn101, with the only cis-proline of the molecule at position 95 [Figure II.8(a)]. This type of turn is also observed in the homologous region of SGPA as depicted in Figure II.8(b). The reverse open turn is distinct from the type VI turn that has a cis-proline at the i+2 position (Richardson, 1981).

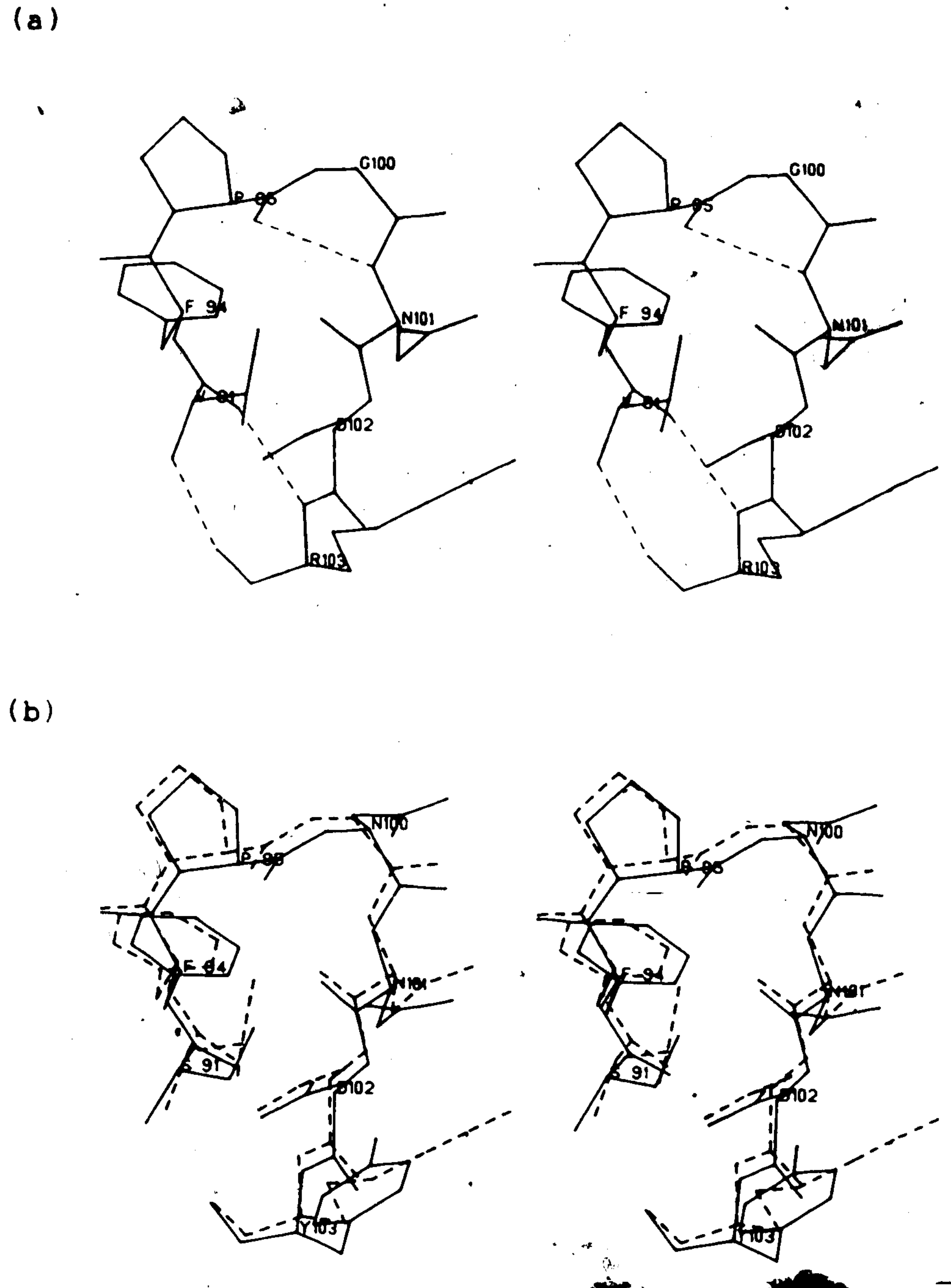

Figure II.8. Reverse Open Turns. (a) The r[illegible]ound cis-Pro95 that is in a reverse open turn c[illegible]mation (Ramachandran & Mitra, 1976). (Hydrogen bonds are shown by broken lines.) (b) The homologous region in SGPA is shown (solid line and labelled residues) superimposed on the equivalent residues of α-lytic protease (broken lines). The superposition was done using all the alpha carbons in each molecule.

In addition to the hydrogen bonds listed in Table II.6, there is a group of main chain-main chain interactions that fall within the limits that we have set for a hydrogen bond but have neverthelbss not been included in the table. These involve residues in what is known as a C_5 conformation. In these residues the N-H and C=O groups are almost coplanar with H···O distances of about 2.3Å and N-H··O angles of about 100°. In most cases these residues are part of a β-sheet and presumably the C_5 conformation is imposed on them by the secondary structure. However, there are other cases in which the reasons for adopting such a conformation are not clear. It is not obvious what the nature of the interaction is in a C_5 conformation and whether it is justified to classify it as a hydrogen bond. The existence of such interactions has been inferred from IR spectra of model peptides in CCl_4 (Avignon *et al*., 1969).

We examined in detail the angular parameters of the main chain hydrogen bonds. Figure II.9 shows the angle at the hydrogen atom as a function of the hydrogen to the acceptor distance. The trend seen in the plot is similar to that observed in small molecule structures (Olovsson & Jonsson, 1976). The linear configuration at the shorter hydrogen to acceptor distances would minimize the unfavourable contact between the nitrogen and oxygen atoms (Hagler *et al*., 1974).

In human lysozyme, Artymiuk and Blake (1981) analyzed the polar angles of the hydrogen bond at the acceptor atom.

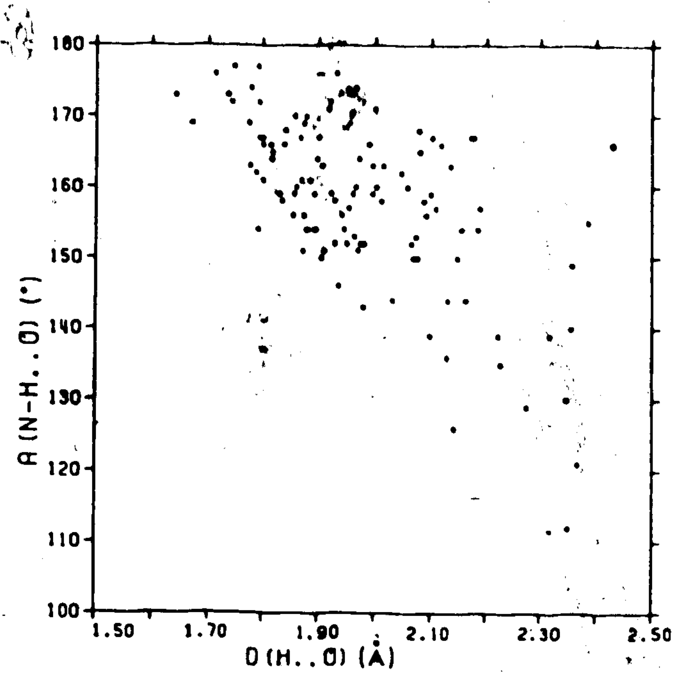

Figure II.9. Hydrogen Bond Angles and Distances. Angle N-H···O as a function of the H···O distance for hydrogen bonds between main chain atoms.

They defined two angles β and γ, where β is the angular position of the hydrogen atom with respect to the plane of the peptide group and γ is the azimuthal angle in the peptide plane with the zero defined by the C=O direction (see Figure II.10 inset).

They observed for human lysozyme a clustering of points corresponding to residues in α-helices, turns and β-sheets into different regions of a β-γ plot. The corresponding plot for α-lytic protease is shown in Figure II.10. A similar clustering of secondary structural units can be seen. In addition, the different types of turns cluster into separate regions. The types I and III turns are in a

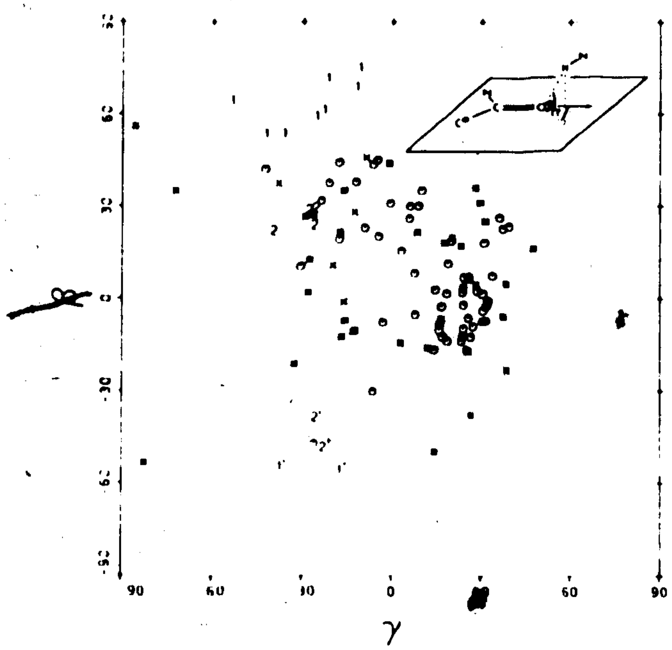

Figure II.10. Angular Parameters at the Main-chain Carbonyl Oxygen. Plot of the polar angles β and γ (defined in the inset) at the carbonyl oxygen of the main chain (Artymiuk & Blake, 1981). The values for the hydrogen bonds involving the main chain atoms are given with the following symbols: (O) β-sheet, (x) α-helix, (*) others. The hydrogen bonds in turns are designated by the corresponding turn type number (Venkatchalam, 1968). Type III turn is included with the type I (shown as 1) and the type III' is included with the type I' (shown as 1'). Type II and II' turns are shown as 2 and 2' respectively.

region (60, -30) in β, γ and type II turns are around (30, -30). All the turns I', II' and III' cluster together at around (-50, -30).

A corresponding pair of polar angles can also be defined at the amide hydrogen atom of peptides. β' is the angular position of the acceptor atom with respect to the peptide plane and γ' is the azimuthal angle in that plane with the zero defined by the N-H direction. The plot of

β'-γ' is shown in Figure II.11. Here again the different secondary structural units seem to be localized into different regions. However, the various types of turns do not cluster into separate groups and the α-helix does not seem to cluster at all. The angular parameters for the helix may not be representative since there is only a short α-helical segment in this molecule.

The possible hydrogen bonds involving the side chains are tabulated in Tables II.8 and II.9. The listing on Table II.8 is much more tentative as it involves donor groups whose hydrogen atom positions are ambiguous (amino terminus,

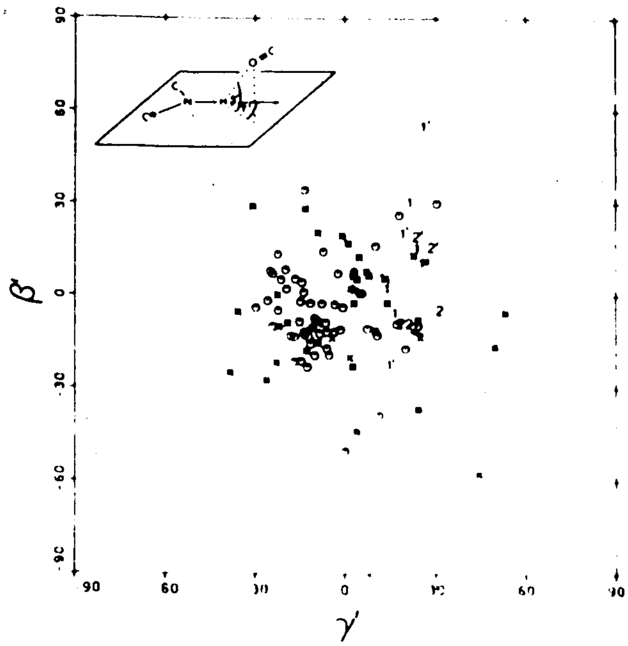

Figure II.11. Angular Parameters at the Main-chain Amide Nitrogen. Plot of the polar angles β' and γ' (defined in the inset) at the amide hydrogen of the main chain. The labelling of the various secondary structural units is as in Figure II.10.

Table II.8

Side-chain Hydrogen Bonds with Ambiguous Hydrogen Atom Positions

Donor	Acceptor	d(DA)[1] (Å)	d(AH) (Å)	a(CDA) (°)	a(DHA) (°)	a(CAD) (°)
Ala 15A N	Gln112 OE1	2.8	1.8	113	175	146
Tyr 33 OH	Thr 54 OG1	2.8	1.8	117	169	106
Ser 34 OG	Ser 34 O *	3.2	2.4	82	141	63
Ser 43 OG	Gly193 O	3.0	2.0	131	149	137
Ser 46 OG	Leu119 O	3.0	2.1	92	154	129
Thr 48 OG1	Thr 48 O *	3.1	2.3	83	141	67
Lys 50 NZ	Ser110 O	2.9	1.9	104	171	139
Lys 50 NZ	Gln112 O	2.6	1.6	116	171	169
Thr 54 OG1	Ser 43 O	3.2	2.2	128	154	147
Thr 87 OG1	Thr 87 O *	3.0	2.1	86	145	70
Ser107 OG	Ser107 O *	3.2	2.4	76	133	75
Thr109 OG1	Thr109 O *	3.3	2.4	85	145	59
Thr125 OG1	Gln208 OE1	3.2	2.3	89	150	109
Ser139 OG	Ala120C O	2.7	1.7	107	174	135
Thr142 OG1	Arg192 O	2.7	1.7	100	164	130
Thr143 OG1	Asp194 OD1	2.6	1.6	114	173	126
Thr163 OG1	Gln182 O	3.2	2.2	116	172	97
Lys165 NZ	Glu129 OE2	3.0	2.0	109	178	124
Lys165 NZ	Ala130 O	2.7	1.7	99	162	143
Tyr171 OH	Ser214 OG	2.8	1.8	113	176	113
Ser195 OG[2]	His 57 NE2	3.1	2.2	85	144	-
Ser198 OG	Gly 44A O	2.6	1.7	119	166	140
Thr201 OG1	Thr201 O *	3.2	2.3	86	146	62
Ser214 OG	Asp102 OD1	2.6	1.6	114	173	127
Ser225 OG	Gln182 OE1	3.0	2.0	102	169	120
Ser226 OG	Ala185 O	2.6	1.6	101	166	165

[1] d,distance; a,angle; D,donor; A,acceptor; C,carbon.
*, an intra-residue interaction.
[2] Not likely to be a hydrogen bond. See text and Table II.11.

Lys, Ser, Thr and Tyr). The hydrogen atoms were placed in the optimal position as described above in the Definition section, and the distances and angles in the table are included to give an idea of the geometry of the interaction. Most of the groups listed in this table also make additional hydrogen bonds with solvent molecules. The possible hydrogen bonds listed in Table II.9 on the other hand involve hydrogen atoms that can be located assuming standard geometry and thus the interactions can be examined with more confidence.

An uncertainty concerning the protonation state of the single histidine residue (His57) arises. Since the crystal was grown at pH 6.1 that is at or near the normal pK_a of histidine (nominally 6.5), the equilibrium favors a partially protonated (i.e. partially positively charged) state of the imidazole ring. From NMR data, Westler *et al.* (1982) have reported pK_a values from 5.9 to 6.5 for this residue of α-lytic protease, depending on the sample preparation. In any case, ND1 of the histidine is protonated; a strong hydrogen bond exists from this atom to Asp102 OD2 (Brayer *et al.*, 1979). At one time there was a controversy about whether the proton was on the histidine or on the aspartate (Hunkapiller *et al.*, 1973). The question now seems to have been resolved in favour of the proton being on the histidine (Bachovchin *et al.*, 1981). The hydrogen bonding interactions found in this study indicate that NE2 of the histidine is also protonated, making the imidazole group positively charged (see the section 'Charge Interactions').

In Tables II.8 and II.9, there are several intra-residue interactions (marked by *'s) which fall within our hydrogen bonding criteria. A five-membered ring-like structure occurs when there is a close approach of the hydroxyl

Table II.9

Side-chain Hydrogen Bonds without Ambiguous Hydrogen Atom Positions

Donor	Acceptor	d(NA)' (Å)	d(AH) (Å)	a(NHA) (°)	a(CAN) (°)	a(CAH) (°)
Arg 48A NE	Tyr238 O	3.3	2.4	146	170	168
Ala 55 N	Thr 54 OG1	2.9	2.1	128	80	96
Gly 56 N	Asp102 OD2	2.8	2.3	108	152	135
His 57 N	Asp102 OD2	2.8	1.9	153	145	143
His 57 ND1	Asp102 OD1	2.7	1.7	166	115	110
Thr 59A N	Thr 59A OG1 *	2.6	2.3	100	68	90
Thr 62 N	Asn 36 OD1	2.9	1.9	170	135	137
Arg 90 NH1	Gly 56 O	3.0	2.1	149	136	140
Arg 90 NE	Gly 56 O	3.2	2.3	148	127	131
Asp102 N	Glu229 OE2	2.9	1.9	169	117	114
Arg103 NH2	Glu229 OE2	3.0	2.2	139	100	88
Arg103 NH1	Gln237 OE1	2.9	1.9	153	171	166
Arg103 NE	Glu229 OE2	2.9	1.9	155	133	134
Gln112 NE2	Thr109 OG1	2.8	1,8	172	160	157
Phe120I N	Ser120H OG	3.0	2.3	126	75	91
Thr125 N	Gln208 OE1	2.9	2.0	168	141	145
Ala130 N	Gln210 OE1	2.9	1.9	178	120	121
Arg138 NH2	Gly156 O	3.1	2.4	127	163	148
Arg138 NH1	Thr143 OG1	2.8	2.1	120	154	140
Arg138 NH1	Cys189 O	2.8	2.0	140	142	155
Arg141 N	Ser 43 OG	3.0	2.1	162	132	138
Arg141 NH2	Thr142 OG1	2.9	1.9	160	125	131
Arg141 NH1	Glu 32 OE2	3.0	2.1	140	141	137
Arg141 NE	Glu 32 OE2	2.8	1.9	150	125	116
Thr142 N	Asp194 OD2	2.9	2.0	148	143	148
Thr143 N	Asp194 OD2	3.1	2.2	164	105	101
Gln158 NE2	Thr143 O	2.9	1.9	163	153	148
Thr161 N	Asn184 OD1	2.9	1.9	161	136	132
Glu174 N	Glu174 OE1 *	2.7	1.8	145	94	107
Gln182 NE2	Thr163 OG1	2.8	1.9	162	125	118
Met190 N	Asn220 OD1	2.9	2.0	148	151	149
Gly191 N	Asp194 OD1	2.7	1.8	147	132	122
Ser198 N	Ser198 OG *	2.6	2.2	103	69	85
Gln208 N	Thr201 OG1	3.0	2.1	153	151	155
Gln210 NE2	Thr125 O	2.8	1.8	162	149	150
Gln210 NE2	Gln208 O	3.0	2.0	173	108	108
Asn217 N	Asn217 OD1 *	2.8	2.3	117	83	95
Asn217 ND2	Gln223 O	3.0	2.1	153	153	145
Gln219 NE2	Asn217 O	2.8	1.9	161	131	130
Asn219D N	Asn219B OD1	3.1	2.2	165	128	129
Asn220 ND2	Ser226 OG	3.1	2.1	168	144	141
Asn220 ND2	Asn217 OD[illegible]	3.0	2.0	165	156	152
Cys220A N	Asn219D OD[illegible]	2.8	1.9	145	129	118
Arg224 NH2	Ile222 O	2.9	2.1	132	162	159
Arg224 NH1	Asn184 O	3.1	2.1	170	1[illegible]	161

(Table II.9 continued)

Donor	Acceptor	d(NA)' (Å)	d(AH) (Å)	a(NHA) (°)	a(CAN) (°)	a(CAH) (°)
Arg224 NE	Ile222 O	3.0	2.2	134	145	140
Ser225 N	Asn217 OD1	2.9	1.9	169	117	121
Arg230 NH2	Glu129 OE1	3.1	2.2	146	139	148
Arg230 NE	Glu129 OE1	3.1	2.2	154	103	98
Gln232 NE2	Ser124 O	2.9	2.0	166	155	156

'd,distance; a,angle; A,acceptor.
*,an intra-residue interaction.

oxygen of serines and threonines to the main chain N-H (Figure II.12). This conformation is similar to the C₇ conformation discussed above, but it lacks the favourable dipole interaction present in the latter. Six-membered ring-like structures can result from an interaction either between the hydroxyl hydrogen atom of serines and threonines to the main chain carbonyl oxygen atom (not shown), or from the main chain N-H to the side chain carbonyl group of aspartate or asparagine residues (Figure II.12). For cases involving a serine or a threonine, the actual hydrogen atom position is not known so that the existence of such an interaction is difficult to assess.

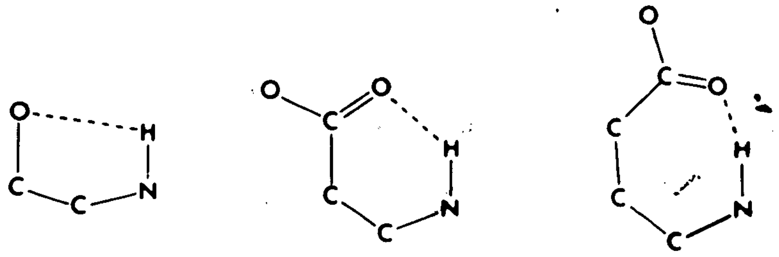

Figure II.12. Intraresidue Interactions. Possible interactions with 5-, 6-, and 7-membered rings.

The last type involves a hydrogen bond between the main chain N-H and the side chain carbonyl oxygen atom of glutamate or glutamine. There is one example of this in α-lytic protease at Glu174 with χ_1 = -56° and χ_2 = 77°. A similar conformation of a glutamate side chain has been observed at the reactive site of the turkey ovomucoid inhibitor domain III (Read *et al.*, 1983). The torsional angles for Glu19I in that structure are χ_1 = 61°, χ_2 = -62°. The most commonly observed conformation for a glutamate side chain has torsional angles χ_1 = -60°, χ_2 = 180° (Janin *et al.*, 1978).

More than two thirds of the interactions listed in Tables II.8 and II.9 occur between side chain and main chain atoms. Of these, the ones involving aspartate and asparagine residues are of particular interest in their role in determining the structure of a protein. There is a high probability for aspartates and especially asparagine residues to be located in a reverse turn (Crawford *et al.*, 1973). In addition these residues can form a reverse turn-like structure, the Asx turn (Rees *et al.*, 1983). In these structures, the Asx side chain CO group takes the role of the main chain carbonyl group of the ith residue in a normal turn. In α-lytic protease, Asx turns are observed at Asn219B and at Asn219D. Asx residues are found frequently also at the end of β-sheets (Richardson, 1981). Here again the side chain CO can replace a main chain carbonyl group resulting in the termination of the β-sheet. Unlike the Asx

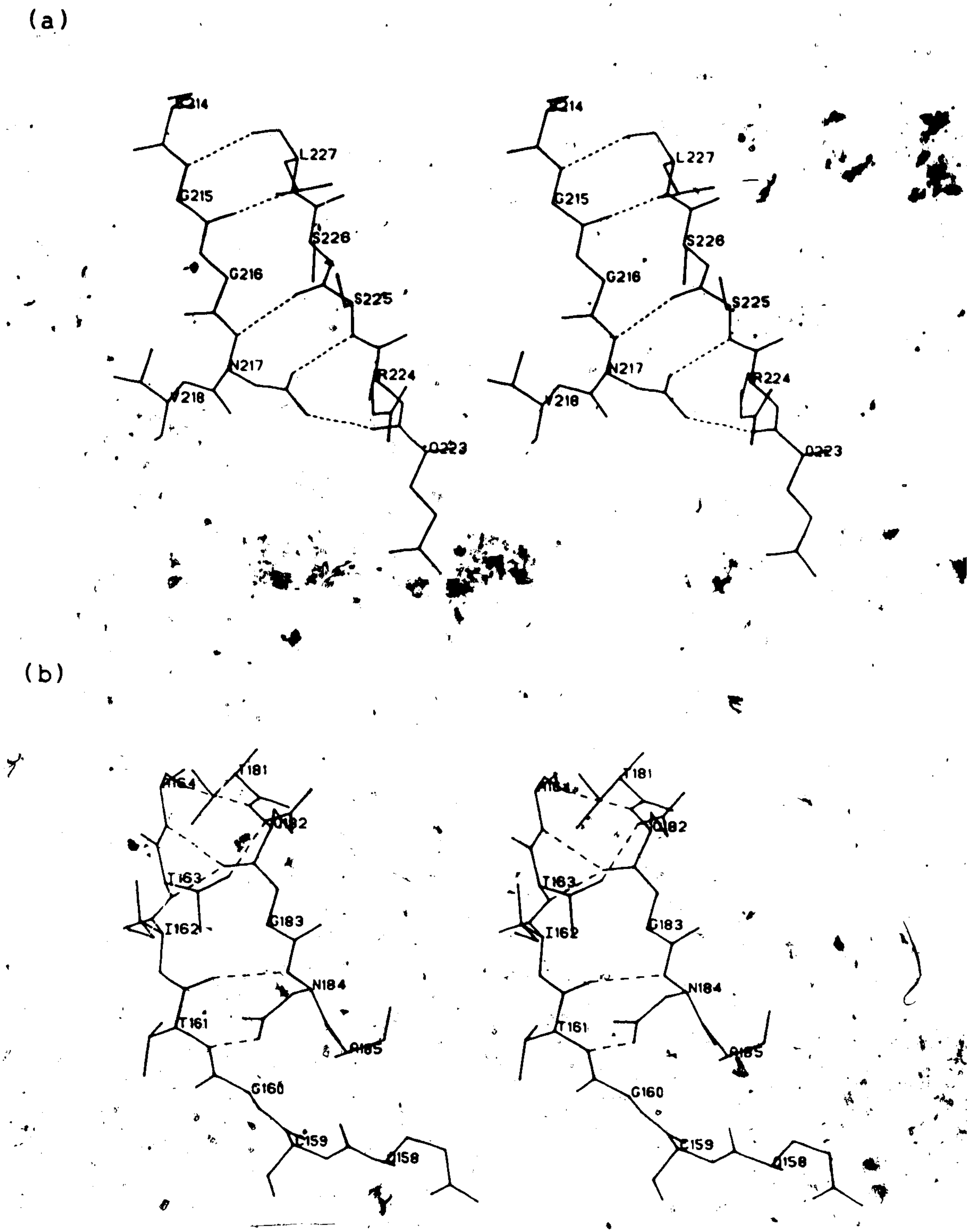
(a)
214
L227
G215
S226
G216
S225
N217
R224
V218
Q223
(b)
A164
T181
Q182
I163
I162
G183
N184
T161
Q185
G160
L159
Q158

Figure II.13

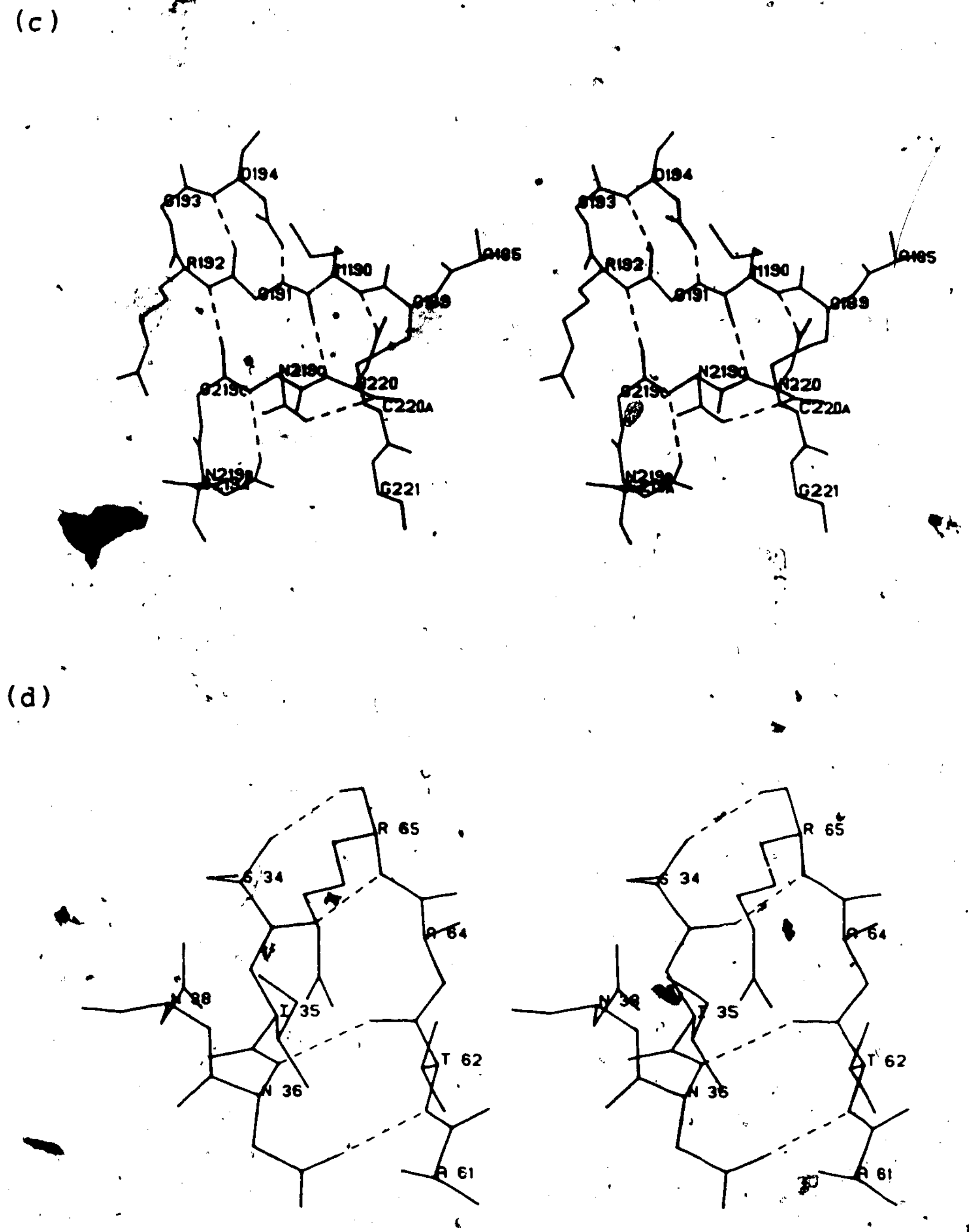

Figure II.13. β-sheet Termination by Asx Residues. (a)-(d) Regions where a β-sheet conformation of the polypeptide chain is terminated by hydrogen bonding to the side chain of Asx residues. The broken lines indicate the hydrogen bonds. (c) also shows two examples of Asx turns (Rees *et al*., 1983), at Asn219B and at Asn219D.

turn, that has the same number of atoms as a reverse turn, the termination of a β-sheet by an Asx residue involves an extra atom (CB) as compared to a β-sheet interaction. The insertion of the additional atom could make that type of interaction less favorable. However the numerous occurrences of this structure in α-lytic protease, at Asn36, Asn184, Asp194, Asn219B and Asn220 (Figures II.13(a)-(d)), indicate that it is an important structural element.

There is a somewhat special hydrogen bond not listed in the tables. This is the interaction between the side chains of Trp199 and Met213. The distance from Trp199 HE1 to Met213 SG is 2.5Å while the distance from the NE1 of the tryptophan side chain to SG is 3.4Å. The angle at the hydrogen is 150°. A cutoff of 2.8Å for H···S distance was used for hydrogen bonds to sulfur acceptors.

There are relatively few direct hydrogen bonds between protein molecules in the crystal and these are given in Table II.10. Most of the intermolecular interactions involve bridging solvent molecules.

Charge Interactions

α-lytic protease is a basic protein with twelve arginines, two lysines, four glutamates, two aspartates, one histidine and the free amino and carboxy termini. All the negatively charged groups interact either directly, or indirectly via solvent bridges, to positively charged groups. The C-terminus forms a hydrogen bonded bridge

Table II.10

Intermolecular Hydrogen Bonds

Donor	Acceptor	d(NA)' (Å)	d(AH) (Å)	a(NHA) (°)	a(CAN) (°)	a(CAH) (°)
Asn 15B ND2	Pro233 O	2.9	2.4	114	113	115
Asn 15B ND2	Ser236 OG	2.8	1.9	167	148	144
Al[illegible]8C N	Glu174 O	2.7	1.7	172	140	139
Ar[illegible]A NH2	Gln237 O	2.8	1.9	158	124	125
A[illegible]A NE	[illegible]r236 O	2.9	2.0	141	171	174
Ser120G N	A[illegible]n219B O	2.8	1.8	162	125	120
Ala131 N	Asn 38 O	2.9	1.9	168	165	16[illegible]
		d(OA)' (Å)	d(AH) (Å)	a(COA) (°)	a(OHA) (°)	a(CAO) (°)
Ser120H OG	Ser241 OG	3.3	2[illegible]	144	174	150

'd,distance; a,angle; A,acceptor

through Wat138 to Arg122NE. Three of the four glutamates interact directly with arginine residues while the remaining Glu174 forms a hydrogen bonded bridge through Wat97 and Wat138 to Arg122 of a neighboring molecule. Both of the aspartate residues are buried. Asp102, which is a member of the catalytic triad Asp·His·Ser, is stabilized by hydrogen bonding to His57ND1 as well as to Gly56N, His57N, and Ser214OG. The other aspartate, Asp194, is also a very important residue as it is conserved in the family of serine proteases with the active site sequence Gly-Asp-Ser-Gly-Gly. The side chain of Asp194 is within 3.4Å of the guanidinium group of Arg138, but there is no hydrogen bond between these groups. Instead the interaction is mediated by hydrogen bonds to Thr143OG (Figure II.14). There is an intricate hydrogen bonding network in the vicinity of these residues,

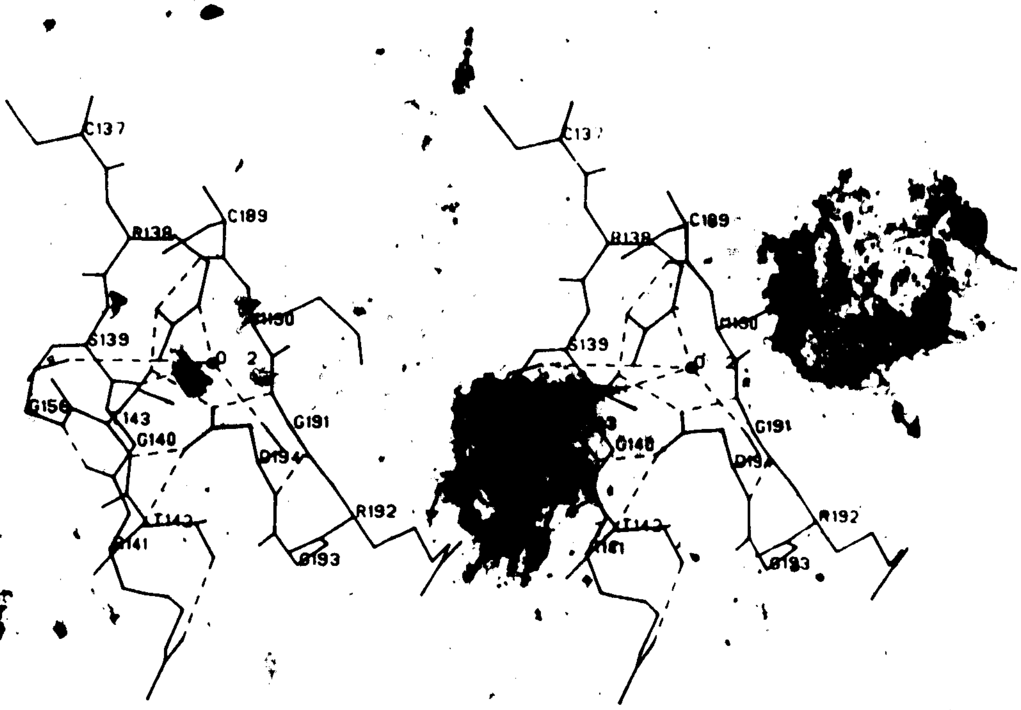

Figure II.14. Environment of Arg138 and Asp194. Hydrogen bonding involving the buried charged residues Arg138 and Asp194. The broken lines show the hydrogen bonds. O2 is a tightly bonded internal water molecule. The interaction between the two charged residues is mediated by Thr143.

including interactions with a strongly bound internal water molecule, Wat2. In the pancreatic serine proteases this aspartate residue interacts with the N-terminus and is important in the activation of the zymogen (Huber & Bode, 1978).

The two sulfate ions in the structure make a number of intermolecular interactions. The possible hydrogen bonds involving these molecules are listed in Table II.11 and Figures II.15(a) and (b) show their environments. Sul 1 is located in the binding site of the enzyme and receives hydrogen bonds from His57NE2 and Ser195OG [Figure II.15(a)].

Table II.11

Hydrogen Bonds to Sulfate Ions

			d(OD)[1] (Å)	d(OH) (Å)	a(OHD) (°)	a(SOD) (°)
Sul 1	S-O1	Wat 87	2.8			102
		Wat 25	2.5			159
	S-O2	Arg122 NH1[2]	2.9	2.[illegible]	123	82
		Wat 51	2.9			146
		Wat108	3.2			121
	S-O3	Wat 87	3.1			90
		Wat 76	3.2			130
		His 57 NE2	2.8	1.9	149	124
		Arg122 NH1[2]	2.9	2.1	135	83
		Ser195 OG	3.1	2.2	151	80
	S-O4	Gly193 N	2.6	1.6	170	120
		Ser195 OG	2.6	1.6	170	[illegible]
Sul 2	S-O1	Wat118	2.9			[illegible]
		Asn 15B N	3.1	2.1	168	123
	S-O2	Wat 58	2.7			116
		Arg230 NH1[2]	3.0	2.0	164	115
	S-O3	Wat104	2.7			126
		Arg230 NH2[2]	2.9	1.9	173	[illegible]
	S-O4	Wat 31	2.5			132

[1] d,distance; a,angle; D,donor

[2] Symmetry related molecule.

In addition, Arg122 from an adjacent molecule donates a hydrogen bond to this ion. The existence of a good hydrogen bond from His57NE2 is a strong indication that this residue is positively charged and thus cannot accept a hydrogen bond from Ser195OG.

The other sulfate ion, Sul 2, is bound near the N-terminus of the protein [Figure II.15(b)]. There is no direct contact with the N-terminal alanine but a long bridging interaction is made via Wat124 and Wat58. The distance between these water molecules is 3.6Å. There is a direct charge interaction from the ion to Arg230 on a neighboring

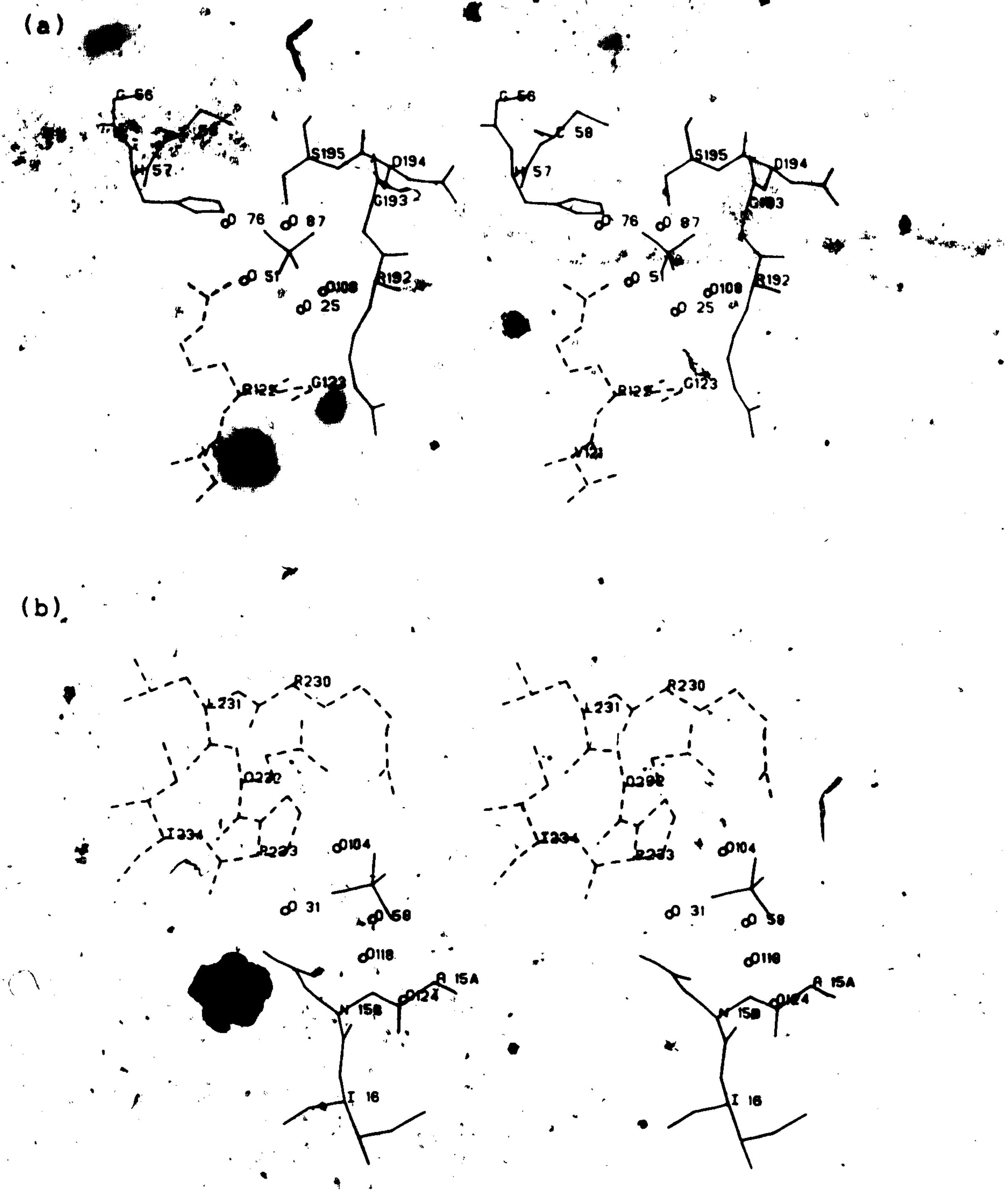

Figure II.15. Environment of the Sulfate Ions. Broken lines indicate amino acids from a neighbouring molecule. Water molecules are given single letter codes, O. The hydrogen bonding interactions are listed in Table II.11.
(a) Sul 1 bound near the active site.
(b) Sul 2 bound near the N-terminus.

molecule.

The arginine and lysine residues that do not have a counterion nearby are all exposed to the bulk solvent. Of these, Arg192 seems to have two distinct side chain conformations in the crystal (Figure II.16). The alternate conformation may be less favourable as the side chain would become buried in the packing interface and it would then be positioned close to Arg122 of a neighboring molecule.

Solvent

Assuming a solvent density of 1.0 g/ml and a partial specific volume of the protein of 0.73 ml/g, it is estimated that there are about 890 water molecules in an asymmetric unit of the crystal of α-lytic protease. Thus the 156

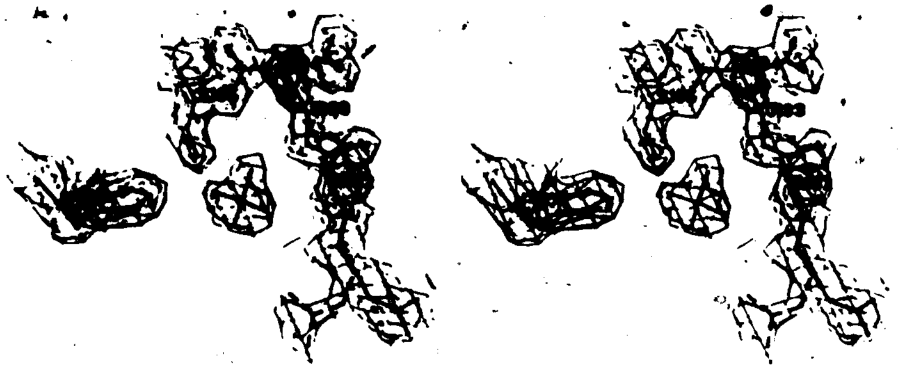

Figure II.16. Disordered Arg192 Residue. The multiple conformation of Arg192 is shown superimposed on the electron density map calculated with coefficients $(2|F_o|-|F_c|, \alpha_c)$. The solid lines indicate the conformation in the refined structure whereas the dashed lines show the possible alternate conformation.

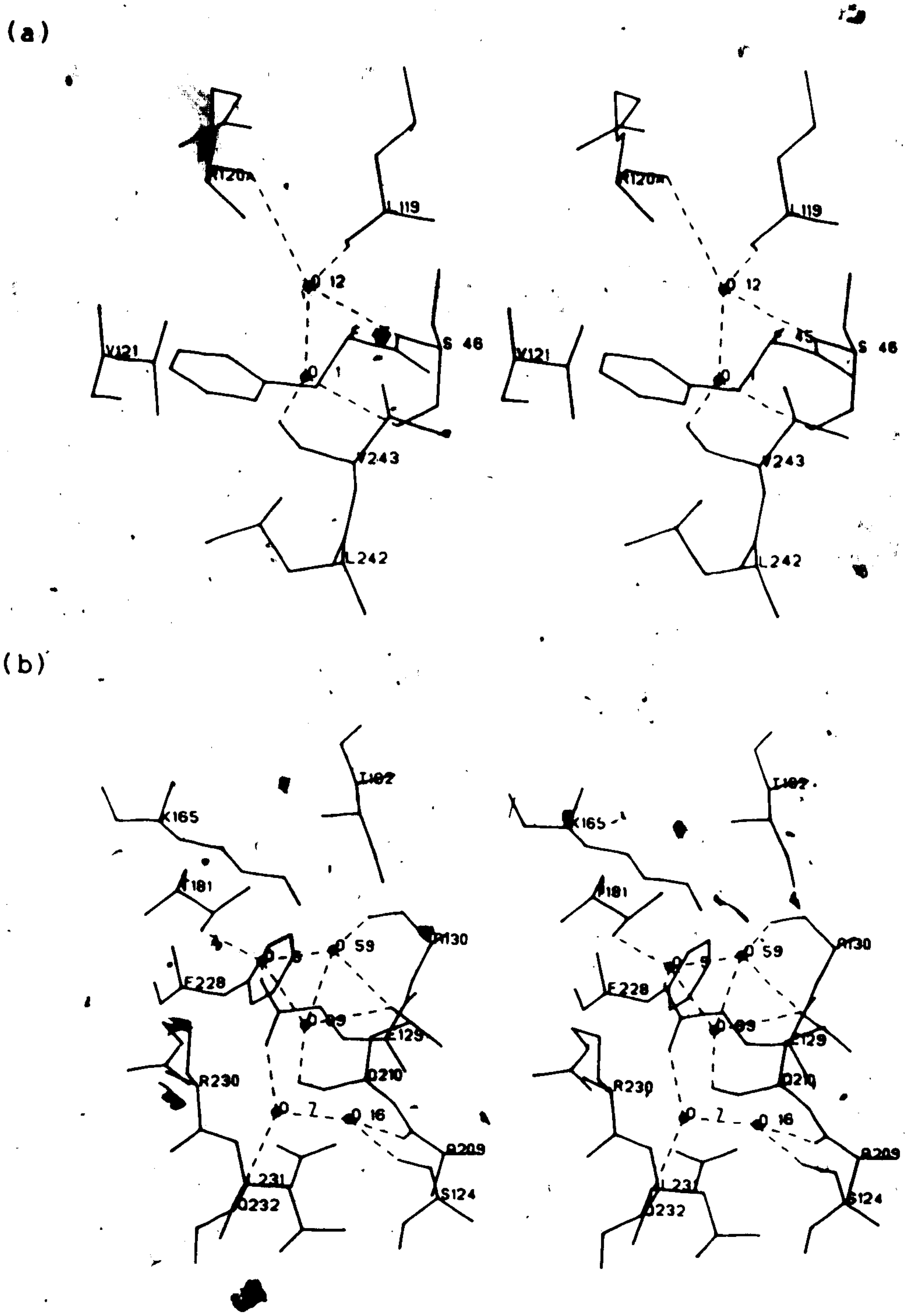

Figure II.17. Oligomeric Internal Water Molecules. The broken lines indicate hydrogen bonds involving the water molecules. A dimer between Wat1 and Wat12 is shown in (a) while a trimer (Wat5, Wat59, Wat99) and a dimer (Wat7, Wat16) are shown in (b).

ordered water molecules represent only 18% of the total solvent content. In contrast, the homologous bacterial serine proteases SGPA and SGPB have 235 and 249 solvent sites respectively (Sielecki & James, 1981; Sawyer *et al.*, unpublished). These results indicate that there may be more solvent sites yet to be found in the α-lytic protease crystal. However, the number of ordered solvent sites will also depend on the crystal packing and the thermal motion of the protein.

There are nine internal water molecules in the α-lytic protease structure. Two of these are isolated molecules inside the protein, while the remaining ones occur as two dimers and one trimer that are hydrogen bonded together (Figure II.17). As expected, most of these water sites refine with very high occupancies and low temperature factors. Their quality factors are among the highest

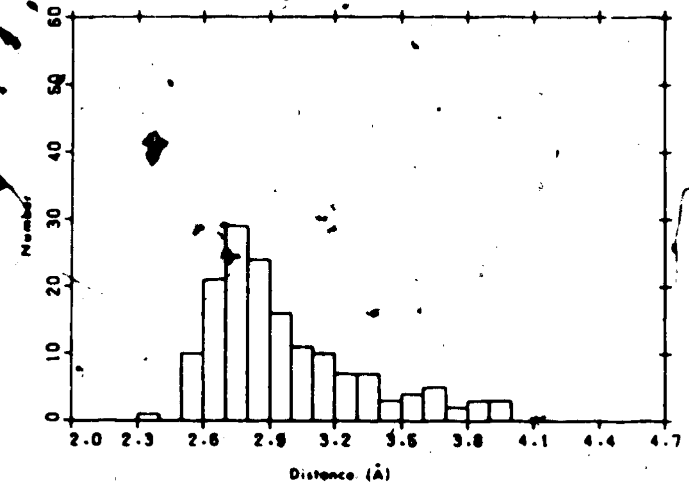

Figure II.18. Water-Protein Contacts. Histogram of the closest water-protein contact distances.

twenty. Two exceptions are two of the water molecules in the trimer that are located at number 59 and 99 in the quality factor ordering.

The analysis of the general features of the solvent structure showed results similar to those found previously in other structures (Blake *et al*., 1983; Watenpaugh *et al*., 1978; Finney, 1979). Figure II.18 shows a histogram of the closest contact distances from solvent to protein. It shows a peak around 2.8Å representing an ideal hydrogen bonding distance. There is no significant evidence of a second shell of ordered water molecules further away from the protein. One water is located 2.4Å from the main chain carbonyl oxygen of Ala202, and there are nine other water molecules with contact distances less than 2.6Å. These were

Table II.12

Classification of Water Molecules by their Hydrogen Bonds

	No.
Total number of solvents per asymmetric unit	156
Solvents with only 1 hydrogen bond to protein	71
Solvents with 2 or more hydrogen bonds to protein	45
Solvents bridging between 2 or more protein molecules	14
Solvents with hydrogen bonds to other solvents only	19
Solvents with no hydrogen bonds	7

not rejected since the estimated error for the protein and the solvent atoms are about 0.10Å and 0.15Å respectively.

Following Blake *et al*. (1983), we grouped the solvents into different categories according to the hydrogen bonds that are made (Table II.12). Of the seven water molecules that make no hydrogen bonds, six of them have polar groups within 4.0Å. The remaining water (Wat126) is located in a region of weak electron density and is not reliable. The breakdown of solvents according to the type of the group or the atom on the protein to which they are bound is shown in Table II.13. The distribution is similar to that observed in lysozyme (Blake *et al*., 1983) and penicillopepsin (James & Sielecki, 1983).

Table II.13

Beakdown of Solvent-Protein Hydrogen Bonds by Groups on the Protein

Hydrogen bond (%)	
To main-chain CO	45
To main-chain NH	16
To side-chain.	39
To oxygen atoms	73
To nitrogen atoms	27
Mean distance to oxygen atoms (Å)	2.91(0.24)
Mean distance to nitrogen atoms (Å)	3.01(0.17)

Bibliography

Artymiuk, P.J. & Blake, C.C.F. (1981) *J. Mol. Biol.* *152*, 737-762.

Avignon, M., Huong, P.V. & Lascombe, J. (1969) *Biopolymers* *8*, 69-89.

Bachovchin, W.W., Kaiser, R., Richards, J.H. & Roberts, J.D. (1981) *Proc. Natl. Acad. Sci. U.S.A.* *78*, 7323-7326.

Barry, C.D., Molnar, C.E., & Rosenberger, F.U. (1976) *Tech. Memo #229* Computer Systems Lab., Washington University, St. Louis, Mo.

Bauer, C.-A., Brayer, G.D., Sielecki, A.R. & James, M.N.G. (1981) *Eur. J. Biochem.* *120*, 289-294.

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

Blake, C.C.F., Pulford, W.C.A. & Artymiuk, P.J. (1983) *J. Mol. Biol.* *167*, 693-723.

Blundell, T.L. & Johnson, L.N. (1976) *Protein Crystallography*, 409-411, Academic Press, London.

Brayer, G.D., Delbaere, L.T,J. & James, M.N.G. (1978) *J. Mol. Biol.* *124*, 261-283.

Brayer, G.D., Delbaere, L.T.J. & James, M.N.G. (1979) *J. Mol. Biol.* *131*, 743-775.

Chambers, J.L., & Stroud, R.M. (1979) *Acta Cryst.* *B35*, 1861-1874.

Christensen, P. & Cook, F.D. (1978) *Int. J. Sys. Bact.* *28*, 367-393.

Crawford, J.L., Lipscomb, W.N., & Schellmann, C.G. (1973) *Proc. Natl. Acad. Sci. U.S.A.* *70*, 538-542.

Cruickshank, D.W.J. (1949) *Acta Cryst.* *2*, 65-82.

Cruickshank, D.W.J. (1954) *Acta Cryst.* *7*, 519.

Cruickshank, D.W.J. (1967) in *International Tables for X-ray Crystallography*, Vol.II, (Kasper, J.S. & Londsdale, K. eds.), 318-340, Kynoch Press, Birmingham, England.

Delbaere, L.T.J., Brayer, G.D. & James, M.N.G. (1979) *Nature*

(London), *279*, 165-168.

Finney, J.L. (1979) In *Water:A Comprehensive Treatise*, (Franks, F. ed.) vol *6*, 47-122, Plenum Press, New York.

Hagler, A.T., Huler, E. & Lifson, S. (1974) In *Peptides Polypeptides and Proteins*, (Blout, E.R., Bovey, F.A., Goodman, M. & Lotan, N. eds.), 35-48, Wiley, New York.

Hendrickson, W.A. (1976) *J. Mol. Biol. 106*, 889-893.

Hendrickson, W.A., & Konnert, J.H. (1980) In *Biomolecular Structure, Function, Conformation and Evolution* (Srinivasan, R. ed.) Vol. I, 43-57, Pergamon Press, Oxford.

Huber, R., & Bode, W. (1978) *Acc. Chem. Res. 11*, 114-122.

Hunkapiller, M.W., Smallcombe, S.H., Whitaker, D.R. & Richards, J.H. (1973) *Biochemistry*, *12*, 4732-4743.

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

James, M.N.G. & Sielecki, A.R. (1983) *J. Mol. Biol. 163*, 299-361.

Janin, J., Wodak, S., Levitt, M. & Maigret, B. (1978) *J. Mol. Biol. 125*, 357-386.

Kaplan, H. & Whitaker, D.R. (1969) *Can. J. Biochem. 47*, 305-316.

Kaplan, H., Symonds, V.B., Dugas, H. & Whitaker, D.R. (1970) *Can. J. Biochem. 48*, 649-658.

Lee, B. & Richards, F.M. (1971) *J. Mol. Biol.* 55, 379-400.

Luzzati, V. (1952) *Acta Cryst. 5*, 802-810.

McLachlan, A.D. & Shotton, D.M. (1971) *Nature New Biology*, *229*, 202-205.

Olovsson, I. & Jonsson, P.-G. (1976) In *The Hydrogen Bond - Recent Developments in Theory and Experiments*, (Shuster, P., Zundel, G. & Sandorfy, C. eds.) Vol. II, Chapter 8, North-Holland Publ. Co., Amsterdam.

Pimentel, G.C. & McClellan, A.L. (1960) *The Hydrogen Bond*, 5-6, Freeman, San Francisco.

Ramachandran, G.N. & Mitra, A.K. (1976) *J. Mol. Biol. 107*, 85-92.

Ramakrishnan, C. & Ramachandran, G.N. (1965) *Biophys. J. 5*, 909-933.

Read, R.J., Fujinaga, M., Sielecki, A.R. & James, M.N.G. (1983) *Biochemistry, 22*, 4420-4433.

Rees, D.C., Lewis, M. & Lipscomb, W.N. (1983) *J. Mol. Biol. 168*, 367-387.

Richardson, J.S. (1981) In *Advances in Protein Chemistry* , (Anfinsen, C.B., Edsall, J.T. & Richards, F.M. eds.) Vol. *34*, 167-339, Academic Press, New York.

Rossmann, M.G. & Argos, P. (1975) *J. Biol. Chem. 250*, 7525-7532.

Sawyer, L., Shotton, D.M., Campbell, J.W., Wendell, P.L., Muirhead, H., Watson, H.C., Diamond, R. & Ladner, R.C. (1978) *J. Mol. Biol. 118*, 137-208.

Sielecki, A.R., Hendrickson, W.A., Broughton, C.G., Delbaere, L.T.J., Brayer, G.D. & James, M.N.G. (1979) *J. Mol. Biol. 134*, 781-804.

Sielecki, A.R. & James, M.N.G. (1981) In *Refinement of Protein Structures*, (Machin, P.A., Campbell, J.W. & Elder, M., eds.) 78-87, Proc. of the Daresbury Study Weekend 15-16 Nov. 1980, Daresbury Laboratory, England.

Sielecki, A.R., James, M.N.G., & Broughton, C.G. (1982) In *Crystallographic Computing* (Sayer, D., ed.) 409-419, Proc. of Intl. Summer School, Carleton University, Ottawa. Oxford University Press, Oxford.

Steitz, T.A. & Shulman, R.G. (1982) *Ann. Rev. Biophys. Bioeng. 11*, 419-444.

van Duijneveldt-van de Rijdt, J.G.C.M. & van Duijneveldt, F.B. (1971) *J. Amer. Chem. Soc. 93*, 5644-5653.

Venkatachalam, C.M. (1968) *Biopolymers 6*, 1425-1436.

Watenpaugh, K.D., Margulis, T.N., Sieker, L.C. & Jensen, L.H. (1978) *J. Mol. Biol. 122*, 175-190.

Westler, W.M., Markley, J.L. & Bachovchin, W.W. (1982) *FEBS Letters 138*, 233-235.

Whitaker, D.R., Roy, C., Tsai, C.S. & Jurasek, L. (1965) *Can. J. Biochem. 43*, 1961-1970.

III. Tonin

Tonin is a serine protease with a possible role to play in the regulation of blood pressure. In vitro, it is capable of producing angiotensin II, a powerful vasoconstricter, from angiotensinogen, angiotensin I or renin tetradecapeptide substrate by the cleavage of a Phe-His bond (Figure III.1) (Grisé *et al*., 1981, Boucher *et al*., 1972).

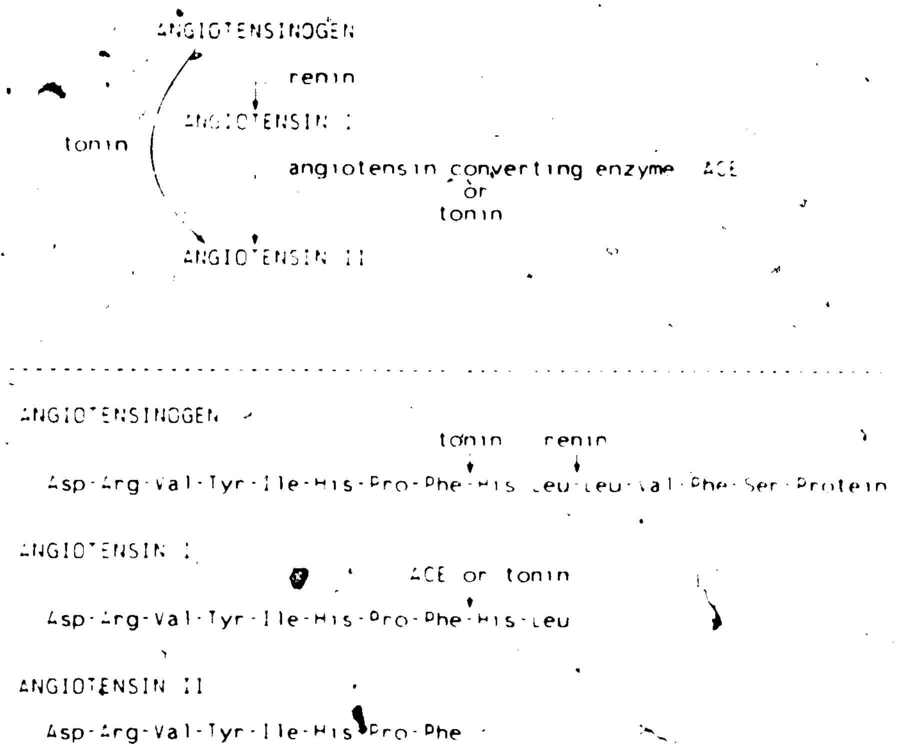
ANGIOTENSINOGEN
renin
ANGIOTENSIN I
tonin
angiotensin converting enzyme ACE
or
tonin
ANGIOTENSIN II
ANGIOTENSINOGEN
tonin
renin
Asp-Arg-Val-Tyr-Ile-His-Pro-Phe-His-Leu-Leu-Val-Phe-Ser-Protein
ANGIOTENSIN I
ACE or tonin
Asp-Arg-Val-Tyr-Ile-His-Pro-Phe-His-Leu
ANGIOTENSIN II
Asp-Arg-Val-Tyr-Ile-His-Pro-Phe

Figure III.1. Reactions Catalyzed by Tonin. Angiotensin II is produced either by the combined actions of renin and angiotensin converting enzyme (ACE) or by tonin. Renin cleaves the N-terminal decapeptide from angiotensinogen to form angiotensin I which is subsequently converted by ACE to angiotensin II. Tonin cleaves at a Phe-His bond in angiotensinogen to produce angiotensin II directly. Tonin also catalyzes the formation of angiotensin II from angiotensin I or from the tetradecapeptide renin substrate (the N-terminal fourteen residues of angiotensinogen).

The production of angiotensin II normally requires two enzymes, renin and angiotensin converting enzyme. Tonin is not affected by pepstatin, a renin inhibitor, or by inhibitors of the angiotensin-converting enzyme: EDTA, SQ14,225 or SQ20,881 (Boucher *et al.*, 1974, Boucher *et al.*, 1977). It is however strongly inhibited by an inhibitor in the plasma thought to be α_1-macroglobulin (Tremblay *et al.*, 1981) which does not inhibit renin or angiotensin-converting enzyme.

The actual physiological function of tonin is not known. Tonin was isolated from the submaxillary gland in rats (Boucher *et al.*, 1972) and was also found in the kidney (Ledoux *et al.*, 1982) and in the prostate (Ashley & McDonald, 1985a). However it has no activity in the plasma due to the presence of an inhibitor as mentioned above.

For small synthetic substrates, tonin shows a trypsin like activity, cleaving at arginyl residues under basic conditions (Thibault & Genest, 1981). On the other hand, studies with substrates such as β-lipotropin, adrenocorticotropin, proopiomelanocortin (Seidah *et al.*, 1979) and substance P (Chrétien *et al.*, 1980) have shown the strange ability of this enzyme to cleave either at phenylalanyl or arginyl residues. There is also an indication of an extended binding region. If the two N-terminal residues in angiotensin I are not present, tonin will not cleave the peptide (Schiller *et al.*, 1976) implying that interactions far from the scissile bond are important in the specificity of the reaction. Recent studies show that these properties of tonin are not unique. Both trypsin and kallikrein can produce angiotensin II from a plasma protein fraction at weakly acidic conditions (Arakawa *et al*., 1976, Arakawa *et al*., 1980, Arakawa & Maruta, 1980). Moreover, tonin can act as a kininogenase, producing bradykinin from low molecular weight (LMW) kininogen (Ikeda & Arakawa, 1984).

Initially the amino acid sequence of tonin was determined by Lazure *et al*. (1984) using protein sequencing methods. The sequence is highly homologous to other serine proteases such as kallikrein and the γ subunit of the nerve growth factor. Like kallikrein and trypsin it has an aspartate residue at position 189[1] so that it should also have a primary specificity for basic residues. The most surprising aspect of the sequence was the presence of a leucyl residue at position 102 where an aspartate residue is normally found. This aspartate is part of the catalytic triad, Asp-His-Ser, in serine proteases, and is believed to be indispensable for their activity. Ashley & McDonald (1985b) have subsequently completed the nucleotide sequence of tonin from rat along with those of several other kallikrein-like proteins. They showed that Lazure *et al*. (1984) had missed a 16 residue fragment and like other serine proteases tonin has an aspartate residue at position 102. The total number of amino acids in tonin is thus 235

[1] The residue numbering used throughout this chapter is derived from the structural alignment of tonin with chymotrypsin.

with a molecular weight of 25548.

Trigonal crystals of tonin were first reported by Hayakawa *et al*. (1978). These crystals diffracted to about 2.5Å resolution. Tetragonal crystals of much better diffracting quality were obtained by including Zn^{2+} in the crystallization solution. Unfortunately, as will be shown later, Zn^{2+} interacts with the enzyme in the region of its active site, distorting it from the native conformation.

The structure solution and refinement of the Zn^{2+} inhibited form of tonin will be described.

A. Materials and Methods

Crystallization and Data Collection

Purified enzyme was kindly provided by Dr. R. Boucher[1] and Dr. J. Genest of the Clinical Research Institute of Montreal. The crystallization was done by the hanging drop method. The crystal used to collect the initial 2.8Å resolution data set was grown under slightly different conditions from that used to collect the 1.8Å resolution data set. The crystal data are summarized in Table III.1. Data collection and processing methods were very similar for the two crystals (Table III.2). Table III.3 gives the percentage of reflections with intensities above one or two times their estimated standard deviations based on counting statistics. The program ORESTES (Thiessen

[1] deceased

Table III.1

Crystal Data

	Crystal 1	Crystal 2
Growth condition	15 ul/ml PEG200[1] 0.1 M cacodylate 10 mM $ZnSO_4$ 1.8 M $(NH_4)_2SO_4$ pH 6.2	0.1 M PIPES[2] 5 mM $ZnSO_4$ 2.1 M $(NH_4)_2SO_4$ pH 6.8
Crystal size	0.80x0.35x0.60mm	0.76x0.48x0.30mm
Space group	$P4_32_12$	$P4_32_12$
a	48.64(1) Å	48.58(1) Å
b	48.64(1) Å	48.58(1) Å
c	201.23(6) Å	200.20(10) Å
Z	8	8

[1] Polyethylene glycol
[2] Piperazine-N,N'-bis[2-ethane sulfonic acid]

& Levy, 1973) was used to scale the data to absolute units and to produce normalized structure factors (|E|). The Wilson plot indicated a scale of 36.5 and an overall temperature factor of 16.1 Å^2 for crystal 2.

Structure Solution

The method of molecular replacement was used to solve the phase problem. For the search model, the refined structure of trypsin (Fehlhammer & Bode, 1975; Bode & Schwager, 1975) was used. The rotational parameters were obtained by using the fast rotation function (Crowther, 1972) with data between 10 and 3.5Å resolution and a radius of integration between 5 and 21Å. The maximum of the function was 4.4σ above the mean. Before doing the translational search, the trypsin search model was rotated according to the angular

Table III.2

Data Collection and Processing

Diffractometer	Enraf-Nonius CAD4
Incident beam	Ni-filtered CuKα,40kV,26mA
Diffracted beam	60 cm crystal-counter He-filled beam path
Scan type	ω scan, continuous
Scan width	0.6° at 0.[illegible]n, (crystal 1) 0.66+0.14$\tan\theta$ at 0.8°/min (crystal 2)
Background measurement	16.7% of the total scan width on each side of the peak
Background correction	averaged in ranges of θ and ϕ
Absorption correction	North *et al*. (1968)
Decay correction	Hendrickson (1976)
Geometric correction	Lorentz and polarization

	Crystal 1	Crystal 2
Minimum d-spacing	2.8Å	1.8Å
Total no. of reflections	6954	26599
No. of unique reflections	6548	23355
Merging R[1]	3.8% (149 refs.)	7.4% (2744 refs.)
No. of refs. with I ≥ 2σ(I)[2]	4707	11023

[1] $R = \Sigma|\bar{I}-I_i|/\Sigma\bar{I}$

[2] $\sigma^2(I) = I+c^2I^2+(t_I/t_{Bk})^2(\Sigma Bk+c^2\Sigma Bk^2)$

I=total intensity
Bk=background counts
t_I=time taken for intensity measurement
t_{Bk}=time taken for background measurement
c=instrument instability=0.01

parameters corresponding to this maximum.

The method used to solve the translational problem was inspired by the work of Harada *et al*. (1981) (see Chapter IV). A systematic search is used in which the model is moved over a grid in the search volume. At each point of the grid, structure factors are calculated and are compared

Table III.3

Number of Observed Reflections from Crystal 2

Resolution range	No. of reflections	I ≥ 1σ (%)	I ≥ 2σ (%)
∞-4.00	2371	90	85
4.00-3.00	3018	82	73
3.00-2.70	1879	78	63
2.70-2.50	1834	72	55
2.50-2.30	2447	67	46
2.30-2.15	2492	67	47
2.15-2.00	3285	60	35
2.00-1.90	2773	49	25
1.90-1.80	3345	38	14

to the observed structure factors by evaluating the correlation coefficient,

$$C = \frac{\Sigma(|F_o|^2-\overline{|F_o|^2})(|F_c|^2-\overline{|F_c|^2})}{[\Sigma(|F_o|^2-\overline{|F_o|^2})^2\Sigma(|F_c|^2-\overline{|F_c|^2})^2]^{1/2}}$$

where $|F_o|$ and $|F_c|$ are the observed and calculated structure factor amplitudes, respectively, and the summation is carried out over a set of reflections. Initially the 140 reflections between 8-10Å resolution and 2Å grid spacing were used. A complete translational search volume for space group $P4_32_12$ is from 0 to 1 in x, 0 to 1/2 in y and 0 to 1/2 in z. The maximum correlation obtained was 0.43, or 4.5σ above the mean. However there were several other peaks along the z direction that were within 1σ of this maximum. Finer searches around these peaks using 0.5Å grid spacings and 903 reflections between 5-10Å resolution discriminated the correct solution. With the higher resolution data the correlation coefficient in the region of the maximum in the previous search became 0.33. The correlation coefficient will, in general, go down with more data but so will the noise level. The corresponding values for other peaks were about 0.23. An improvement in the algorithm increased the speed of computation and allowed the recalculation of the translation function over the total search area using data between 4 and 5Å resolution and a 1Å grid spacing and confirmed the correctness of the solution. A peak which was 8σ above the mean clearly showed the correct position of the molecule in the crystal.

For certain space groups, there is an enantiomorphic space group for which the extinctions are identical and a unique assignment is not possible from the diffraction pattern. $P4_12_12$ and $P4_32_12$ are such a pair and it was during the translational search that the ambiguity of the space group became resolved. The translational searches done in the enantiomorphic space group $P4_12_12$ failed to give a significant signal. The R factor ($\Sigma||F_o|-|F_c||/\Sigma|F_o|$) for data between 10-2.8Å resolution based on the rotated and translated trypsin search model in the correct space group $P4_32_12$ was 0.48.

Refinement

At the time the refinement was started, only partial sequence information was available (Lazure *et al.*, 1981). Therefore a combined tonin-trypsin model was constructed by replacing, wherever possible, the amino acids of the molecular replacement trypsin model by those of tonin. The resulting structure consisted of 223 amino acid residues of which 146 residues, in three fragments, corresponded to the sequence of tonin. The restrained-parameter least-squares refinement program of Hendrickson & Konnert (1980) was used. Manual corrections to the model were made on the MMS-X interactive graphics system (Barry *et al.*, 1976) using the M3 program written by C. Broughton (Sielecki *et al.*, 1981). Reinterpretations were done mainly based on electron density maps computed with coefficients $2|F_o|-|F_c|$, and calculated phases α_c. The progress of the refinement was very slow and initially there was much concern about the extent of model bias since the only source of phase information was the model. Many parts of the molecule roughly fitted the electron density whereas other parts that did not fit could not be reinterpreted based on the current maps. The problem remained even after most of the sequence (except for a 16 residue fragment that had been missed) became available (Lazure *et al.*, 1984). At this stage the R factor was 0.43 for the data with $I \geq 2\sigma(I)$ in the 6.0-2.8Å resolution shell. Various approaches were tried in order to overcome model bias and to establish confidence in the model. These included weighting of map coefficients by figures of merit, free atom refinement in which all restraints were removed, refinement of a polyglycine model based on the main chain atoms of the tonin-trypsin structure, and partial difference maps with coefficients, $|F_o|-|F_c|$, α_c, in which parts of the model were removed before calculating structure factors. All these procedures did not seem to give any more information than the maps with coefficients $2|F_o|-|F_c|$. In the end, after many cycles of refinement, the strategy that worked the best was to remove all parts of the model that could not be interpreted from the electron density map and then to refine the resulting partial structure. This consisted of 185 residues or about 80% of the whole molecule. The R factor at that point was 0.34 for the 6.0-2.0Å data. The missing amino acid residues were added back into the model only when the associated electron density became apparent and interpretable. Additional improvement in the phases and the quality of the electron density were obtained when a peak near the active site region was interpreted as a Zn^{2+} ion. It clarified the position of a loop containing two histidine residues that bind the Zn^{2+} ion. As the refinement progressed the resolution of the data used was slowly increased, individual temperature factors, B, were refined and solvent molecules were added.

When the refinement was almost complete (R = 0.17 for 8.0-1.8Å data), the positions of the 190 solvent molecules in the structure were checked in a manner similar to that done for α-lytic protease (Fujinaga *et al.* (1985). First the solvent sites were ordered according to a quality factor defined by occupancy²/B (James & Sielecki, 1983). This factor was defined to reflect the fact that the atoms with high occupancies and low B values are the most reliable.

Table III.4

r.m.s. Deviations from Ideal Geometry at the End of Refinement

Distance restraints (Å)	
Bond distance	0.016(0.013)[1]
Angle distance	0.031(0.018)
Planar 1-4 distance	0.027(0.018)
Plane restraint (Å)	0.018(0.015)
Chiral-volume restraint (Å³)	0.139(0.080)
Non-bonded contact restraint (Å)	
Single torsion contact	0.220(0.250)
Multiple torsion contact	0.192(0.250)
Possible hydrogen bond	0.207(0.250)
Peptide torsion angle restraint (°)	3.2(2.0)

[1] The values of σ, in parentheses, are the input estimated standard deviations that determine the relative weights for the corresponding restraints (see Hendrickson & Konnert, 1980).

The solvent molecules with the 2/3 lowest quality factors were removed from the model. A few cycles of refinement were carried out preceding the computation of a difference map from which new solvent positions were determined. Additional rounds of refinement and examination of difference maps resulted in a total of 149 solvent sites, 41 fewer than originally assigned.

Most of the refinement was done using data with $I \geq 2\sigma(I)$. The last few cycles of refinement were carried out including data with $I \geq 1\sigma(I)$. The parameters of the last cycle as well as deviations from ideal geometry are given in Table III.4. Figure III.2 shows the R-factor as a function of resolution. The change from the initial molecular replacement model can be seen in Figure III.3, where the α carbons of the trypsin search model and those of the refined

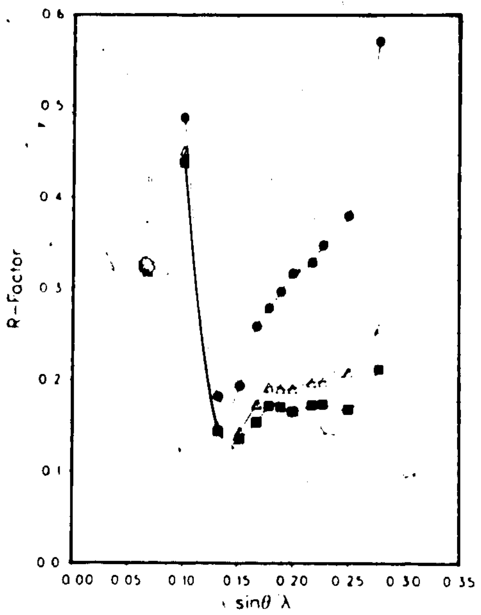

Figure III.2. Variation of the R-factor with Resolution. R-factor in ranges of $\sin\theta/\lambda$. ● all data; Δ data with $I \geq \sigma(I)$; ■ data with $I \geq 2\sigma(I)$. The overall R-factor is 0.196 for data between 8-1.8Å resolution with $I \geq \sigma(I)$.

tonin structure are superimposed. The numbering of the amino acids in this figure and throughout the chapter follows that of chymotrypsinogen and is based on the tertiary structural alignment of tonin with α-chymotrypsin. For this purpose, the structure of α-chymotrypsin in the molecular complex with the third domain of the turkey ovomucoid inhibitor refined at 1.8Å resolution was used (Read, R. J, Fujinaga, M., Sielecki, A., Ardelt, W., Laskowski, M. Jr., & James, M. N. G., unpublished results).

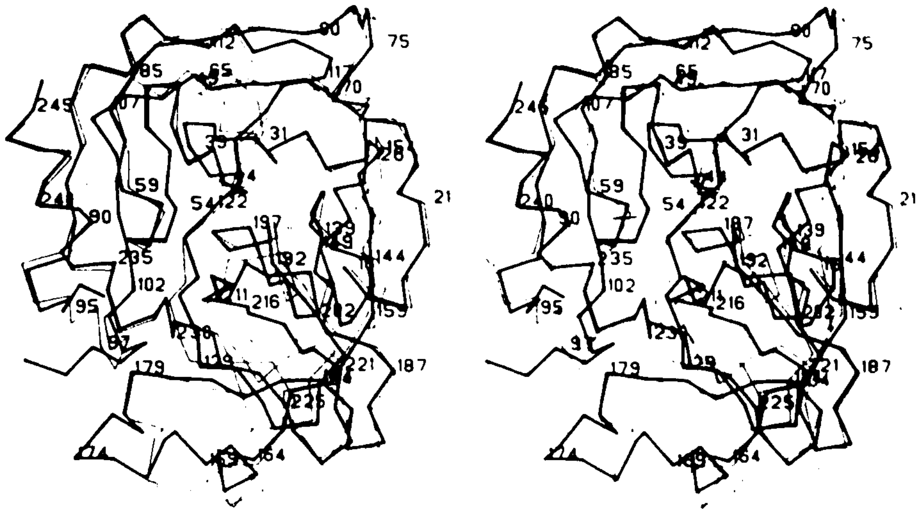

Figure III.3. Comparison of the Molecular Replacement Solution with the Final Refined Structure. The trypsin model obtained with molecular replacement is shown superimposed on the final refined structure of tonin. The molecules are represented by line segments connecting the α-carbon atoms. Tonin drawn with thick lines and is labeled every five residues, and trypsin is shown with thin lines.

Quality of the Structure

There is yet no known way of determining accurately the errors in atomic coordinates resulting from a sparse-matrix restrained refinement. Therefore, errors in the coordinates were estimated using several different approaches. Luzzati (1952) calculated the mean error in the coordinates from the variation of the R-factor with resolution. The expected variation assumes that the discrepancy between $|F_o|$ and $|F_c|$ results only from coordinate errors and that the observed and the model structures have the same scattering power. For the structure of tonin the mean error calculated in this way is about 0.35Å that corresponds to a root-mean-square

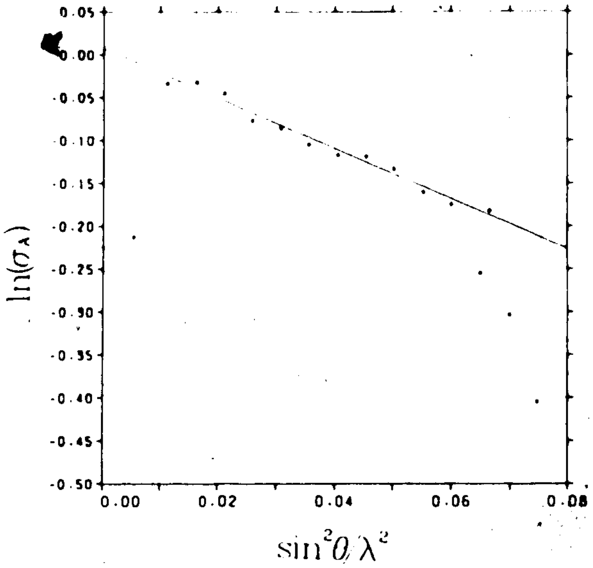

Figure III.4. Estimation of the Error with a σ_A Plot. Plot of $\ln(\sigma_A)$ vs. $\sin^2\theta/\lambda^2$. for estimating the coordinate error by the method of Read (1986). The slope of the regression line is related to the mean coordinate error ($|\Delta r|$) by, slope=$\pi^3|\Delta r|^2$. The first and the last three points were rejected in determining the regression line.

(r.m.s.) error of 0.38Å.

A similar method which uses the agreement between $|F_o|$ and $|F_c|$ to derive the mean error is that of Read (1986). The error is obtained from the slope of a line fitted to the plot of $\ln(\sigma_A)$ vs $\sin^2\theta/\lambda^2$ (Figure III.4), where σ_A is related to the sharpness of the phase probability distribution (Srinivasan & Chandrasekaran, 1966). This approach allows for a difference in the scattering power between the model and the real structure and is insensitive to scaling errors. The mean error obtained by this method is 0.31Å ($\sigma_{r.m.s.}$ = 0.34Å).

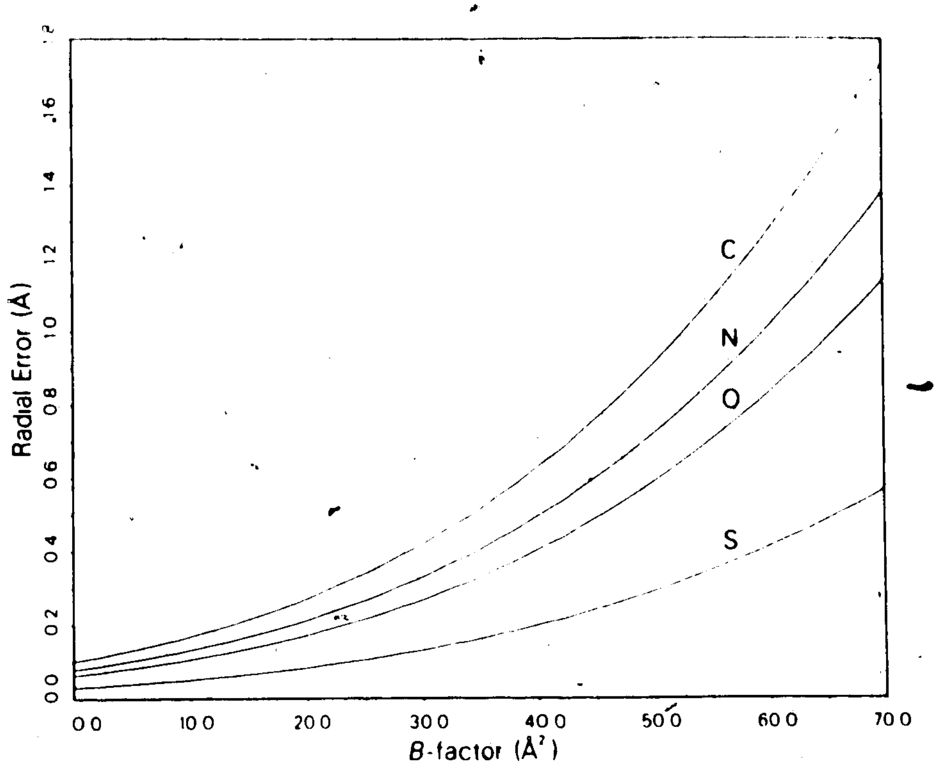

Figure III.5. Estimation of the Error with the Method of Cruickshank. Coordinate errors calculated from the equations given by Cruickshank (1949, 1954, 1967). The radial error was calculated from $\sigma_r{}^2=3\sigma_x{}^2$. Errors for carbon, nitrogen, oxygen, and sulfur atoms as a function of the temperature factor, B, are plotted.

Unlike the above two methods which give only an overall value for the error, the formulae derived by Cruickshank (1949, 1954, 1967) allow for the calculation of coordinate errors as a function of atom type and temperature factor. Despite the fact that the equations were derived for difference Fourier refinement with resolved atoms, they seem to give reasonable results (Read *et al.*, 1983, Fujinaga *et al.*, 1985). Figure III.5 shows the plots of the calculated errors as a function of temperature factor, B, for four

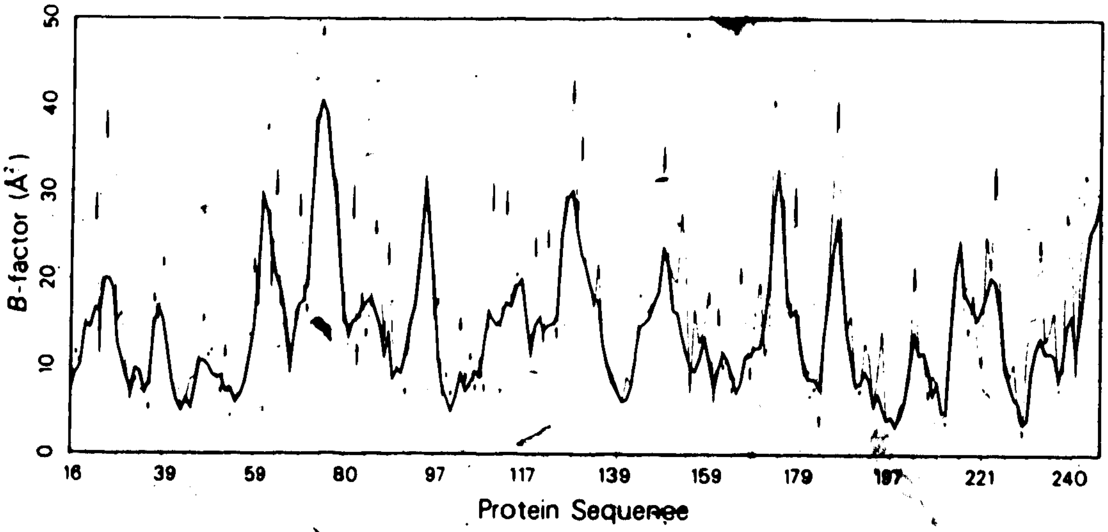

Figure III.6. The Variation of the B-factor Along the Polypeptide Chain. The average temperature factors of the main chain atoms are shown with thick lines and the average over the side chain atoms are shown with thin lines.

different atom types. The variation of the temperature factor along the polypeptide chain is depicted in Figure III.6. These two figures can be used to obtain the errors associated with each residue. The r.m.s. error for all the atoms in the structure is 0.25Å. Thus the three different methods all indicate that the overall error is about 0.3Å. This value is larger than those found for other structures refined at high resolution where the errors are in the range 0.1 to 0.2Å. The larger error in the tonin coordinates is probably due to the lower effective resolution of the data (Table III.3).

Qualitative information about the accuracy of the refined model can be obtained by examining the fit of the

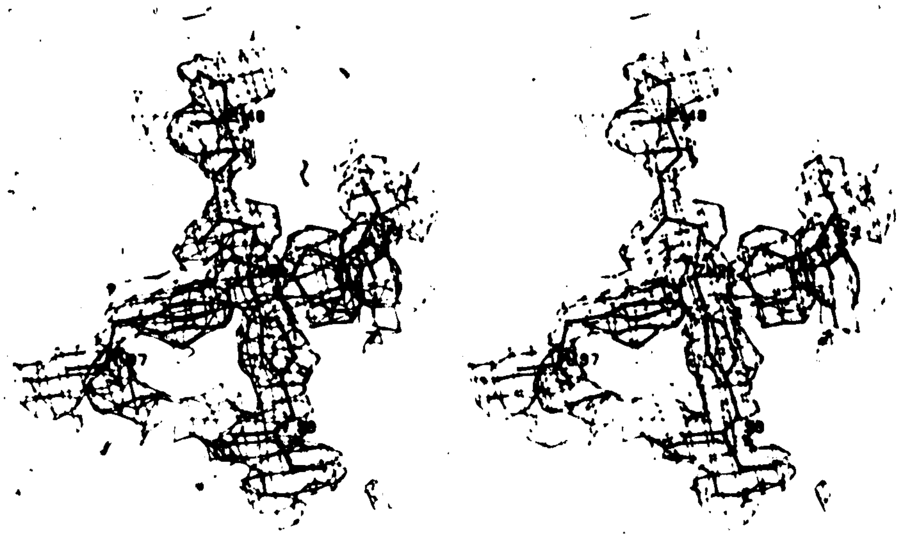

Figure III.7. The Zn^{2+} Ion Environment. The coordination of the Zn^{2+} ion found near the active site of the enzyme with His57 being one of the ligands. The superposed electron density shows the map calculated with coefficients, $2|F_o|-|F_c|$, α_c contoured at 40 $e\AA^{-3}$.

molecular model to the electron density map. For the most part the agreement between the model and the electron density is very good. As an example, Figure III.7 shows the environment of the Zn^{2+} ion bound in the vicinity of the active site. Nevertheless there are other regions that are not as well defined. An eight residue segment from Ile95C to Gln95J could not be located. The following residue Pro95K could not be fitted as a prolyl residue and was refined as an alanine. Likewise, there was no electron density associated with the side chains of Arg86 and so the atoms beyond CB were removed from the model. In addition, there are some other residues whose side chains could not be fitted well. Table III.5 summarizes all the regions of the

Table III.5

Poorly Determined Regions

Lys21	CD, CE, NZ	low density
Glu23	CG, CD, OE1, OE2	low density
Lys24	CE, NZ	no interpretable density
Asn63	OD1, ND2	no interpretable
Gln65	CD, OE1, NE2	spherical density, maybe rotational disorder
Leu73	side chain	poor fit but in density
Phe74	CB	no interpretable density, ring has some density
Lys75	CE, NZ	no interpretable density
Arg82	NE, CZ, NH1, NH2	no interpretable density
Arg86	side chain	refined as Ala, no interpretable density beyond CD
Gln87	CD, OE1, NE2	low density
Ile95C to Gln95J		no interpretable density
Pro95K		refined as Ala, not possible to fit a Pro
Glu110	CD, OE1, OE2	no interpretable density
Thr125	OG1, CG2	poor fit
Lys128 to Glu129		disconnected, probably mobile (see Figure III.6)
Lys131	CG, CD, CE, NZ	no interpretable density
Val150	CB, CG1, CG2	poor fit
Glu166	CG, CD, OE1, OE2	poor fit
Lys173	CG, CD, CE, NZ	low density
Met186	CB, CE	no interpretable density for CB, poor fit for CE
Glu186A	CB, CG, CD, OE1, OE2	no interpretable density
Lys222	CE, NZ	low density, probably there is an alternate conformation
Lys223	CG, CD, CE, NZ	no interpretable density
Lys239	CE, NZ	low density
Lys243	CG, CD, CE, NZ	no interpretable density

model where the agreement with the electron density are poor.

B. Results and Discussion

Overall Description

The refined structure of tonin consists of 227 amino acid residues out of the 235 in the complete molecule. The side chains of two residues, Arg86 and Pro95K were omitted beyond the β-carbon atom. The model also includes one Zn^{2+} ion and 149 water molecules. There is one cis peptide at Pro219. The model of tonin is shown in Figure III.8(a) while the main chain atoms and hydrogen bonding among them are shown in Figure III.8(b).

The secondary structural assignments were done according to the definitions given by Kabsch & Sander (1983), in which electrostatic interactions between NH and CO groups are calculated according to,

$$E = q_1 q_2 (1/d(ON)+1/d(CH)-1/d(OH)-1/d(CN))*332$$

with q_1=0.42e and q_2=0.20e, e is the unit electron charge and d(AB) is the distance between atoms A and B in Å. The dimension factor, 332, gives E in unit of kcal/mol. Interactions with E<-0.5 kcal/mol are considered as hydrogen bonds. The output of their program is reproduced on Table III.6. It shows that the protein consists mainly of extended β-sheets with three helical regions. There is a short 3_{10} helix from Ala56 to Cys58 that contains the active

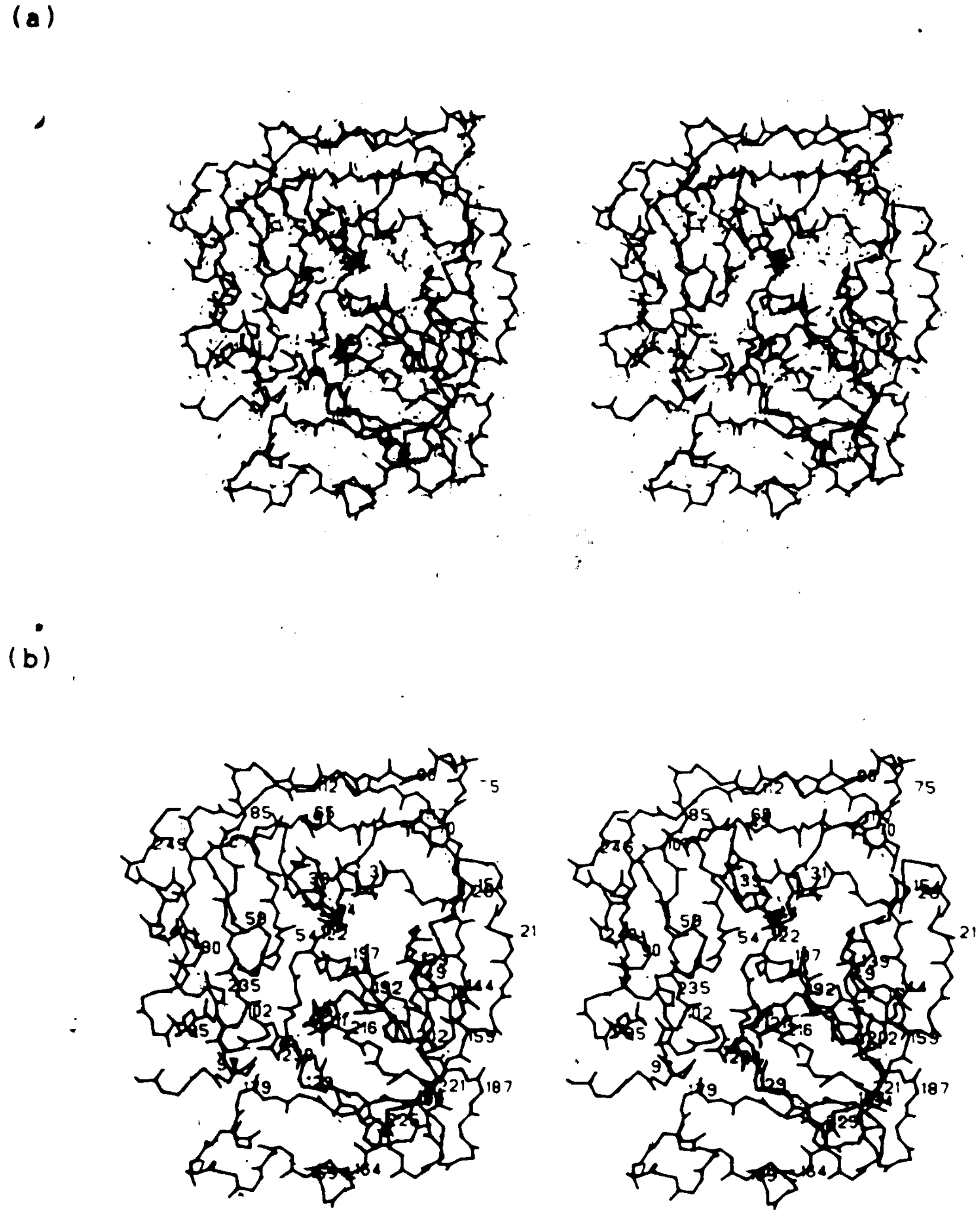

Figure III.8. The Refined Structure of Tonin. (a) The complete molecule with the polypeptide backbone shown in thick lines.
(b) Same view as in (a) but without the side-chain atoms. Amino acids are labeled every five residues and hydrogen bonds among the backbone atoms, as defined by the criterion of Kabsch & Sander (1983), are shown with broken lines.

Table III.6

Secondary Structure Assignments

RESIDUE		ACC	SUM	3-T	4-T	5-T	BND	CHR	BR1	BR2	SHT
I	16	0									
V	17	1	B					+	A		A
G	18	5	S				S	+			
G	19	3						-			
Y	20	*	E					-	B		[illegible]
K	21	*	E					[illegible]	B		[illegible]
a	22	1						-			
E	23	*		>				[illegible]			
K	24	*	T	3			S	-			
N	25	5	T	3			S	+			
S	26	5	S	<			S	+			
Q	27	3		>				+			
P	28	1	T	3			S	+			
W	29	5	T	3			S	+			
Q	30	1	E	<				+	I		C
V	31	0	E					-	I		C
A	32	0	E					-	I	J	C
V	33	0	E					-	I	J	C
I	34	5	E					+	I	J	C
N	35	3	S				S	-			
E	39	*	S				S	+			
Y	40	*	S				S	-			
L	41	4	E					+	I		C
b	42	1	E					-	I		C
G	43	0	E					-	I		C
G	44	0	E					-	I		C
V	45	0	E					-	I	K	C
L	46	0	E					+		K	C
I	47	2	E				S	+			
D	48	6	E	>			S	-		K	C
P	49	4	T	3			S	+			
S	50	2	T	3			S	+			
W	51	1	E	<				-	K	L	C
V	52	0	E					-	K	L	C
I	53	0	E					+	K	L	C
T	54	0	E					-		L	C
A	55	0		>				-			
A	56	0	G	>			S	+			
H	57	5	G	3			S	+			
b	58	0	G	<			S	+			
Y	59	*		<				+			
S	60	1						-			
N	61	*	S				S	+			
N	63	*						-			
Y	64	1						-			
Q	65	7	E					-	J		C
V	66	0	E					-	J	M	C

(Table III.6 continued)

RESIDUE	ACC	SUM	3-T	4-T	5-T	BND	CHR	BR1	BR2	SHT
L 67	3	E					+	J	M	C
L 68	0	E					+		M	C
G 69	0	S				S	+			
R 70	5						+			
N 71	2	S				S	+			
N 72	2	B	>			S	-	O		D
L 73	2	T	3			S	+			
F 74	*	T	3			S	+			
K 75	*	S	<			S	-			
D 77	*						+			
E 78	2		>				-			
P 79	*	T	3			S	+			
P 79A	7	T	3			S	+			
A 80	4		<				+			
Q 81	2	E					-	M		C
R 82	*	E					-	M		C
R 83	2	E					-	M		C
L 84	8						-			
V 85	3	E					-	N		C
R 86	7	E				S	-			
Q 87	*	E					-	N		C
S 88	3	E					-	N		C
F 89	4	E					-	N		C
R 90	9	E					-	N		C
H 91	2		>				-			
P 92	9	T	3			S	+			
D 93	*	T	3			S	+			
Y 94	2		<				-			
I 95	6						-			
P 95A	6									
L 95B	*									
I 95C	Not visible in the final electron density map.									
V 95D										
T 95E										
N 95F										
D 95G										
T 95H										
E 95I										
Q 95J										
P 95K	*									
V 96	*						-			
H 97	4						+			
D 98	4						-			
H 99	6	S	>			S	+			
S100	1	T	3			S	+			
N101	2	T	3			S	+			
D102	0		<				+			
L103	0						+			
M104	0	E					-	L	N	C

(Table III.6 continued)

RESIDUE	ACC	SUM	3-T	4-T	5-T	BND	CHR	BR1	BR2	SHT
L105	0	E					-	L	N	C
L106	0	E					-	L	N	C
H107	3	E					-	L	N	C
L108	1	E					-		N	C
S109	6	S				S	+			
E110	*	S				S	-			
P111	6						-			
A112	1						-			
D113	*						-			
I114	9						+			
T115	4						-			
G116	4	S				S	+			
G117	0	S				S	+			
V118	0						+			
K119	*						-			
V120	5						-			
I121	2						-			
D122	*						-			
L123	3						-			
P124	1						-			
T125	*						+			
K128	*	S				S	-			
E129	*						-			
P130	2						-			
K131	*		>				-			
V132	*	T	3			S	+			
G133	5	T	3			S	+			
S134	1		<				-			
T135	7	E					-	C		B
c136	0	E					-	C	D	B
L137	2	E					-	C	D	B
A138	0	E					-	C	D	B
S139	0	E					+	C		B
G140	0	E					-	C		B
W141	0	S				S	+			
G142	0	S				S	-			
S143	0						-			
T144	6	S				S	+			
N145	6	S				S	-			
P146	5	S				S	+			
S147	8	S				S	+			
E148	*	S				S	-			
M149	*						+			
V150	7						-			
V151	5						-			
S152	2						-			
H153	8	S				S	+			
D154	4	B				S	-	O		D
L155	1						-			

(Table III.6 continued)

RESIDUE	ACC	SUM	3-T	4-T	5-T	BND	CHR	BR1	BR2	SHT
Q156	2	E					-	B	C	B
a157	0	E					-	B	C	B
V158	0	E					-		C	B
N159	6	E					+		C	B
I160	0	E					-		C	B
H161	7	E					-	E	C	B
L162	1	E					-	E		B
L163	3	E					-	E		B
S164	5		>				-			
N165	3	G	>			S	+			
E166	*	G	3			S	+			
K167	*	G	<			S	+			
d168	0		X				-			
I169	*	G	>			S	+			
E170	7	G	>			S	+			
T171	1	G	<			S	+			
Y172	*	G	<			S	+			
K173	*	S	X	>		S	-			
D174	*	T	3	4		S	-			
N175	*	T	>	4		S	+			
V176	1	G	X	4			+			
T177	5	G	>	<		S	+			
D178	*	G	<			S	+			
V179	3	G	<			S	+			
M180	1	E	<				-		F	B
L181	1	E					-		F	B
d182	0	E					+	E	F	B
A183	0	E					+	E	F	B
G184	0	E				S	-	E		B
E185	4		>				-			
M186	4	T	3			S	+			
E186A	*	T	3			S	-			
G186B	1		<				+			
G187	4	S				S	+			
K188	8						+			
D189	1	B					-	A		A
T190	0						-			
e191	0		>				-			
A192	2	T	3			S	+			
G193	1	T	3			S	+			
D194	0		X				+			
S195	1	T	3			S	+			
G196	0	T	3			S	+			
G197	0		<				-			
P198	0	E					-		G	B
L199	0	E					-	D	G	B
I200	2	E					-	D	G	B
c201	0	E	>			S	-	D	G	B
D202	*	T	3			S	-			

(Table III.6 continued)

RESIDUE	ACC	SUM	3-T	4-T	5-T	BND	CHR	BR1	BR2	SHT
G207	7	T	3			S	+			
V208	5	E	<			S	-	G		B
L209	0	E					+	G		B
Q210	1	E					+			
G211	0	E					-	G	H	B
I212	0	E					-	G	H	B
T213	0	E					-		H	B
S214	0						-			
G215	2						-			
G216	2						+			
A217	8						-			
T218	6	S				S	+			
P219	9	S				S	+			
e220	1						+			
A221	1	S				S	+			
K222	*		>				-			
P222A	7	T	3			S	+			
K223	*	T	3			S	+			
T224	8		<				+			
P225	1						-			
A226	2	E					-	F		B
I227	1	E					+	F		B
Y228	0	E					-	F	H	B
A229	1	E					-	F	H	B
K230	3	E	>				-		H	B
L231	0	G	>	>		S	+			
I232	3	G	>	4		S	+			
K233	*	G	<	4		S	+			
F234	1	G	<	>		S	+			
T235	3	H	<	X		S	+			
S236	7	H		>		S	+			
W237	5	H		>		S	+			
I238	0	H		X		S	+			
K239	8	H		X		S	+			
K240	*	H		X		S	+			
V241	2	H		X		S	+			
M242	4	H		<		S	+			
K243	*	H		<		S	+			
E244	*	H		<		S	+			
N245	4			<						
P246	*									

RESIDUE = Amino acids are given in one-letter code except for Cys residues which are labeled a, b, ç, etc. to indicate the SS-pairs.

ACC = Accessible area, given as the possible number of water molecules in contact. * = more than 9 water molecules. Also can be interpreted as the solvated surface area

(Table III.6 continued)

in units of 10 Å^2.

SUM = Structure Summary
H = 4-helix (α-helix)
B = residue in isolated β-bridge
E = extended strand in β-ladder
G = 3-helix (3_{10}-helix)
I = 5-helix (π-helix)
T = hydrogen bonded turn
S = bend
In case of overlap, priority is given to the structure first in the above list.

3-T, 4-T, 5-T = Hydrogen bonding patterns for turns and helices
> = backbone CO makes H-bond(i,i+n)
< = backbone NH makes H-bond(i-n,i)
X = both CO and NH make H-bond
3, 4, 5 = residues bracketed by H-bonds

BND = Bend
S = five-residue bend

CHR = Chirality
The sign of the dihedral angle defined by CA i-1 to i+2

BR1,BR2 = Bridge1, Bridge2
Name of β-ladder in which the residue i participates
A, B, C, etc. = antiparallel
a, b, c, etc. = parallel
Ladders named sequentially from N- to C-terminus

SHT = Sheet
Name of β-sheet in which the residue participates
A, B, C, etc.

See Kabsch & Sander (1983) for more details.

site residue, His57. This helix is also found in other serine proteases and its dipole moment helps to stabilize the negative charge of Asp102 (Moult *et al.*, 1985). There are several 3_{10} helices from Asn165 to Val179, interrupted at Cys168 and at the turn from Lys173 to Val176. Finally, there is a long helix at the C-terminus that begins as a 3_{10} helix at Leu231 becoming an α-helix at Thr235 and continuing until Glu244.

A turn, as defined by Kabsch & Sander (1983), exists when there is a hydrogen bond between the carbonyl group of the ith residue and the amide group of the i+n residue, where n=3, 4, or 5. All the turns in tonin are of the type n = 3 and they are listed in Table III.7 with the associated hydrogen bonding energies (Kabsch & Sander, 1983) and ϕ-ψ angles. The ones with the lower energies conform more

Table III.7

Turns in Tonin

			Energy (kcal/mol)	ϕ_2	ψ_2	ϕ_3	ψ_3	Type'
β-turns								
Glu23	–	Ser26	-1.5	-53	129	48	45	II
Gln27	–	Gln30	-2.5	-58	-20	-90	-7	I
Asp48	–	Trp51	-1.1	-49	-28	-108	-2	I
Asn72	–	Lys75	-1.1	-63	-28	-115	46	I
Glu78	–	Ala80	-0.6	-72	-38	-93	7	I
His91	–	Tyr94	-1.6	-50	-36	-91	-2	I
His99	–	Asp102	-0.9	-78	144	65	26	II
Lys131	–	Ser134	-2.6	-54	136	92	-11	II
Lys173	–	Val176	-2.8	41	-115	-132	31	II'
Glu185	–	Gly186B	-2.2	-55	-28	-75	-18	I
Cys191	–	Asp194	-2.1	-59	128	102	-18	II
Asp194	–	Gly197	-1.9	-47	134	91	-12	II
Cys201	–	Val208	-2.2	52	36	86	-3	I'
Lys222	–	Thr224	-1.6	-61	142	63	23	II
Asx turns								
Tyr40	–	Asn35						
Ser50	–	Asp48						
Phe74	–	Asn72						
Ser100	–	Asp98						
Ser147	–	Asn145						

'Classification according to Venkatachalam (1968).

closely to the idealized classification of turns defined by Venkatachalam (1968). The table also shows the location of 'Asx turns' (Rees *et al.*, 1983).

Sibanda & Thornton (1985) have analyzed what they have called β-hairpins in proteins. These structures consist of two antiparallel strands connected by a loop region. In tonin, there are two β-hairpins, each with two residues in the loop. One of these is linked by a type I' turn, Cys201 to Val208. Type I' and type II' turns are most commonly observed in two residue loop β-hairpins (Sibanda & Thornton, 1985). However the other β-hairpin involves a type I turn

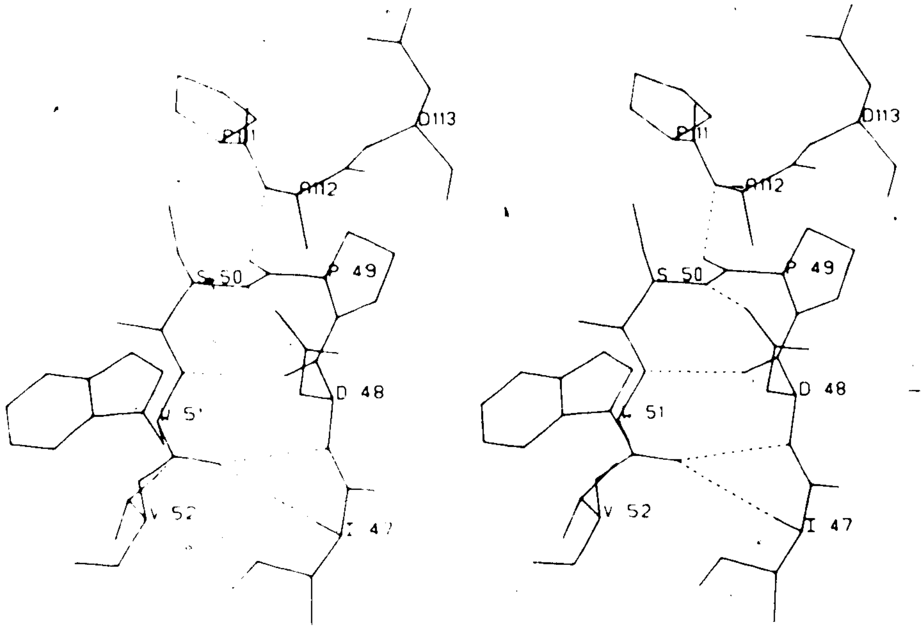

Figure III.9. Type I' Turn from Asp48 to Trp51. This type of turn is not favoured for an antiparallel β-hairpin (Sibanda & Thornton, 1985) but may be necessitated by the presence of a prolyl residue at the i+1 position. The Asx turn from Asp48 and the hydrogen bond from Ala112 stabilize the conformation of the turn.

from Asp48 to Trp51. This type of turn, that is thought to be incompatible with an antiparallel β-hairpin (Sibanda & Thornton, 1985), is most likely imposed by the presence of Pro49 at the i+1 position and stabilized by the hydrogen bond from Ala112 N to Pro49 O and the Asx turn between Asp48 OD1 and Ser50 N. In fact, four out of the five Asx turns in this structure occur in turn regions implying an important role in helping the main chain to form a bend. This trend has been noted in other proteins by Baker & Hubbard (1984).

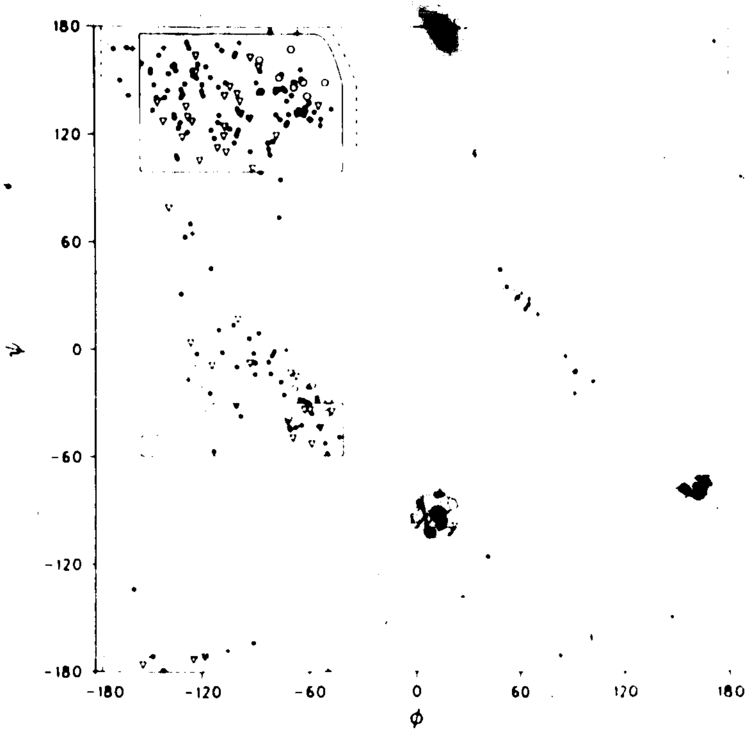

Figure III.10. ϕ-ψ Angles in Tonin. The symbols correspond to the following: (o) proline, (∇) β-branched amino acids, (+) glycine, (•) others. The fully allowed conformational regions (solid lines) and the areas of acceptable van der Waals contact for $\tau(C^{\alpha})$ of 115° (Ramakrishnan & Ramachandran, 1965) are shown.

The conformation of the molecule is shown as a ϕ-ψ plot (Ramakrishnan & Ramachandran, 1965) in Figure III.10. Most of the amino acid residues have conformations that fall within the region of acceptable van der Waals contact. Two exceptions are Asn35 with (ϕ,ψ) angles of (-158°,-134°) and Asp174 with (41°, -115°). The environment of these residues, along with the associated electron density are shown in Figures III.11(a) and (b). Asn35, involved in an Asx turn (Table III.7), is well ordered with an average main chain temperature factor of 16Å^2 and the fit to the electron density is very good. On the other hand Asp174, in a type II' turn (table III.7), is highly mobile with an average main chain temperature factor of 31Å^2 and the corresponding electron density is poor. However, it is still clear that the path of the polypeptide chain in this region is the correct one. Attempts to refit in a different conformation were unsuccessful. Both of these regions will be discussed in more detail in the next section.

Comparison with Kallikrein

The high sequence homology between tonin and kallikrein (Lazure *et al*., 1984; Ashley & McDonald, (1985b) suggests that their tertiary structures should be similar. The refined coordinates of tonin were superimposed on those of porcine kallikrein (Bode *et al*., 1983) with a program written by W. Bennett based on the method of Rossmann & Argos (1975). A probability cutoff of 0.005 was used

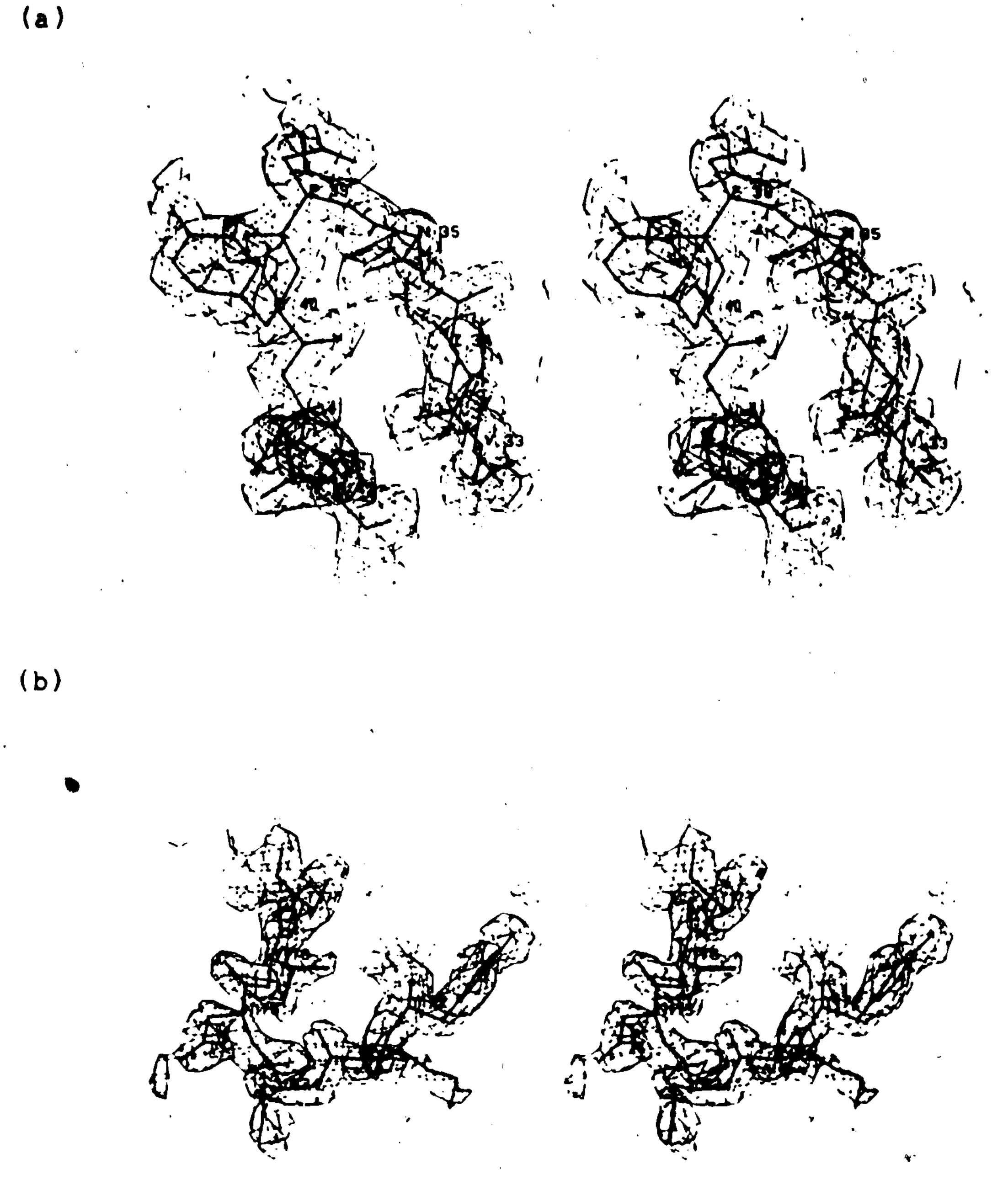

Figure III.11. Regions with Abnormal ϕ-ψ Angles. (a)Asn35 and its surroundings superimposed on an electron density map calculated with coefficients $2|F_o|-|F_c|$, α_c. The map is contoured at 0.40 eÅ^{-3}. (b)Region around Asp174 superimposed on an electron density map as in (a).

without the progression rule. The two structures closely resemble one another as seen in Figure III.12. There are 203 α-carbon atom pairs that superimpose within 1.9Å with an r.m.s difference of 0.6Å. Table III.8 shows the sequence of tonin aligned with that of kallikrein according to the tertiary structural homology. This alignment is identical to the previous alignment of these proteins based on the sequence information alone (Ashley & McDonald, 1985b). As previously stated, the sequence numbering used is based on the tertiary structural alignment of tonin and α-chymotrypsin. This numbering differs slightly from that used by Bode *et al*. (1983) for kallikrein. In particular, since the polypeptide chain of kallikrein is cleaved at

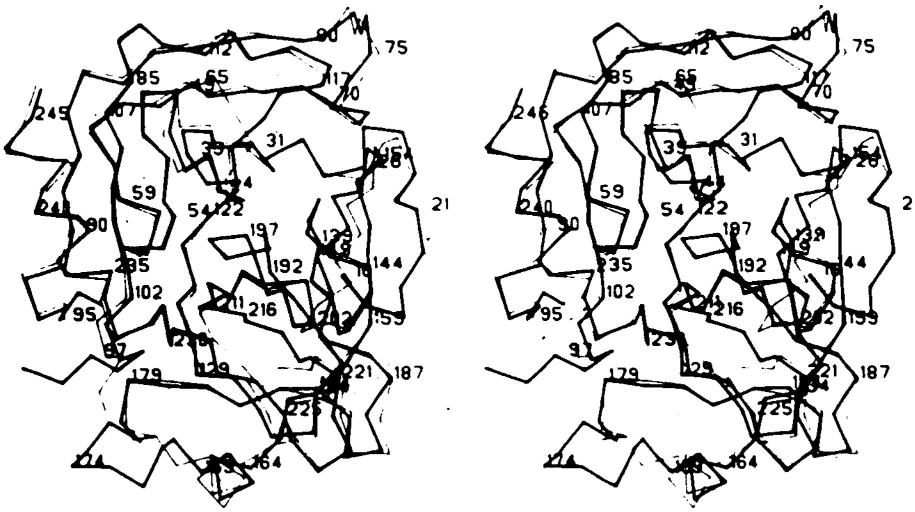

Figure III.12. Structures of Tonin and Kallikrein. The structure of tonin (thick lines) is compared to that of kallikrein (thin lines). The molecules are represented by their α-carbon atoms connected by line segments. The amino acids of tonin are labeled every five residues.

Table III.8

Sequence Alignment of Tonin with Kallikrein

	16	17	18	19	20	21	22	23	24	25	26	27	28	29
TN	I	V	G	G	Y	K	C	E	K	N	S	Q	P	W
KA	I	I	G	G	R	E	C	E	K	N	S	H	P	W

	30	31	32	33	34	35	36	38	39	40	41	42	43	44
TN	Q	V	A	V	I	N	-	-	E	Y	L	C	G	G
KA	Q	V	A	I	Y	H	Y	S	S	F	Q	C	G	G

	45	46	47	48	49	50	51	52	53	54	55	56	57	58
TN	V	L	I	D	P	S	W	V	I	T	A	A	H	C
KA	V	L	V	N	P	K	W	V	L	T	A	A	H	C

	59	60	61	63	64	65	66	67	68	69	70	71	72	73
TN	Y	S	N	N	Y	Q	V	L	L	G	R	N	N	L
KA	K	N	D	N	Y	E	V	W	L	G	R	H	N	L

						A								
	74	75	77	78	79	79	80	81	82	83	84	85	86	87
TN	F	K	D	E	P	F	A	Q	R	R	L	V	R	Q
KA	F	E	N	E	N	T	A	Q	F	F	G	V	T	A

									A	B	C	D	E	F
	88	89	90	91	92	93	94	95	95	95	95	95	95	95
TN	S	F	R	H	P	D	Y	I	P	L	I	V	T	N
KA	D	F	P	H	P	G	F	N	L	S	-	-	-	-

	G	H	I	J	K									
	95	95	95	95	95	96	97	98	99	100	101	102	103	104
TN	D	T	E	Q	P	V	H	D	H	S	N	D	L	M
KA	-	-	-	A	D	G	K	D	Y	S	H	D	L	M

	105	106	107	108	109	110	111	112	113	114	115	116	117	118
TN	L	L	H	L	S	E	P	A	D	I	T	G	G	V
KA	L	L	R	L	Q	S	P	A	K	I	T	D	A	V

(Table III.8 continued)

	119	120	121	122	123	124	125	128	129	130	131	132	133	134
TN	K	V	I	D	L	P	T	K	E	P	K	V	G	S
KA	K	V	L	E	L	P	T	Q	E	P	E	L	G	S

														A
	135	136	137	138	139	140	141	142	143	144	145	146	147	147
TN	T	C	L	A	S	G	W	G	S	T	N	P	S	-
KA	T	C	E	A	S	G	W	G	S	I	E	P	G	P

	B													
	147	148	149	150	151	152	153	154	155	156	157	158	159	160
TN	-	E	M	V	V	S	H	D	L	Q	C	V	N	I
KA	D	D	F	E	F	P	D	E	I	Q	C	V	Q	L

	161	162	163	164	165	166	167	168	169	170	171	172	173	174
TN	H	L	L	S	N	E	K	C	I	E	T	Y	K	D
KA	T	L	L	Q	N	T	F	C	A	D	A	H	P	D

													A	B
	175	176	177	178	179	180	181	182	183	184	185	186	186	186
TN	N	V	T	D	V	M	L	C	A	G	E	M	E	G
KA	K	V	T	E	S	M	L	C	A	G	Y	L	P	G

	187	188	189	190	191	192	193	194	195	196	197	198	199	200
TN	G	K	D	T	C	A	G	D	S	G	G	P	L	I
KA	G	K	D	T	C	M	G	D	S	G	G	P	L	I

	201	202	203	207	209	210	211	212	213	214	215	216	217	218
TN	C	D	G	V	L	Q	G	I	T	S	G	G	A	T
KA	C	N	G	M	W	Q	G	I	T	S	W	G	H	T

					A									
	219	220	221	222	222	223	224	225	226	227	228	229	230	231
TN	P	C	A	K	P	K	T	P	A	I	Y	A	K	L
KA	P	C	G	S	A	N	K	P	S	I	Y	T	K	L

	232	233	234	235	236	237	238	239	240	241	242	243	244	245
TN	I	K	F	T	S	W	I	K	K	V	M	K	E	N
KA	I	F	Y	L	D	W	I	D	D	T	I	T	E	N

(Table III.8 continued)

	246
TN	P
KA	P
	. . . .

The sequence alignment of tonin (TN) with that of kallikrein(KA). The alignment was done based on the structural overlap of the two molecules based on the algorithm of Rossmann & Argos (1975). A structural alignment of each enzyme with α-chymotrypsin was used to derive the sequence numbering shown. The solid and dotted lines under the sequences indicate regions of structural similarity. Residues that are within 1.0Å of each other after the superposition of the two structures are underlined with solid lines whereas the dotted lines indicate the additional fifteen residues that are within 1.9Å of each other.

Ser95B and it is not known how much of the chain is missing, the residues in this region have been numbered 95A, 95B, 95Y, 95Z (Fiedler & Fritz, 1981). On the other hand, the complete sequence of tonin is known so the corresponding residues were labeled consecutively instead.

As expected from the close tertiary structural homology, the secondary structures of tonin and kallikrein are almost identical. The main differences occur where there are insertions or deletions. There are six regions in tonin in which the path of the polypeptide chain differs greatly from that in kallikrein (Table III.8). These are residues Asn25-Ser26, Glu39-Leu41, Pro95A-Asn101, Ser147-Ser152, Ile169-Asp178, and Gly215-Pro219. All but the first region are located near the active site of the enzyme are probably regions involved in substrate specificity. It was also found in comparing the structures of kallikrein and trypsin that the differences in the two structures were mainly in the regions surrounding the active site (Bode *et al*., 1983).

It is instructive, in light of the recent interest in comparative model building of protein structures based on sequence homology (Greer, 1981; Read *et al*., 1984), to look at these regions in more detail. In such model building studies, it will be the parts of the molecule that determine the specificity that will be the most important but also the most difficult to predict.

The first region, around Asn25 and Ser26, is in a turn (Figure III.13) located far from the active site. The

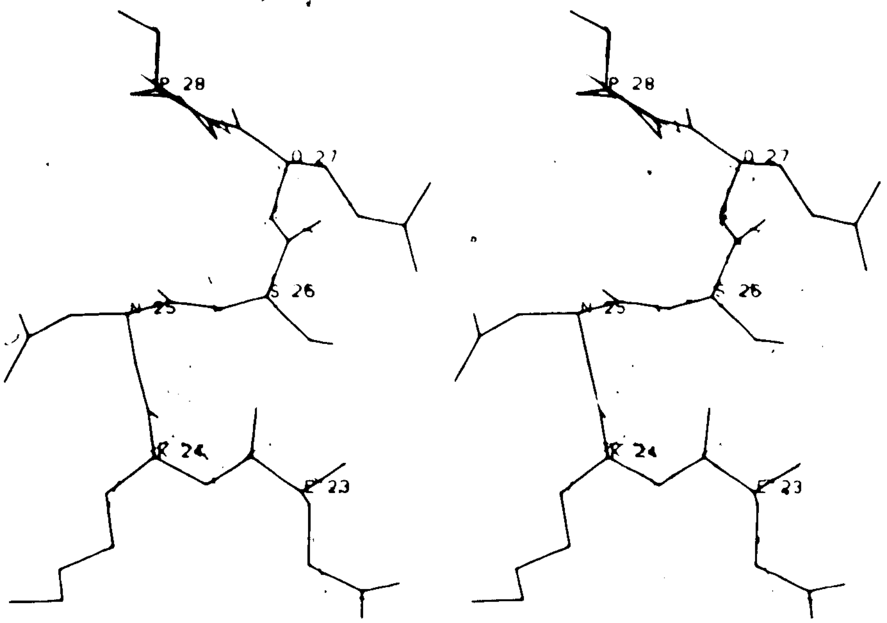

Figure III.13. Comparison of Tonin and Kallikrein from Residue 23 to 28. Tonin is shown in thick lines and kallikrein in thin lines. The small differences in the structures is due to the change from a type I turn in kallikrein to a type II turn in tonin.

relatively small differences in the two structures are a result of changing from a type I turn in kallikrein to a type II turn in tonin. Since the amino acids in the turn Glu23-Ser26 are the same in the two structures, the conformation of the turn must be dictated by the environment outside of these residues. The conformation in tonin in this region is in fact very similar to the corresponding part in trypsin with the sequence, Gly23-Ala24-Asn25-Thr26.

The next segment, Glu39 to Leu41, involves an insertion of two residues in kallikrein with respect to the tonin

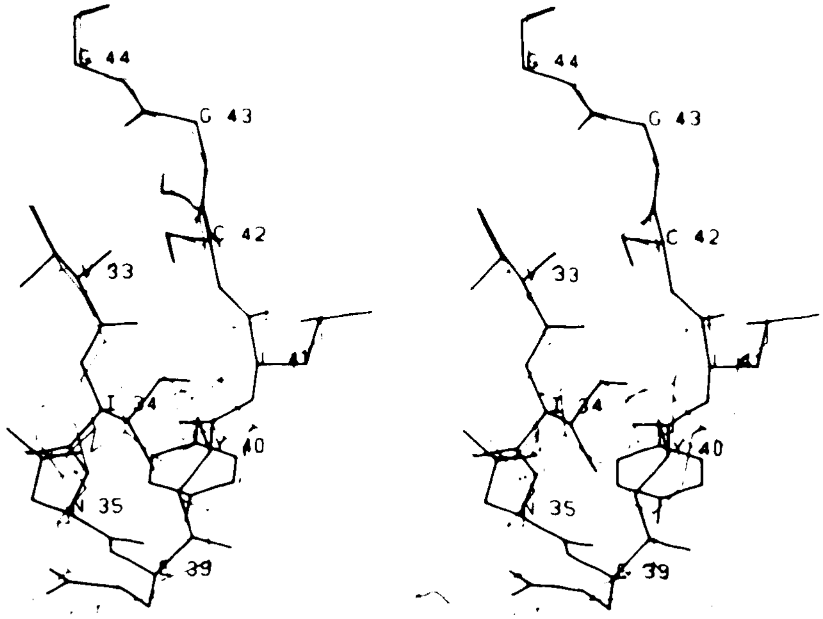

Figure III.14. Comparison of Tonin and Kallikrein from Residue 33 to 44. Tonin is shown in thick lines and kallikrein in thin lines. There is a two residue insertion in kallikrein with respect to tonin.

sequence (Figure III.14). In kallikrein, this region is a β hairpin turn with a classic β-bulge (Richardson *et al.*, 1978) at residues Gln41 and Cys42. The shorter loop in tonin forms a three residue β-hairpin with a hydrogen bond from Asn35 N to Tyr40 O. This type of hydrogen bond between the NH group of a residue to the CO group of a residue that is two residues ahead is rarely observed (Kabsch & Sander, 1983). In addition Asn35 is involved in an Asx turn to Tyr40, providing additional stability to this unusual bend. The side chains of Tyr40 in tonin and Phe40 in kallikrein occupy totally different positions, contrary to what one might expect from the homology in the sequence.

The region of the kallikrein loop, around Pro95A to Asn101, is expected to be quite different in the two structures because of the additional seven residues in tonin with respect to kallikrein (Table III.8). In addition the kallikrein molecule is cleaved in this loop and the position of the ends of the chains could only be inferred from the structure of kallikrein-pancreatic trypsin inhibitor complex (Chen & Bode, 1983). This region in tonin is also not defined with residues Ile95C to Gln95J totally absent from the electron density. Figure III.15 shows the superposition of the kallikrein loop regions from the two enzymes. Further perturbations may be introduced in the tonin structure by the Zn^{2+} ion which binds to His97 and His99.

The loop containing the residue Ser147 to Ser152 in tonin adopts a conformation similar to the corresponding

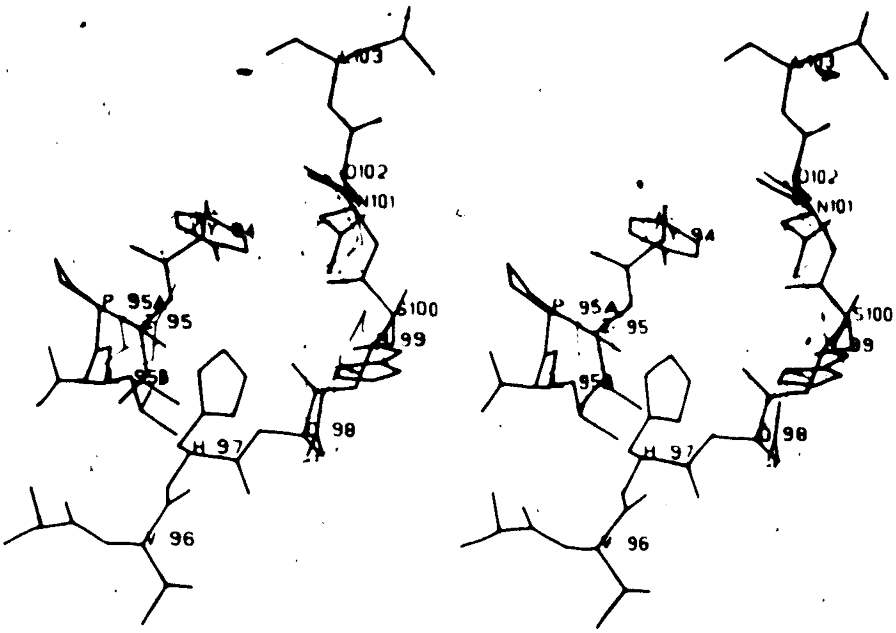

Figure III.15. Comparison of Tonin and Kallikrein in the Region of the Kallikrein Loop. Tonin is shown in thick lines and kallikrein in thin lines. Eight residues from Ile95C to Gln95J in tonin could not be located in the electron density map. Pro95K was refined as as alanyl residue. Kallikrein is cleaved between Ser95B and Ala95J and the positions of residues Leu95A to Gly96 were not determined (Bode *et al.*, 1983).

loop in kallikrein which has two extra residues (Figure III.16). The conformation in tonin is quite similar to the corresponding region in trypsin where the number of residues involved is the same (Figure III.3). The side chain of Glu148 is one of the ligands of the Zn^{2+} ion along with three other histidyl residues from another molecule.

The largest differences between the two structures occur in the region of residues Ile164 to Asp178 (Figure

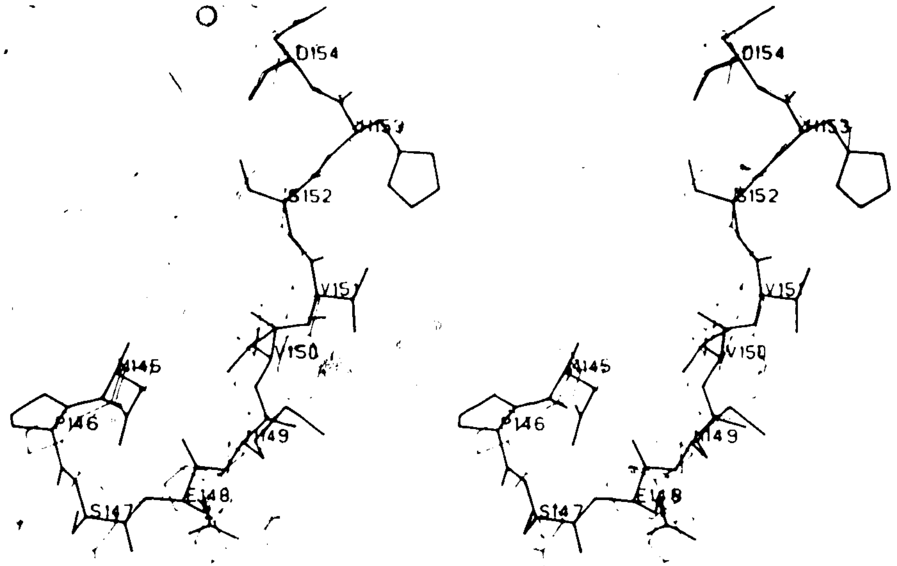

Figure III.16. Comparison of Tonin and Kallikrein from Residue 145 to 154. Tonin is shown in thick lines and kallikrein in thin lines. There is an insertion of two residues, Pro147A and Asp147B, in kallikrein with respect to tonin.

III.17). This is also the part that may be responsible for the specificity that tonin has for residues far from the scissile bond in the P_7 and P_8 positions. There is a change in the secondary structure in this region, from an α-helix in kallikrein (Asn165 to Ala171) to two short 3_{10} helices in tonin (Asn165 to Lys167 and Ile169 to Thr171). In kallikrein the helix is followed by a bend at Pro173, then a stretch of extended polypeptide chain and a type I turn starting at Thr177. In contrast the 3_{10} helix in tonin is followed by a type II' turn starting at Lys173 and another short 3_{10} helix from Val176 to Val179. The type II' turn

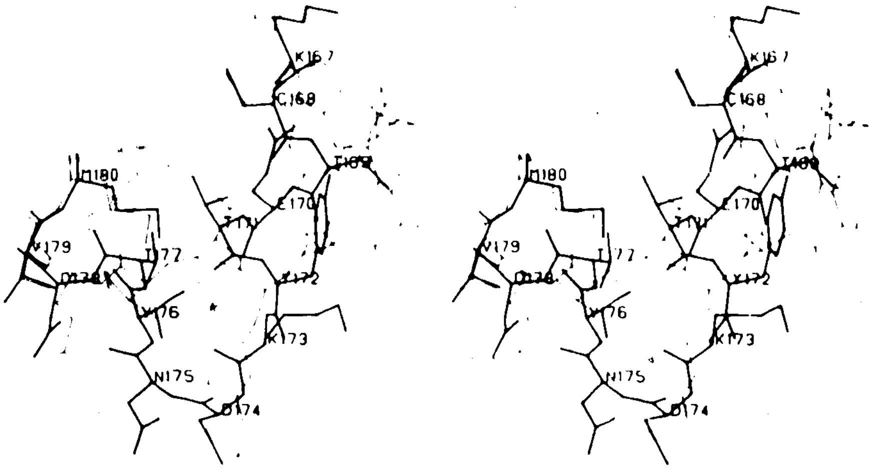

Figure III.17. Comparison of Tonin and Kallikrein from Residue 167 to 180. Tonin is shown in thick lines and kallikrein in thin lines. There are extensive changes in the secondary structure in this region, even though the number of residues are the same.

places Asp174 in an unfavourable conformation (Figure III.10) as discussed previously. A glycyl residue at this position is normally expected in a type II' turn (Sibanda & Thornton, 1985). Since the are no insertions or deletions in this region, the observed differences in the two structures would be very hard to predict.

The conformation of residues Gly215 to Pro219 was totally unexpected and differs drastically from that in the corresponding region in all other serine proteases of known structure (Figure III.18). This is a segment of polypeptide chain that normally forms the opening of the primary specificity pocket and is also involved in secondary binding. In tonin, however, the polypeptide is shifted effectively

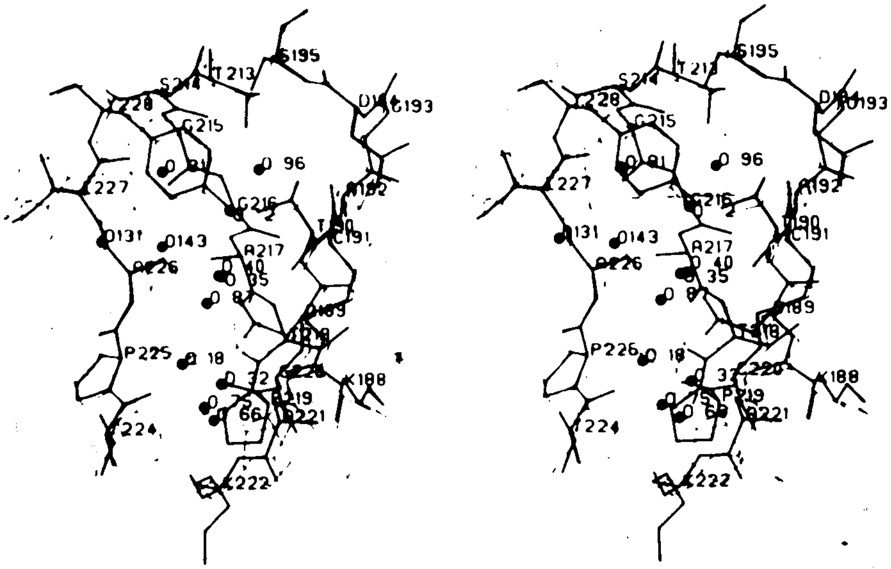

Figure III.18. Comparison of Tonin and Kallikrein in the Binding Loop. Tonin is shown in thick lines and kallikrein in thin lines. There are large differences in the segment from Ser214 to Cys220. Many water molecules are found in the primary specificity pocket around the carboxyl group of Asp189.

closing off the normal opening of the pocket. The new conformation has resulted in another opening between this strand and the strand containing residues Pro225-Tyr228 (Figures III.8, III.11 and III.17). It is not clear the reason for the very different conformation in tonin. There would not be any close contacts if this segment had the conformation found in kallikrein. The occurrences of two glycine residues in a row must make this region more flexible and it is possible that the hydrogen bond from Thr218 to the same residue in a symmetry related molecule causes the polypeptide to shift. Aside from this segment, the binding pocket of the two enzymes are very similar. Asp189 that give these enzymes a primary specificity for basic residues occupy equivalent positions in the two structures. The serine at position 226 that is hydrogen bonded to Asp189 in kallikrein is changed to an alanyl residue in tonin. The corresponding position in trypsin is occupied by a glycine.

Figure III.19 shows the regions of the active site. The main differences between the two structures are in the binding loop (Gly215 to Pro219) discussed above, and the conformation of the side chain of His57. The imidazole ring has shifted in order that it can act as a ligand for the Zn^{2+} ion along with the side chains of His97 and His99. This change has been achieved without losing the hydrogen bonded interaction to the Asp102 carboxyl group from the ND1 of the histidine. There is a small change in the orientation of the Asp102 side chain, perhaps to accommodate the new position of His57.

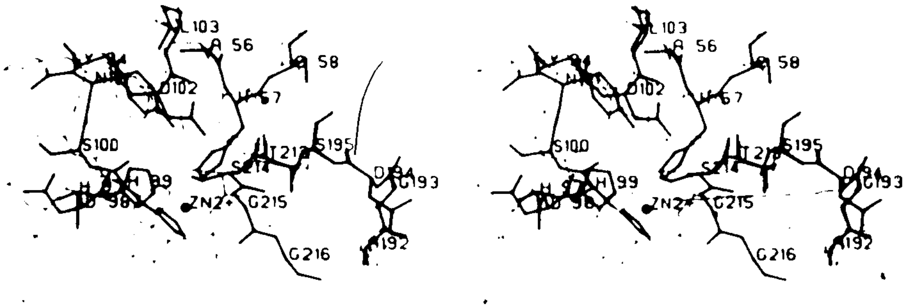

Figure III.19. Comparison of the Catalytic Residues of Tonin and Kallikrein. Tonin is shown in thick lines and kallikrein in thin lines. The side chain of His57 is rotated away, from the position found in other serine proteases, to bind the Zn^{2+} ion.

The Zn^{2+} Ion and its Environment

There are many lines of evidence for the presence of Zn^{2+} ion in the structure. First it occupies a region of very high electron density and refines well as a Zn^{2+} ion with a temperature factor of 6.1 Å^2 and an occupancy of 0.85. Zinc is the heaviest atom in the crystallization medium, all the other atoms being much lighter. Secondly the coordination is roughly tetrahedral with bonding distances expected for Zn^{2+} (Table III.9) with four ligands. The relative positions of the ligands can also be seen in Figure III.7. The coordination is provided by three imidazole groups from His57, His97 and His99 and the carboxyl group of Glu148 from a neighboring molecule. Finally the distortion of the conformation of the active

Table III.9

Geometry of the Zn Coordination

Distances (Å)	
Zn ··········His57 NE2	2.04
Zn ··········His97 NE2	2.05
Zn ··········His99 NE2	2.06
Zn ··········Glu148' OE2	2.07
Angles (°)	
His57 NE2 ···Zn ··His97 NE2	99
His57 NE2 ···Zn ··His99 NE2	99
His57 NE2 ···Zn ··Glu148' OE2	101
His97 NE2 ···Zn ··His99 NE2	100
His97 NE2 ···Zn ··Glu148' OE2	144
His99 NE2 ···Zn ··Glu148' OE2	99

'from a symmetry related molecule

site histidine (His57) is expected to prevent the enzyme from functioning. Indeed the presence of Zn^{2+} inhibits the activity of tonin to cleave p-tosyl-L-arginine methyl ester at pH 6.5 (unpublished result). The possibility that the Zn^{2+} bound form of tonin can cleave angiotensin I is eliminated because that reaction is not inhibited by EDTA (Boucher *et al.*, 1972).

Conclusions

It is expected that the native form of tonin, without a bound Zn^{2+} ion, will resemble more closely the structure of kallikrein in the region of the active site. The side chain of His57 should be rotated close to Ser195 so that it can participate in the catalytic reaction. The binding loop from Ser214 to Cys220 should also have the conformation that is found in the corresponding segment in kallikrein. This conformation will allow the productive binding of a substrate.

Assuming that the native conformation of tonin is as described above, it is still not possible to explain fully the abnormal behavior of this enzyme. The ability of tonin to cleave a substrate at a phenylalanyl residue can be partially understood by noting that the upper part of the primary binding pocket is designed to bind apolar atoms. This is the region that binds the aliphatic parts of arginine and lysine. The benzyl moiety of benzamidine also interacts with this region in kallikrein (Bode *et al.*, 1983)

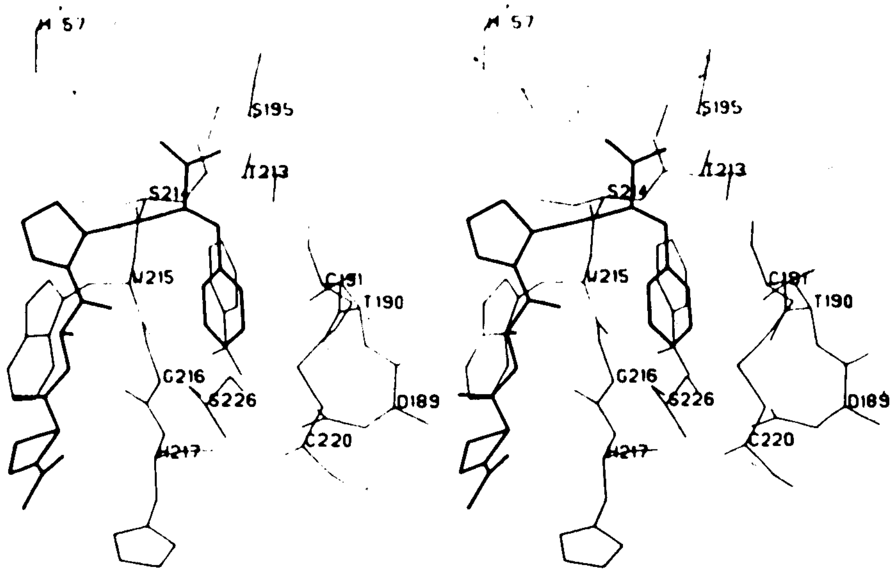

Figure III.20. Superposition of a Tetrapeptide Product on Kallikrein. The structure of a product acetyl-Pro-Ala-Pro-Phe complexed to SGPA (James *et al*., 1980) was superposed on the structure of kallikrein (Bode *et al*., 1983) by aligning the enzyme structures by the method of Rossmann & Argos (1975). Shown are the active site region of kallikrein (thin lines) with a benzamidine molecule bound in the P1 pocket. The position of the phenyl group on the tetrapeptide (thick lines) overlap the benzamidine molecule, indicating that the upper part of the binding pocket accepts aromatic groups.

and in trypsin (Bode & Schwager, 1975). When the structure of a bacterial serine protease, *Streptomyces griseus* protease (SGPA) with a bound product acetyl-Pro-Ala-Pro-Phe (James *et al*., 1980) is superposed on the structure of kallikrein, the phenyl ring of the bound product in SGPA overlaps the position of the benzamidine ring in the kallikrein structure. Of course the binding of phenylalanine in the primary specificity pocket of tonin, kallikrein, or trypsin will result in burying the charged side chain of Asp189 at the bottom of the pocket. It is not clear how such an interaction can become favourable near neutral pH where the hydrolysis of angiotensin I at a phenylalanyl residue by tonin is known to occur (Thibault & Genest, 1981).

Bibliography

Arakawa, K., Ikeda, M., Fukuyama, J., & Sakai, T. (1976) *J. Clin. Endocr. Metab. 42*, 599-602.

Arakawa, K. & Maruta, H. (1980) *Nature (London) 288*, 705-706.

Arakawa, K., Yuki, M., & Ikeda, M. (1980) *Biochem. J. 187*, 647-653.

Ashley, P.L. & MacDonald, R.J. (1985a) *Biochemistry 24*, 4520-4527.

Ashley, P.L. & MacDonald, R.J. (1985b) *Biochemistry 24*, 4512-4520.

Baker, E.N. & Hubbard, R.E. (1984) *Prog. Biophys. Molec. Biol. 44*, 97-179.

Barry, C.D., Molnar, C.E., & Rosenberger, F.U. (1976) *Tech. Memo #229 Computer Systems Lab.*, Washington University, St. Louis, Mo.

Bode, W. & Schwager, P. (1975) *J. Mol. Biol. 98*, 693-717.

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

Boucher, R., Saidi, M., & Genest, J. (1972) In *Hypertension '72* (Genest, J. & Koiw, E., eds), pp. 512-523, Springer-Verlag, New York.

Boucher, R., Asselin, J., & Genest, J. (1974) *Circ. Res. 34-35*(suppl I), I203-I212.

Boucher, R., Demassieux, S., Garcia, R., Gutkowska, J., & Genest, J. (1977) *Union Med. Canada 106*, 502-507.

Chen, Z. & Bode, W. (1983) *J. Mol. Biol. 164*, 283-311.

Chrétien, M., Lee, C.M., Sandberg, B.E.B., Iversen, L.L., Boucher, R., Seidah, N.G., & Genest, J. (1980) *FEBS Letters 113*, 173-176.

Crowther, R.A. (1973) In *The Molecular Replacement Method* (Rossmann, M.G., ed.), pp. 173-178, International Science Review 13, Gordon & Breach Inc., N.Y.

Cruickshank, D.W.J. (1949) *Acta Cryst. 2*, 65-82.

Cruickshank, D.W.J. (1954) *Acta Cryst. 7*, 519.

Cruickshank, D.W.J. (1967) In *International Tables for X-ray Crystallography*, Vol.II, (Kasper, J.S. & Londsdale, K. eds), pp. 318-340, Kynoch Press, Birmingham, England.

Fehlhammer, H. & Bode, W. (1975) *J. Mol. Biol. 98*, 683-692.

Fiedler, F. & Fritz, H. (1981) *Hoppe-Seyler's Z. Phsiol. Chem. 362*, 1171-1175.

Fujinaga, M., Delbaere, L.T.J., Brayer, G.D., & James, N.M.G. (1985) *J. Mol. Biol. 184*, 479-502.

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

Grisé, C., Boucher, R., Thibault, G., & Genest, J. (1981) *Can. J. Biochem. 59*, 250-255.

Harada, Y., Lifchitz, A., Berthou, J., & Jolles, P. (1981) *Acta Cryst. A37*, 398-406.

Hayakawa, K., Kelly, J.A., & James, M.N.G. (1978) *J. Mol. Biol. 123*, 107-111.

Hendrickson, W.A. (1976) *J. Mol. Biol. 106*, 889-893.

Hendrickson, W.A., & Konnert, J.H. (1980) In *Biomolecular Structure, Function, Conformation and Evolution* (Srinivasan, R. ed.), Vol. I, pp. 43-57, Pergamon Press, Oxford.

Ikeda, M. & Arakawa, K. (1984) *Hypertension 6*, 222-228.

James, M.N.G. & Sielecki, A.R. (1983) *J. Mol. Biol. 163*, 299-361.

Kabsch, W. & Sander, C. (1983) *Biopolymers 22*, 2577-2637.

Lazure, C., Seidah, N.G., Thibault, G., Genest, J., & Chrétien, J. (1981) In *Proceedings of the seventh American Peptide Symposium* (Rich, D.H. & Gross, E. eds.) pp. 517-519, Pierce Chemical, New York.

Lazure, C., Leduc, R., Seidah, N.G., Thibault, G., Genest, J., & Chrétien, M. (1984) *Nature (London) 307*, 555-558.

Ledoux, S., Gutkowska, J., Garcia, R., Thibault, G., Cantin, M., & Genest, J. (1982) *Histochemistry 76* 329-339.

Luzzati, V. (1952) *Acta Cryst. 5*, 802-810.

Moult, J., Sussman, F., & James, M. N. G. (1985) *J. Mol. Biol. 182* 555-566.

North, A.C.T., Phillips, D.C., & Mathews, F.S. (1968) *Acta Cryst. A24*, 351-359.

Ramakrishnan, C. & Ramachandran, G.N. (1965) *Biophys. J. 5*, 909-933.

Read, R.J. (1986) *Acta Cryst. A42* 140-149.

Read, R.J., Fujinaga, M., Sielecki, A.R., & James, M.N.G. (1983) *Biochemistry 22*, 4420-4433.

Rees, D.C., Lewis, M. & Lipscomb, W.N. (1983) *J. Mol. Biol. 168*, 367-387.

Richardson, J.S., Getzoff, E.D., & Richardson, D.C. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 2574-2578.

Rossmann, M.G. & Argos, P. (1975) *J. Biol. Chem. 18*, 7525-7532.

Schechter, I., & Berger, A. (1967) *Biochem. Biophys. Res. Commun. 27*, 157-162.

Seidah, N.G., Chan, J.S.D., Mardini, G., Benjannet, S., Chrétien, M., Boucher, R., & Genest, J. (1979) *Biochem. Biophys. Res. Commun. 86*, 1002-1013.

Schiller, P.W., Demassieux, S., & Boucher, R. (1976) *Circ. Res. 39*, 629-632.

Sibanda, B.L. & Thornton, J.M. (1985) *Nature (London) 316*, 170-174.

Sielecki, A.R., James, M.N.G., & Broughton, C.G. (1981) In *Crystallographic Computing* (Sayer, D., ed.) Proc. of Intl. Summer School, Carleton University, Ottawa, pp. 409-419, Oxford University Press, Oxford.

Srinivasan, R. & Chandrasekaran, R. (1966) *Indian J. Pure Appl. Phys. 4*, 178-186.

Thibault, G. & Genest, R. (1981) *Biochim. Biophys. Acta 660*, 23-29.

Thiessen, W.E., & Levy, H.A. (1973) *J. Applied Crystallogr. 6*, 309.

Tremblay, J., Thibault, G., Gutkowska, J., Boucher, R., & Genest, J. (1981) *Can. J. Biochem. 59*, 256-261.

Venkatachalam, C.M. (1968) *Biopolymers, 6*, 1425-1436.

IV. Experiences with a New Translation Function

The physical process of diffraction of x-rays by a crystal can be described by a mathematical device called the Fourier transform (FT). The transform and its inverse are defined for the one-dimensional case as follows:

$$g(t) = (1/2\pi)\int_{-\infty}^{\infty} G(w)\exp(iwt)dw$$

$$G(w) = \int_{-\infty}^{\infty} g(t)\exp(-iwt)dt$$

The Fourier transform defines a relationship between two functions and converts one into the other.

$$g \overset{FT}{\longleftrightarrow} G$$

An analogous transform can be defined for the three-dimensional case and it is used to model the relationship between the positions of the electrons that scatter the x-rays and the diffracted beam as follows.

$$\rho \overset{FT}{\longleftrightarrow} \mathbf{F}$$

ρ is the electron density and **F** represents the diffracted x-ray beam and is called the structure factor. **F** is a complex quantity consisting of an amplitude and a phase. The domain of ρ is called the real space and that of **F** is called the reciprocal space. If **F** is known then ρ can be calculated using the Fourier transform and determination of crystal structures would be trivial. However only the amplitude of **F**, which is proportional to the intensity of the diffracted beam, can be measured and the phase angle cannot be directly determined. This is the 'phase problem' of crystallography. Values for the phase angles must somehow be obtained in order for a crystal structure to be solved. The phase problem can be overcome by different methods depending on the nature of the structure. For a small molecule, with less than about 100 atoms, either Patterson methods or direct methods usually lead to a solution.

Patterson methods rely on the presence in the structure of an atom that is much heavier than the other atoms. The position of this heavy atom is first located by using the Patterson function (Patterson, 1935). This function is defined by,

$$P(u) = (1/V)\Sigma|F_O(h)|^2\cos2\pi(h\cdot u)$$

where V is the crystal cell volume, $|F_O(h)|$ are the measured structure factor amplitudes, h are the reciprocal lattice vectors, and u are position vectors given in the fractional coordinate system. The summation is done over the reciprocal lattice points. The function can be calculated prior to having any phase information since it uses only the observed structure factor amplitudes. The function has a peak at every position corresponding to all the inter-atomic vectors in the crystal and the size of the peak is proportional to the scattering power of the atoms at the ends of the vector. Therefore the peaks corresponding to vectors between heavy atoms can be picked out and the position of the heavy atom determined. If the heavy atom makes up a substantial fraction of the total scattering matter in the crystal, the phases calculated using only the heavy atom will approximate the true phase and the positions of the lighter atoms can then be determined.

Direct methods, as the name implies, is a means of obtaining the phase information directly from the relationships that exist among the structure factor amplitudes. It is based on probabilistic formulae that result from assumptions about positivity and atomicity of the electron density. As the number of atoms in the molecule increases, these relationships become less reliable and the structure solution becomes more difficult. A detailed account of the subject can be found in Ladd & Palmer (1980). Many small molecules are now solved using direct methods with the available program packages (eg. Main *et al*., 1980) that allow automated structure solution.

For protein molecules, the above methods do not work due to the large number of atoms involved. The usual approach to the solution of the phase problem is to use the method of multi-isomorphous replacement (MIR). It was first successfully used for the structure solution of hemoglobin (Green *et al*., 1954). The method relies on the replacement or the addition of a heavy atom without perturbing the rest of the structure, that is the structure remains isomorphous. The knowledge of the position of the heavy atom and the changes in the intensities of the diffracted beam that result due to the inclusion of the heavy atom provide phasing information.

Let F_P and F_{PH} be the structure factors of the protein and the protein with a heavy atom, respectively. If F_H is the contribution to the structure factor due to the heavy atom, then

$$F_{PH} = F_P + F_H$$

The position of the heavy atom can be determined using the Patterson function in a manner similar to the small molecule case so that both the amplitude and the phase of F_H can be determined. The relationship between the observed amplitudes, $|F_{PH}|$ and $|F_P|$, and F_H can be understood most easily from an Argand diagram (Figure IV.1). It shows that the phase angle of F_P is restricted to two possible values. In an ideal case two isomorphous heavy atom derivatives with

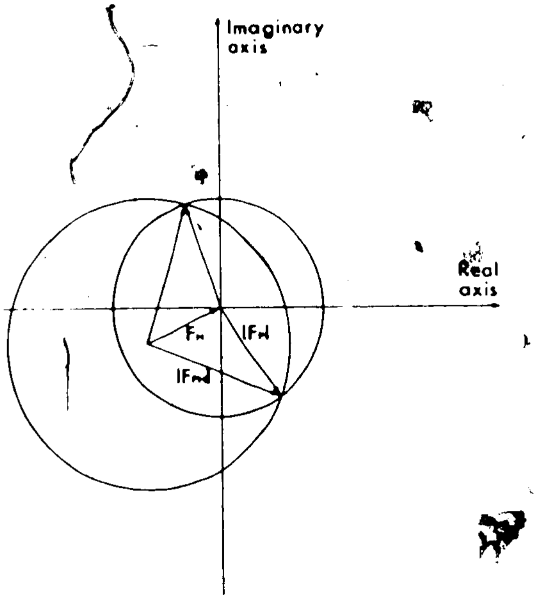

Figure IV.1. The Isomorphous Replacement Method. Argand diagram showing the relationship among $|F_P|$, $|F_{PH}|$ and F_H. The knowledge of the phase of F_H restricts the phase of F_P to one of two values since $F_{PH}=F_P+F_H$.

the heavy atoms occupying different positions are required to assign unique phases to the native structure factors, F_P. In practice, due to the errors in the data and non-isomorphism, more than two derivatives are required hence the name, multi-isomorphous replacement. In addition, the preparation and screening of heavy atom derivatives is often a very time consuming process and success is not guaranteed.

An alternative approach is available if the unknown structure is related to a known structure. This is the technique of molecular replacement (Rossmann, 1972). Consider a protein of unknown structure for which there are crystallographic data. If there is a protein whose structure is known and whose sequence is similar to the protein of interest, then these proteins are expected to have similar three-dimensional structures and the known structure will serve as an approximate model of the unknown one in overcoming the phase problem. This is done by positioning the model in the crystal cell to superimpose on the unknown structure. The phases, α_C, can then be calculated from this model using the equation for the structure factor given by

$$\begin{aligned} \mathbf{F}_C(\mathbf{h}) &= |\mathbf{F}_C(\mathbf{h})|\exp(i\alpha_C) \\ &= \sum_j f_j \exp[2\pi i(\mathbf{h}\cdot\mathbf{x}_j)] \end{aligned}$$

where $\mathbf{F}_C$ is the calculated structure factor, f_j are the atomic scattering factors and $\mathbf{x}_j$ are the coordinates of the atoms in the model. These calculated phases are then used as approximations of the true phases. Molecular replacement is then concerned with finding the three rotational and three translational parameters that specify the orientation and position, respectively, of the molecule in the crystal cell with respect to the symmetry elements.

Many of the techniques used in molecular replacement are based on the Patterson function and it is useful to go into more detail here. As mentioned before, the Patterson function has peaks at positions corresponding to inter-atomic vectors. These vectors can be thought of as being made up of two groups. One is the set consisting of intra-molecular vectors or vectors between atoms of the same molecule, and the other is the set of inter-molecular vectors or those between atoms of different molecules. The intra-molecular vectors contain only the information concerning the orientation of the molecule whereas the inter-molecular vectors also have information about the position of the molecule in the crystal. Consideration of these two groups of vectors allows the determination of the orientation first, independently of the position.

The three rotational parameters are usually determined from the rotation function proposed by Rossmann and Blow (1962). It is based on the the idea of maximizing the agreement of the intra-molecular vectors between the observed and calculated Patterson functions. Even though the inter- and intra-molecular vectors in the observed Patterson function cannot be separated, by placing the model structure in an artificially large unit cell, the two sets of vectors for the calculated Patterson can be distinguished. The function given by Rossmann & Blow (1962) is given by,

$R(\theta_1,\theta_2,\theta$ [illegible] $P_C(u')P_O(u)du$

where θ_1, θ_2, θ_3 specify the orientation of the model, usually in terms of Eulerian angles, and u' is in a rotated coordinate system corresponding to these angles. P_C and P_O are the Patterson functions of the model and observed structures, respectively. The limits of integration are chosen to include all the intra-molecular vectors of P_C and to reject the inter-molecular vectors. The function measures the overlap of the two Patterson functions as the model is rotated and should be maximal at the orientation where the agreement between the intra-molecular vectors of tne model and of the observed structure is highest. The actual computation of this function is done in reciprocal space. Crowther (1972) proposed a fast algorithm which expands the Patterson function in terms of spherical harmonics, allowing for the calculation to be performed using the fast Fourier transform (FFT) which is an algorithm devised by Cooley & Tukey (1965) for carrying out certain types of Fourier transforms very efficiently. The rotational parameters are thus obtained independently of the position of the model in the cell. Once the orientation is known the translational parameters can be determined.

It has been more difficult to find a satisfactory solution to the translation problem as indicated by the numerous translation functions that exist (Tollin, 1966; Crowther & Blow, 1967; Hendrickson & Ward, 1976; Harada *et al.*, 1981; Langs, 1985). The one of Crowther and Blow (1967) is often used. It is similar to the rotation function in that the correlation of the observed and the model Patterson functions are calculated with a product function. In this case, however, the model Patterson function consists of the intermolecular vectors of molecules related by a symmetry operation.

$$T(t) = \int_V P_{ij}(u,t)P_o(u)du$$

where P_{ij} is the Patterson due to symmetry related molecules i and j of the model and P_o is the observed Patterson. The intermolecular vector between molecules i and j is given by t. The integral is taken over the cell volume, V.

The expression is evaluated in reciprocal space using FFT. The performance of the function can be improved by subtracting out the intra-molecular vector contributions from P_o. This is done by substituting for $|F_o(h)|^2$ in calculating P_o, the expression,

$$|F_o(h)|^2-\sum_j|F_M(hR_j)|^2$$

where F_M are the structure factors due to the model structure and R_j are the rotation matrices of the crystallographic symmetry operations. The summation is over all the symmetry operations. Tollin (1966,1969) has shown that his

Q-function is similar to Crowther and Blow's translation function but expressed in terms of the sum function (Buerger, 1959).

Another commonly used procedure is to calculate the R-factor ($R=\Sigma||F_o|-|F_c||/\Sigma|F_o|$) as the model is translated in the cell. This is not computed as efficiently as is the translation function since FFT cannot be used but with increasing computer speeds this is not a serious drawback. On the other hand it is sensitive to errors in scaling of $|F_o|$ to the absolute scale that $|F_c|$ is on. Such an error may result in an incorrect solution. Nevertheless many closely related structures have been solved using this method (Rossmann, 1980).

More recently Harada *et al.* (1981) have introduced a function that combines a product function, similar to the one defined by Crowther & Blow (1967), and an overlap function that measures the amount that the atoms of symmetry related molecules of the model overlap. They began with a correlation coefficient defined as,

$$C' = \frac{\Sigma|F_o|^2|F_c|^2}{[\Sigma|F_o|^4\Sigma|F_c|^4]^{1/2}}$$

and then derived a quantity that is more easily computed. The numerator of the new function is a measure of the agreement between observed and calculated structures and is given by,

$$TO(r) = \frac{\Sigma|F_o(h)|^2|F_c(h,r)|^2}{\Sigma|F_o(h)|^4}$$

where r is the position of the molecule in the cell. The expression can be shown to be proportional to the product function of the observed and calculated Patterson functions since by Parseval's theorem (Bracewell, 1965),

$$TO(r) \propto \int_V P_o(u)P_c(u,r)du$$

Unlike Crowther and Blow's translation function (1967), the calculated Patterson, P_c, corresponds to a complete set of vectors between all symmetry related molecules.

The denominator of the function is an overlap function and is defined as,

$$O(r) = \frac{\Sigma|F_c(h,r)|^2}{N\Sigma|F_m(h)|^2}$$

where N is the number of symmetry operations and F_m is the molecular structure factor or the contribution to the structure factor due to one molecule. This function is proportional to the origin peak in the calculated Patterson function, representing vectors from every atom to itself and any other atom pairs that are very close together, and thus would be a measure of the overlap of the molecules in the cell. The correct structure should not have atoms that are very close together and so the overlap will be minimal for such a structure. The complete function is then,

$$T'(r) = \frac{TO(r)}{O(r)}$$

The function is maximal when the agreement between the inter-molecular vectors of the observed and calculated Patterson functions is large and the overlap among the molecules is small. Harada *et al*. have shown that the function can be evaluated using FFT.

A. The Correlation Coefficient

We were inspired by the work of Harada *et al*. (1981) to use the correlation coefficient for the solution of the translation problem. Unlike them we chose to work with the standard linear correlation coefficient defined as follows:

$$C = \frac{\Sigma(|F_o|^2-\overline{|F_o|^2})(|F_c|^2-\overline{|F_c|^2})}{[\Sigma(|F_o|^2-\overline{|F_o|^2})^2\Sigma(|F_c|^2-\overline{|F_c|^2})^2]^{1/2}}$$

This quantity varies from -1 to 1 unlike the correlation coefficient, C', defined by Harada *et al*. (1981) that is limited by 0 and 1. Like the R-factor, the correlation coefficient cannot be computed using FFT methods but unlike the R-factor, it is insensitive to scaling errors.

Using Parseval's theorem, one can show that C' is equal to the corresponding quantity for Patterson functions. i.e.

$$C' = \frac{\int_V P_o P_c du}{[\int_V P_o{}^2 du \int_V P_c{}^2 du]^{1/2}}$$

and similarly, for summation over a narrow range of resolution, C can be thought of as a correlation of origin-removed Patterson functions.

$$C = \frac{\int_V P_o' P_c' du}{[\int_V P_o'^2 du \int_V P_c'^2 du]^{1/2}}$$

where P_o' and P_c' are origin-removed Patterson functions. It is not immediately obvious whether C includes some measure of molecular overlap since the origin peaks are removed. Overlapping atoms will lead to lower values in both the numerator and the denominator of the expression for C.

Alternatively the correlation coefficient can be interpreted in reciprocal space as a measure of the phase error. A group in Madras has worked out probability distributions for a pair of structures, where one is the observed structure and the other is a model or a trial structure approximating the observed one (Srinivasan & Parthasarathy, 1976). They considered the case of comparing an observed structure with a partial model with errors. For a non-centrosymmetric space group the probability distribution of a pair of normalized structure factor amplitudes is as follows:

$$P(|\mathbf{E}_o|,|\mathbf{E}_c|) = \frac{4|\mathbf{E}_o||\mathbf{E}_c|}{1-\sigma_A^2}\exp\left[-\left[\frac{|\mathbf{E}_o|^2+|\mathbf{E}_c|^2}{1-\sigma_A^2}\right]\right] I_0\left[\frac{2\sigma_A|\mathbf{E}_o||\mathbf{E}_c|}{1-\sigma_A^2}\right]$$

where

$\sigma_A = \sigma_1 D$

$\sigma_1{}^2 = (\sum_j^M f_j{}^2)/(\sum_j^N f_j{}^2)$

where M = no. of atoms in the model
N = no. of atoms in the observed structure

$D = \langle\cos(2\pi h \cdot \Delta r_j)\rangle$

Δr_j = coordinate error

$I_0(X)$ = zero order modified Bessel function (Watson, 1958)

The conditional probability distribution for the phase error, $\alpha=\alpha_0-\alpha_c$, given the normalized structure factor amplitudes is,

$$P(\alpha;|E_o|,|E_c|) = K\exp\left[\frac{2\sigma_A|E_o||E_c|\cos\alpha}{1-\sigma_A{}^2}\right]$$

where

$$K = \left[2\pi I_0(2\sigma_A|E_o||E_c|/(1-\sigma_A{}^2))\right]^{-1}$$

This is a unimodal distribution with the maximum at $\alpha=0$ and the width determined by σ_A. The distribution becomes sharper as σ_A becomes larger.

The behavior of various discrepancy indices with respect to σ_A were also examined. One in particular, the normalized Booth-type index using intensities, has a simple relationship to σ_A. The index is defined as,

$$B^{R_1}(I) = \frac{\Sigma(I_o - I_c/\sigma_1{}^2)^2}{\Sigma I_o{}^2}$$

$$= \frac{<(|E_o|^2 - |E_c|^2)^2>}{<|E_o|^4>}$$

This index is equal to $1+\sigma_A{}^2$ assuming that Wilson's statistics hold.

Hauptman (1982), working on a related problem came up with the identical distribution for a pair of normalized structure factor amplitudes. He went on to show that,

$$\sigma_A{}^2 = \frac{<(|E_1|^2 - <|E_1|^2>)(|E_2|^2 - <|E_2|^2>)>}{[<(|E_1|^2 - <|E_1|^2>)^2><(|E_2|^2 - <|E_2|^2>)^2>]^{1/2}}$$

$$\simeq C$$

If the summation is done over a narrow range of resolution, then a correlation coefficient calculated for |F| will be the same as that calculated for |E|. Under this condition, finding the position for the molecular model in the unit cell that maximizes the correlation coefficient is equivalent to minimizing the phase error.

The calculation of correlation coefficients has been implemented in a program called BRUTE (for the brute force technique of finding a solution). It moves the search model over a grid of points in the crystal cell. At each point the symmetry related positions are generated and the structure factors are calculated. The amplitudes of these calculated structure factors are then compared to the observed values using the correlation coefficient, C, as well as the conventional R-factor. The calculation of structure factors, $\mathbf{F}_c$, are done rapidly by the use of molecular scattering factors (Lipson & Cochran, 1957).

Let $\mathbf{x}_{jk} = \mathbf{R}_j\mathbf{x}_{0k}+\mathbf{T}_j$

where $\mathbf{R}_j$, $\mathbf{T}_j$ are the rotation matrix and the translation vector, respectively, of the symmetry operation j of the space group, and $\mathbf{x}_{0k}$ are the coordinates of the atoms of an oriented search model in an asymmetric unit with k = 1 to the number of atoms (NATM). Then for a shift Δ in the coordinates,

$$\begin{aligned} x_{jk}' &= \mathbf{R}_j(x_{0k}+\Delta)+\mathbf{T}_j \\ &= \mathbf{R}_j\mathbf{x}_{0k}+\mathbf{T}_j+\mathbf{R}_j\Delta \\ &= \mathbf{x}_{jk}+\mathbf{R}_j\Delta \end{aligned}$$

and

$$\begin{aligned} F(h) &= \sum_{j}^{NSYM}\sum_{k}^{NATM} f_k\exp2\pi i(h\cdot \mathbf{x}_{jk}') \\ &= \sum_j\sum_k f_k\exp2\pi i(h\cdot(\mathbf{x}_{jk}+\mathbf{R}_j\Delta)) \\ &= \sum_j\sum_k f_k\exp2\pi i(h\cdot x_{jk})\exp2\pi i(h\cdot \mathbf{R}_j\Delta) \\ &= \sum_j G_{Mj}(\mathbf{h})\exp2\pi i(h\cdot \mathbf{R}_j\Delta) \end{aligned}$$

where

NSYM = no. of symmetry operations

$$G_{Mj}(\mathbf{h}) = \sum_{k}^{NATM} f_k\exp2\pi i(\mathbf{h}\cdot x_{jk})$$

= molecular scattering factor

The molecular scattering factors, G_{Mj}, are calculated once for the first grid point and stored for use at subsequent grid points. Therefore the overall calculation time becomes essentially independent of the number of atoms in the model. The computation time is then a function of the number of symmetry operations, the number of reflections, and the number of grid points.

In addition to the basic computations described above, the program has been modified to incorporate some useful features. First, the program allows for the inclusion of a set of atoms whose positions are fixed. Their contributions to the structure factors are added to the part due to the moving set of atoms. This is useful when the orientation and the position of a part of a molecule are both known. Second, the program can make adjustments in the orientation of the search model. A rotational search can be done over a set of grid points corresponding to rotations about each of an orthogonal set of axes. The resulting rotations are nearly mutually orthogonal for small changes in the angles. When combined with a translational search, a six-dimensional search is possible. However for each new orientation, a whole set of molecular scattering factors, G_{Mj}, must be recalculated so that the computation time becomes prohibitively long except for small 6-D searches.

A possible alternative approach would be to use a rigid-body refinement program such as CORELS (Sussman *et al.*, 1977). The utility of using CORELS in improving the molecular replacement solution has been described by Leslie (1985). This is especially true when there are many rigid groups involved. On the other hand, it does have a limit on the size of adjustments it can make to the structure whereas there is no such limitation when using BRUTE.

The program has been written for the Floating Point Systems 164 Attached Processor (FPS164) which is currently driven by the host computer, Amdahl 5870. FPS164 is a parallel pipeline machine which is capable of fast operations on long arrays. BRUTE was written to maximize vector usage and takes advantage of the assembler subroutines supplied with the system for doing the vector and matrix operations. A version of the program written totally in standard FORTRAN also exists.

B. Practical Experiences

In the course of the development of the program, several structures have been solved and a 'standard' procedure has evolved. Some of these structures were trivial to solve as they were known structures in a different crystal environment. Others which used homologous molecules as the search models were more difficult. A unique solution to the translation problem has always been obtained whenever the rotation function result was unambiguous. The program itself started out as an R-factor search program which was modified to look at the correlation of |F|'s and then of $|F|^2$'s as it does now.

The procedure that is outlined here was derived by trial and error and is by no means a rigorous one. It was used for the most difficult case that was solved successfully, the pepsinogen structure (James & Sielecki, 1986).

The first thing to consider when doing a translational search is to choose the resolution of the data to be used which in turn dictates the maximum grid size that must be used. Finer grids must be used for higher resolution data. It seems that about 1/4 of the minimum d-spacing is an appropriate size for the grid interval. It is recommended that a second search which is offset from the first by 1/2 grid interval in each direction be done to insure that no peaks are missed. The choice of the resolution of data to be used depends on the similarity of the model to the unknown structure. Lower resolution data should be used for models which are not very homologous to the unknown molecule. Data between 4 and 5 Å resolution and 1 Å grid intervals have been used with success.

It was found that rotational parameters can and should be refined using BRUTE before any translational searches are done. This is done by specifying only the P1 symmetry and adjusting the angular parameters, in the neighborhood of the orientation obtained from the rotation function, until the correlation is maximized.

Once the orientation has been optimized the complete translational search can be performed. The search area must be the maximum subset of the unit cell that contains a single permissible origin. It is an area that contains a complete set of unique vectors relative to an origin. This means that for example in P1 in which every point can be an origin, there is no need for a translational search. For polar space groups the search area is a plane perpendicular to the polar axis, and for space groups of higher symmetry, a three-dimensional volume must be searched. After a solution has been found, all of the angular and translational parameters can be refined with a 6-D search in the neighborhood.

Since the correlation coefficient is related to the phase error, one should attempt to obtain the highest correlation possible. This may be done by using a different model or by deleting regions of the molecule that are likely to be different or by substituting the amino acid side chains to correspond to the ones in the observed structure. Once the best model is obtained, phases can be derived and an electron density map calculated. There is always a danger, in such a map, of model bias. It has long been known that maps calculated with model phases tend to reproduce the model (Ramachandran & Srinivasan, 1961). However, when the electron density resembles the model it is not often possible to distinguish whether it is due to model bias or due to the model resembling the observed structure. The best way to reduce model bias is to combine the molecular replacement phases with others derived by an independent method such as MIR. There are also a variety of map coefficients and weighting schemes available for reducing model bias (Main, 1979; Read, 1986a). It may also be helpful to make a partial difference map in which parts of the model which are questionable are deleted before the phase calculation.

The capabilities of the program can be demonstrated by examining three of the more difficult cases that have been successfully solved. These are two serine proteases, tonin (Chap. III) and *Streptomyces griseus* trypsin (SGT) (Read, 1986b; Read *et al.*, 1984), and the aspartyl protease zymogen, pepsinogen (James & Sielecki, 1986).

Tonin is a serine protease that is closely related to kallikrein (Bode *et al.*, 1983). It is isolated from the submaxillary gland of rats (Boucher *et al.*, 1972). Its structure has been solved using molecular replacement methods with bovine trypsin (Fehlhammer & Bode, 1975) as the search model. Had the structure of kallikrein been known at the time the structure solution of tonin was being done, it would have served as a much better search model than trypsin. The structures of tonin and kallikrein are very similar with an r.m.s. deviation of 0.6Å for the 203 α-carbon atom pairs that can be superposed to within 1.9Å (Chapter III). The molecular replacement solution was obtained with only a slight difficulty but the interpretation of the resulting electron density map was not straightforward and the refinement of the structure proceeded slowly.

SGT is a bacterial protease whose structure is more similar to that of bovine trypsin than to other bacterial serine proteases. Consequently the structures of trypsin and chymotrypsin were used as molecular replacement models (Read, 1986b). Trouble was encountered in obtaining an unambiguous translation solution. It turned out that the orientation obtained from the rotation function was sufficiently far from the correct one to prevent the translation search from working. The problem was overcome by including in BRUTE the capability to adjust the orientation. The phases derived from the trypsin and chymotrypsin models were combined with MIR phase information. However the heavy atom phases did not make a significant contribution due to their low quality and due to the overestimation of the accuracy of the molecular replacement phases. During the refinement of the structure, a method was developed by Read (1986a) to get a better estimate of the phase errors and to produce a map with reduced model bias. These techniques facilitated the refinement process.

The structure solution of the mammalian aspartyl protease zymogen, pepsinogen, is the most difficult one that has been successfully solved with BRUTE. As the search model, the refined structure of a fungal aspartyl protease, penicillopepsin (James & Sielecki, 1983) was used. The correlation obtained from the translational search was low and the electron density map obtained from the molecular replacement phases was deemed too difficult to interpret. The molecular replacement phases were combined with the MIR phases which were of slightly higher quality than in the case of SGT. A better estimate of the error of the molecular replacement phases was obtained using the procedure of Read (1986). The combined phases consisted mainly of the MIR phase information and the electron density was more interpretable. In this case the additional information provided by molecular replacement was small and may not have been necessary.

A summary of the structure solutions of these three proteins is given in Table IV.1. The experience with these cases shows that fairly dissimilar proteins can be used as models for molecular replacement. However as the structures become more different, less phasing information is obtained to the point that it no longer allows the development of the observed structure from the starting model. That is not to say that in such cases, molecular replacement will be useless. It can still be used for finding or confirming heavy atom sites to be used for MIR. It will also be useful in resolving the ambiguities in the phases obtained from a single isomorphous replacement. Finally, a molecular replacement model will aid in the interpretation of the electron density map.

Table IV.1

Summary of Molecular Replacement Results

	Space Group and Cell Parameters	Search Model	Sequence Homology[1]	Structural Homology[2]	Rotation Function[3]	BRUTE[4]	R Factor[5]
Tonin (235a.a.)[6]	$P4_32_12$ a=48.64Å b=48.64Å c=201.2Å	Trypsin (223a.a.)	40%	0.89Å (191a.a.)	4.4	0.4	48%
SGT (233a.a.)	$C222_1$ a=72.04Å b=50.86Å c=120.4Å	Trypsin (223a.a.)	33%	0.86Å (168a.a.)	6.0	0.34	52%
Pepsinogen (370a.a.)	C2 a=105.8Å b=43.40Å c=88.60Å β=91.4°	Penicillo- pepsin (323a.a.)	35%	1.63Å (275a.a.)	5.7	0.28	49%

[1] Percentage of identical residues

[2] r.m.s. deviations of equivalent α-carbon atoms after superposition by algorithm of Rossmann and Argos(1975) with probability cutoff of 0.005. Number in brackets is the number of residues considered equivalent.

[3] Number of standard deviations above the mean. The rotation functions were calculated with data between 10 and 3.5Å resolution.

[4] Maximum correlation obtained using data between 5 and 4Å resolution

[5] For 2.8Å data.

[6] a.a. = amino acids

Bibliography

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

Boucher, R., Saidi, M., & Genest, J. (1972). In *Hypertension '72* (Genest, J. & Koiw, E., eds), pp. 512-523, Springer-Verlag, New York.

Bracewell, R. (1965) in *The Fourier Transform and Its Applications*, McGraw-Hill, New York.

Buerger, M. J. (1959) *Vector Space*, Wiley, New York.

Cooley, J. W. & Tukey, J. W. (1965) *Math. Comput. 19* 297-301.

Crowther, R. A. (1972) in *The Molecular Replacement Method* (Rossmann, M. G., ed.) International Science Review 13, pp173-178, Gordon & Breach, New York.

Crowther, R. A. & Blow, D. M. (1967) *Acta Cryst. 23* 544-548.

Fehlhammer, H. & Bode, W. (1975). *J. Mol. Biol. 98* 683-692.

Green, D. W., Ingram, V. M., & Perutz, M. F. (1954) *Proc. Roy. Soc. A225* 287-307.

Harada, Y., Lifchitz, A., Berthou, J., & Jolles, P. (1981) *Acta Cryst. A37* 398-406.

Hauptman, H. (1982) *Acta Cryst. A38* 289-294.

Hendrickson, W. A. & Ward, K. B. (1976) *Acta Cryst. A32* 778-780.

James, M. N. G. & Sielecki, A. R. (1986) *Nature (London) 319* 33-38.

Ladd, M. F. C. & Palmer, R. A. (1980) *Theory and Practice of Direct Methods in Crystallography*, Plenum Press, New York.

Langs, D. A. (1985) *Acta Cryst. A41* 578-582.

Leslie, A. G. W. (1985) in *Molecular Replacement* Proceedings of the Daresbury Study Weekend, 15-16 Feb., 1985 (Machin, P. A., ed.) pp78-81, Daresbury Laboratory, Daresbury, Warrington.

Lipson, H. & Cochran, W. (1957) *The Determination of Crystal Structures*, p235, Bell, London.

Main, P. (1979) *Acta Cryst. A35* 779-785.

Main, P., Fiske, S. J., Hull, S. E., Lessinger, L., Germain, G., Declercq, J.-P., & Woolfson, M. M. (1980) *MULTAN80* Univs. of York, England and Louvain, Belgium.

Patterson, A. L. (1935) *Z. Kristallogr. 90* 517-542.

Ramachandran, G. N. & Srinivasan, R. (1961) *Nature (London) 190* 159-161.

Read, R. J. (1986a) *Acta Cryst. A42* 140-149.

Read, R. J. (1986b) *Ph. D. Thesis*, University of Alberta.

Read, R. J., Brayer, G. D., Jurášek, L., & James, M. N. G. (1984) *Biochemistry 23* 6570-6575.

Rossmann, M. G. (1972) *The Molecular Replacement Method* International Science Review 13, Gordon & Breach, New York.

Rossmann, M. G. (1980) in *Theory and Practice of Direct Methods in Crystallography* (Ladd, M. F. C. & Palmer, R. A., eds.) pp361-417, Plenum Press, New York.

Rossmann, M. G. & Argos, P. (1975) *J. Biol. Chem. 250* 7525-7532.

Rossmann, M. G. & Blow, D. M. (1962) *Acta Cryst. 15* 24-31.

Srinivasan, R. & Parthasarathy, S. (1976) *Some Statistical Applications in X-Ray Crystallography*, Pergamon Press, Oxford.

Sussman, J. L., Holbrook, S. R., Church, G. M., & Kim, S.-H. (1977) *Acta Cryst. A33* 800-804.

Tollin, P. (1966) *Acta Cryst. 21* 613-614.

Tollin, P. (1969) *Acta Cryst. A25* 376-377.

Watson, G. N. (1958) *A Treatise on the Theory of Bessel Functions*, Cambridge University Press.

V. General Discussion

The determination and refinement of the 3-dimensional structure of proteins are essential in understanding their activity and their architecture. It is only after accurate coordinates have been determined, with high resolution data and extensive refinement, that inferences concerning the details of the reactivities of proteins can be made. In addition, the comparison of homologous structures show us the parts of the molecule which are important to their common function as well as indicating how differences in reactivities can arise.

The structures of two serine proteases, α-lytic protease and tonin, have been refined at high resolution. Their structures have been analysed and compared to structures of homologous enzymes. These structures provide accurate coordinates for interpretation of results using other methods such as NMR, and for structural and mechanistic studies using computer simulation techniques. The analysis of the interactions in the proteins reveal characteristics and patterns that help to generalize the features of protein structures. In both enzymes, the basic folding of the polypeptide chain is seen to be relatively constant in the core of the protein. Compensatory changes occur in the sequence to maintain the same folding. In the case of tonin, the structure that was determined is not in its native form due to a bound Zn^{2+} ion. However the existence of a highly homologous enzyme, kallikrein, allows for the prediction of the native structure of tonin.

Homologous protein structures not only provide structural information but also can be used to solve the structures of related molecules. This is done using the methods of molecular replacement as was done in structure solution of tonin. One of the difficult steps in the method is the positioning of the oriented molecule in the crystal cell, otherwise known as the translation problem. During the course of the determination of the tonin structure a new method of overcoming the translation problem was developed. It has been subsequently improved upon as a consequence of problems encountered in solving other structures. The utility of the method is demonstrated by the number of difficult structures that have been solved with it.