THE UNIVERSITY OF ALBERTA

X-ray Crystallographic Studies on Serine Proteinases and their Protein Inhibitors

by

Randy John Read

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

SPRING 1986

ISBN 0-315-30201-1

COPYRIGHT STATUS FORM

Lg #2

Name of American Chemical Society Publication

Author(s)

Ms No

Ms Title

Received

This manuscript will be considered with the understanding you have submitted it on an exclusive basis You will be notified of a decision as soon as possible

rint or
ype
uthor's
lame and
ddress

[THIS FORM MAY BE REPRODUCED]

COPYRIGHT TRANSFER

The undersigned, with the consent of all authors, hereby transfers, to the extent that there is copyright to be transferred, the exclusive copyright interest in the above cited manuscript (subsequently referred to as the "work") to the **American Chemical Society** subject to the following (Note: if the manuscript is not accepted by ACS or if it is withdrawn prior to acceptance by ACS, this transfer will be null and void and the form will be returned.):

A. The undersigned author and all coauthors retain the right to revise, adapt, prepare derivative works, present orally, or distribute the work provided that all such use is for the personal noncommercial benefit of the author(s) and is consistent with any prior contractual agreement between the undersigned and/or coauthors and their employer(s).

B. In all instances where the work is prepared as a "work made for hire" for an employer, the employer(s) of the author(s) retain(s) the right to revise, adapt, prepare derivative works, publish, reprint, reproduce, and distribute the work provided that all such use is for the promotion of its business enterprise and does not imply the endorsement of the American Chemical Society.

C. Whenever the American Chemical Society is approached by third parties for individual permission to use, reprint, or republish specified articles (except for classroom use, library reserve, or to reprint in a collective work) the undersigned author's or employer's permission will also be required.

D. No proprietary right other than copyright is claimed by the American Chemical Society

E. For works prepared under U.S. Government contract or by employees of a foreign government or its instrumentalities, the American Chemical Society recognizes that government's prior nonexclusive, royalty-free license to publish, translate, reproduce, use, or dispose of the published form of the work, or allow others to do so for noncommercial government purposes. State contract number: ____________

SIGN HERE FOR COPYRIGHT TRANSFER [Individual Author or Employer's Authorized Agent (work made for hire)]

Print Author's Name

Print Agent's Name and Title

Original Signature of Author on Behalf of All Authors (in Ink)

Date

Original Signature of Agent (in Ink)

CERTIFICATION AS A WORK OF THE U.S. GOVERNMENT

This is to certify that **ALL** authors are or were bona fide officers or employees of the U.S. Government at the time the paper was prepared, and that the work is a "work of the U.S. Government" (prepared by an officer or employee of the U.S. Government as a part of official duties), and, therefore, it is not subject to U.S. copyright. (This section should NOT be signed if the work was prepared under a government contract or coauthored by a non-U.S. Government employee.)

INDIVIDUAL AUTHOR OR AGENCY REPRESENTATIVE

Print Author's Name

Print Agency Representative's Name and Title

Original Signature of Author (in Ink)

Date

Original Signature of Agency Representative (in Ink)

FOREIGN COPYRIGHT RESERVED (NOTE: If your government permits copyright to be transferred, refer to section E and sign this form in the top section.)

□ **If ALL authors are employees of a foreign government that reserves its own copyright as mandated by national law, DO NOT SIGN THIS FORM. Please check this box as your request for the FOREIGN GOVERNMENT COPYRIGHT FORM (Blue Form) which you will be required to sign. If you check this box, mail this form to:** Copyright Administrator, Books and Journals Division, American Chemical Society, 1155 Sixteenth Street, N.W., Washington, D.C. 20036, U.S.A.

Co-author Permission Form

Permission is hereby granted to Randy J. Read to use material from

"Critical Evaluation of Comparative Model Building of *Streptomyces griseus* Trypsin"

by R. J. Read, G. D. Brayer, L. Jurášek and M. N. G. James,

published in *Biochemistry* 23: 6570-6575 (1984).

in his thesis, entitled

X-ray Crystallographic Studies on Serine Proteinases and their Protein Inhibitors.

Signed Gary D. Brayer

Name Gary D. Brayer

Date Sept 6/85

Co-author Permission Form

Permission is hereby granted to Randy J. Read to use material from

"Critical Evaluation of Comparative Model Building of *Streptomyces griseus* Trypsin"
by R. J. Read, G. D. Brayer, L. Jurášek and M. N. G. James,
published in *Biochemistry* 23: 6570-6575 (1984),

in his thesis, entitled
X-ray Crystallographic Studies on Serine Proteinases and their Protein Inhibitors.

Signed ....

Name Lubomir Jurášek

Date 6 Sep 85

Co-author Permission Form

Permission is hereby granted to Randy J. Read to use material from

"Critical Evaluation of Comparative Model Building of *Streptomyces griseus* Trypsin"

by R. J. Read, G. D. Brayer, L. Jurášek and M. N. G. James,

published in *Biochemistry* 23: 6570-6575 (1984),

in his thesis, entitled

X-ray Crystallographic Studies on Serine Proteinases and their Protein Inhibitors.

Signed Michael James

Name Michael N. G. James

Date September 3, 1985

Department of Biochemistry
University of Alberta
Edmonton, Alberta
CANADA
T6G 2H7
October 2, 1985

Dr. S. G. Richardson
Rights and Permissions
Elsevier Science Publishers
P. O. Box 1527
1000 BM Amsterdam
THE NETHERLANDS

Dear Dr. Richardson,

Thank you for your letter of September 19 concerning permission to use material from the chapter in *Proteinase Inhibitors* that I wrote with Professor James.

Having not yet written the relevant parts of my doctoral thesis, I find it difficult to be specific about which parts of the chapter I wish to use. The parts dealing with my own work would be used almost verbatim, while the review of other work would be extensively rewritten as background, introductory material. Since there is a time constraint, especially considering the turnaround time on correspondence with Europe, I will nonetheless be as specific as I can at this point.

The chapter contains 4 major sections, excluding references. (For your convenience, a copy of the title page and table of contents is enclosed.) I would like to use material from Sections 1 and 2, though not from sub-sections 2.1.4 and 2.5. There are 13 figures; of these, I may wish to use Figures 2 (a and b), 4, 5, 6 (b and c), 7, 8, 9 and 10 (a and b). Finally, there are 5 tables of which I may use or adapt Tables 2, 3, 4 and 5.

The title of my thesis, which you will probably need for your permission form, will be *X-ray Crystallographic Studies on Serine Proteinases and their Protein Inhibitors.*

I trust that granting permission for the use of this material will not be a problem for you. After all, my thesis will have quite a limited readership.

Thank you for your prompt attention to this matter.

Yours sincerely,

Randy Read

Randy Read

Elsevier Biomedical Press B.V.
Amsterdam, date:
Permission granted subject to
permission from the author(s)
and to full acknowledgement
of the source.

4/11/85

INTRODUCTION TO THE PROTEIN INHIBITORS: X-RAY CRYSTALLOGRAPHY

Randy J. Read and Michael N.G. James
Medical Research Council of Canada Group in Protein
Structure and Function
Department of Biochemistry, University of Alberta
Edmonton, Alberta, Canada T6G 2H7

Table of Contents

Co-author Permission Form

Permission is hereby granted to Randy J. Read to use material from

"Introduction to the Protein Inhibitors: X-ray Crystallography"

by R. J. Read and M. N. G. James,

to be published in *Proteinase Inhibitors*, eds. A. J. Barrett and G. S. Salvesen, Elsevier Science Publishers, Amsterdam,

in his thesis, entitled

X-ray Crystallographic Studies on Serine Proteinases and their Protein Inhibitors.

Signed ....Michael James........

Name ....Michael N. G. James........

Date ....September 3, 1985........

COPYRIGHT STATUS FORM

2/4 #3

Name of American Chemical Society Publication

Author(s)

Ms No

Ms Title

Received

This manuscript will be considered with the understanding you have submitted it on an exclusive basis You will be notified of a decision as soon as possible

rint or
ype
uthor's
ame and
ddress

[THIS FORM MAY BE REPRODUCED]

COPYRIGHT TRANSFER

The undersigned, with the consent of all authors, hereby transfers, to the extent that there is copyright to be transferred, the exclusive copyright interest in the above cited manuscript (subsequently referred to as the "work") to the **American Chemical Society** subject to the following (Note if the manuscript is not accepted by ACS or if it is withdrawn prior to acceptance by ACS, this transfer will be null and void and the form will be returned.):

A. The undersigned author and all coauthors retain the right to revise, adapt, prepare derivative works, present orally, or distribute the work provided that all such use is for the personal noncommercial benefit of the author(s) and is consistent with any prior contractual agreement between the undersigned and/or coauthors and their employer(s)

B In all instances where the work is prepared as a "work made for hire" for an employer, the employer(s) of the author(s) retain(s) the right to revise, adapt, prepare derivative works, publish, reprint, reproduce, and distribute the work provided that all such use is for the promotion of its business enterprise and does not imply the endorsement of the American Chemical Society.

C. Whenever the American Chemical Society is approached by third parties for individual permission to use, reprint, or republish specified articles (except for classroom use, library reserve, or to reprint in a collective work) the undersigned author's or employer's permission will also be required.

D No proprietary right other than copyright is claimed by the American Chemical Society

E. For works prepared under U.S. Government contract or by employees of a foreign government or its instrumentalities, the American Chemical Society recognizes that government's prior nonexclusive, royalty-free license to publish, translate, reproduce, use, or dispose of the published form of the work, or allow others to do so for noncommercial government purposes. State contract number ______________

SIGN HERE FOR COPYRIGHT TRANSFER [Individual Author or Employer's Authorized Agent (work made for hire)]

______________ Print Author's Name

______________ Print Agent's Name and Title*

______________ Original Signature of Author on Behalf of All Authors (in Ink)

______________ Date

______________ Original Signature of Agent (in Ink)

CERTIFICATION AS A WORK OF THE U.S. GOVERNMENT

This is to certify that **ALL** authors are or were bona fide officers or employees of the U.S. Government at the time the paper was prepared, and that the work is a "work of the U.S. Government" (prepared by an officer or employee of the U.S. Government as a part of official duties), and, therefore, it is not subject to U.S. copyright. (This section should NOT be signed if the work was prepared under a government contract or coauthored by a non-U S Government employee.)

INDIVIDUAL AUTHOR OR AGENCY REPRESENTATIVE

______________ Print Author's Name

______________ Print Agency Representative's Name and Title

______________ Original Signature of Author (in Ink)

______________ Date

______________ Original Signature of Agency Representative (in Ink)

FOREIGN COPYRIGHT RESERVED (NOTE. If your government permits copyright to be transferred, refer to section E and sign this form in the top section.)

☐ **If ALL authors are employees of a foreign government that reserves its own copyright as mandated by national law, DO NOT SIGN THIS FORM. Please check this box as your request for the FOREIGN GOVERNMENT COPYRIGHT FORM (Blue Form) which you will be required to sign. If you check this box, mail this form to:** Copyright Administrator, Books and Journals Division, American Chemical Society, 1155 Sixteenth Street, N.W. **Washington, D.C. 20036, U.S.A.**

Co-author Permission Form

Permission is hereby granted to Randy J. Read to use material from

"Structure of the Complex of *Streptomyces griseus* Protease B and the Third Domain of the Turkey Ovomucoid Inhibitor at 1.8-A Resolution"

by R. J. Read, M. Fujinaga, A. R. Sielecki and M. N. G. James,

published in *Biochemistry* 22: 4420-4433 (1983).

in his thesis, entitled

'X-ray Crystallographic Studies on Serine Proteinases and their Protein Inhibitors.

Signed ....

Name Masao Fujinaga ....

Date September 3, 1985 ....

Co-author Permission Form

Permission is hereby granted to Randy J. Read to use material from

"Structure of the Complex of *Streptomyces griseus* Protease B and the Third Domain of the Turkey Ovomucoid Inhibitor at 1.8-A Resolution"

by R. J. Read, M. Fujinaga, A. R. Sielecki and M. N. G. James,

published in *Biochemistry* 22: 4420-4433 (1983),

in his thesis, entitled

X-ray Crystallographic Studies on Serine Proteinases and their Protein Inhibitors.

Signed ....

Name Anita R. Sielecki

Date September 3, 1985

Co-author Permission Form

Permission is hereby granted to Randy J. Read to use material from

"Structure of the Complex of *Streptomyces griseus* Protease B and the Third Domain of the Turkey Ovomucoid Inhibitor at 1.8-A Resolution"

by R. J. Read, M. Fujinaga, A. R. Sielecki and M. N. G. James,

published in *Biochemistry* 22: 4420-4433 (1983).

in his thesis, entitled

X-ray Crystallographic Studies on Serine Proteinases and their Protein Inhibitors.

Signed Michael James

Name Michael N. G. James

Date September 3, 1985

International Union of Crystallography

Transfer of Copyright Agreement

Title of Article (*Please type or use capital letters*)

IMPROVED FOURIER COEFFICIENTS FOR MAPS USING PHASES FROM PARTIAL STRUCTURES

WITH ERRORS

Authors (*Please type or use capital letters*)

RANDY J. READ

Copyright to the above article is hereby transferred (except as otherwise required by national laws*) to the International Union of Crystallography, effective if and when the article is accepted for publication in *Acta Crystallographica* or *Journal of Applied Crystallography*. However, the authors reserve the following:

(1) All proprietary rights other than copyright, such as patent rights.
(2) The right to use all or part of this article in future works prepared by or on behalf of the author.
(3) In the case of a 'work for hire', the right of the employer to incorporate all or part of this article in future works.
(4) The royalty-free right of the employer to make copies of this article, without further permission, for his own use but not for resale. Any identification or notice appearing on pages to be reproduced must not be deleted.

Signature Randy Read

Name and position, if not author

Date May 29, 1985

Signature

Name and position, if not author

Date

This Transfer of Copyright must be signed by at least one of the authors (who agrees to inform the others, if any) or, in the case of a 'work made for hire', by the employer.

The signed statement must be received by the Editor or Co-editor to whom the article is submitted before the article can be accepted for publication. Requests for further information should be sent to the Executive Secretary of the Union.

* In this case please give further details overleaf.

For use of the International Union of Crystallography only.

T. E. Ref.	Co-editor Ref.	Issue	Journal		

This form may be photocopied.

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR Randy John Read

TITLE OF THESIS X-ray Crystallographic Studies on Serine Proteinases and their Protein Inhibitors

DEGREE FOR WHICH THESIS WAS PRESENTED Doctor of Philosophy

YEAR THIS DEGREE GRANTED SPRING 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) Randy Read

PERMANENT ADDRESS:

5531 Dalcastle Dr. NW

Calgary, Alberta

DATED December 19 1985

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 on Serine Proteinases and their Protein Inhibitors submitted by Randy John Read in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Michael James
....
Supervisor

Brian D Sykes
....

Wayne F. Anderson
....

....

....
External Examiner

Date: December 19, 1985

To Jack and Anne Read, my parents

Abstract

Three crystal structures of serine proteinases, two complexed to a protein proteinase inhibitor, have been solved and refined at high resolution. In addition, methods have been developed to make better use of the phase information given by a partial structural model.

Streptomyces griseus trypsin (SGT) is a bacterial serine proteinase that is more homologous to mammalian than to other bacterial enzymes. The structure of SGT has been solved by molecular replacement, using the mammalian pancreatic serine proteinases bovine trypsin and α-chymotrypsin as models. Because these proteins have low homology to SGT compared to most successful replacement models, new strategies were required for molecular replacement to succeed. The model of SGT has been refined at 1.7Å resolution to a final *R*-factor of 0.161; the correlation coefficient between observed and calculated structure factors is 0.908. The structure of SGT had been predicted in two models on the basis of its expected homology to the pancreatic serine proteinases. An evaluation of these models demonstrates the effect of several sources of error on such comparative model-building. The objective of comparative model-building is often to explain substrate specificity, or to suggest potential highly specific drugs. The unique parts of modelled proteins that are most important for such purposes are, however, the most poorly determined by the model-building procedure.

The structures of two complexes of the third domain of the ovomucoid inhibitor from turkey (OMTKY3), with *Streptomyces griseus* Protease B (SGPB) and with α-chymotrypsin, have been solved by molecular replacement, using the native proteinase structures as models. Both structures have been refined at 1.8Å resolution to final R-factors of 0.145 for the SGPB complex and 0.168 for the α-chymotrypsin complex. The conformation of OMTKY3 differs in the two complexes, indicating that conformational flexibility is required for its broad specificity. The mechanism of inhibition can be understood in terms of the thermodynamic parameters K_{assoc}, the inhibitor:enzyme association constant, and K_{hyd}, the equilibrium constant for inhibitor hydrolysis. These parameters can be rationalized in terms of the observed structures of these and other complexes.

Unrefined or partially refined models of macromolecules are generally incomplete and typically have large coordinate errors. In such cases, phase probability equations appropriate for a perfect partial structure lead to inaccurate estimates of phase probabilities. Therefore, one must use equations that have been derived allowing for errors in the partial structure. A method is given to estimate the parameter σ_A in these phase probability expressions from the observed and calculated structure factor amplitudes. From the variation of σ_A with resolution, one can estimate the rms coordinate error for the model. Electron density maps calculated using partial structure phases are biased towards

the partial structure. When there are coordinate errors, a new expression for the non-centric Fourier coefficients $[(2m|\mathbf{F}_N|-D|\mathbf{F}_P^C|)\exp(i\alpha_P^C)]$ is required to suppress this model bias. Judged by correlation coefficients comparing electron density maps with the correct and the partial structure maps, these coefficients are superior to others currently in use. With a few assumptions, related Fourier coefficients are developed to reduce bias in combined phase electron density maps.

9

Acknowledgement

I would like to thank Michael James for his careful guidance and for his friendship during my tenure as a graduate student. Looking back at my years under his supervision, I can appreciate the judgement required to teach and supervise, while encouraging the student to develop as an independent investigator.

I value the friendship and cooperative spirit that exist in this laboratory, and I have enjoyed working with everyone here. Anita Sielecki, in particular, has helped to create the warm atmosphere in the lab. In addition, Anita has been my guide to the careful refinement of protein structures. Masao Fujinaga has been my frequent collaborator; because of the critical eye he brings to his work, these collaborations have always been rewarding. He also introduced me to serious jazz and to long-distance hiking. Much of the work described in Chapter III, I should note here, was done by Masao, Anita and Mike James in a collaborative project. I have found John Moult very stimulating to work with, as a source of many good ideas, as a sounding-board for my ideas, and as a dependable running partner. Osnat Herzberg and Cathy McPhalen have been untiring in their crusade to find bugs in programs declared bug-free. I offer my grudging thanks. Cathy, in addition, has set a record pace in supplying new crystal structures for lab celebrations. On a more serious note, Osnat assisted me in learning about heavy-atom refinement. Koto Hayakawa's quiet

competence in technical support is gratefully acknowledged. She grew all of the crystals used in this work, except for the large SGT crystal used for high resolution data collection – and in that case she told me exactly what to do. Colin Broughton was a never-ending source of novel ideas, and taught me much of what I know about computer programming. Finally, I inherited the SGT project from Gary Brayer, who left me several sets of data and the initial characterization of the crystal.

Our association with Wayne Anderson's laboratory has also been enjoyable. I would like to thank in particular David Bacon, for assistance in the use of his excellent computer graphics programs, and Alastair Muir, for a careful proofreading of the first two chapters of this thesis.

A number of friends and colleagues have been important to me over the years. I will not risk offending some by attempting to list them; I hope they know who they are. I will, however, make an exception for the members of my family, who have been very supportive of my chosen career.

Finally, I wish to acknowledge the friendship and encouragement of Carol Woo. Her patience and companionship over the last several years, especially these last hectic months, have helped me to keep my perspective and have made my task much easier.

This research has been supported by the Medical Research Council of Canada, through a grant to the MRC Group in Protein Structure and Function, and by the Alberta

Heritage Foundation for Medical Research, through a Research Allowance to me. I have been supported as a graduate student by Studentships from the MRC, and by a Dissertation Fellowship from the University of Alberta.

Table of Contents

List of Tables

Table Page

List of Figures

Figure Page

Figure Page

List of Symbols and Abbreviations

a, b, c	unit cell axes
B	thermal motion parameter
BT	bovine trypsin
BTan	anhydrotrypsin
BTn	trypsinogen
CHT	α-chymotrypsin
CI-2	barley chymotrypsin inhibitor 2
d_{min}	minimum interplanar spacing of diffraction data
e	electron
$\|\mathbf{E}\|$	normalized structure factor amplitude
$\|\mathbf{E}_c\|$	normalized calculated structure factor amplitude
$\|\mathbf{E}_o\|$	normalized observed structure factor amplitude
E_H	lack-of-closure error
f	atomic scattering factor
$\mathbf{F}$	structure factor (vector quantity)
$\mathbf{F}_c$	calculated structure factor
$\mathbf{F}_H$	heavy atom structure factor
$\mathbf{F}_P$	protein structure factor (except Chapter IV)
$\mathbf{F}_{PH}$	protein plus heavy atom structure factor
$\mathbf{F}_{PH^+}$, $\mathbf{F}_{PH^-}$	Friedel pair of derivative data
$\|\mathbf{F}\|$	structure factor amplitude
$\|\mathbf{F}_o\|$	observed structure factor amplitude
GTM2	SGT mersalyl derivative data set
GTN2	SGT low resolution native data set

GTN3	SGT high resolution native data set
GTP1	SGT platinum derivative data set
GTUA	first SGT uranyl acetate derivative data set
GTU2	second SGT uranyl acetate derivative data set
I	intensity
IV	isoleucylvaline
JG-SGT	J. Greer's comparative model of SGT
k_{cat}	rate constant for catalysis
K_M	Michaelis constant
LJ-SGT	L. Jurášek's comparative model of SGT
m	figure of merit
m_c	m for calculated (model) phase
m_{comb}	m for combined phase
MIR	multiple isomorphous replacement
NMR	nuclear magnetic resonance
OMJPQ3	domain 3 of Japanese quail ovomucoid inhibitor
OMSVP3	domain 3 of silver pheasant ovomucoid inhibitor
OMTKY3	domain 3 of turkey ovomucoid inhibitor
PSTI	pancreatic secretory trypsin inhibitor
PT	porcine trypsin
PTI	pancreatic trypsin inhibitor
R	standard crystallographic residual
rms	root-mean-square
SGPA	*Streptomyces griseus* Protease A
SGPB	*Streptomyces griseus* Protease B

SGT	*Streptomyces griseus* trypsin
SSI	*Streptomyces* subtilisin inhibitor
STI	soybean trypsin inhibitor
α_c	calculated structure factor phase
α_{comb}	combined phase
$\sigma(x)$	standard deviation in x
σ_A	phase probability parameter
Σ_N	measure of scattering matter in crystal
Σ_P	measure of scattering matter in model
Σ_Q	measure of missing scattering matter

Note: terms and notation used in Chapter IV are defined in Table IV.1.

I. Introduction

A direct image of molecular structure at the atomic level is provided by the technique of X-ray crystallography. Such information is of great value in many fields. For instance, the precise disposition of functional groups in an enzyme creates an environment in which a particular chemical reaction is accelerated. The crystallographic study of enzymes and their complexes with substrates and inhibitors is a prerequisite to the detailed understanding of enzyme action. Since many drugs are enzyme inhibitors, such structural knowledge can also provide the basis for the rational design of drugs. As our knowledge of the relationships among amino acid sequence, three-dimensional structure and protein function increases, even the design of proteins with new functions may become possible.

It is beyond the scope of this discussion to give a background description of protein crystallography that would be informative to someone who is not already familiar with the concepts. For the non-crystallographer, the introduction to protein crystallography in Chapter 6 of "Proteins: Structures and Molecular Properties" by Creighton (1983) is highly recommended, as it introduces the concepts and terminology in a fairly readable fashion. The text by Blundell and Johnson (1976) is a standard reference work; it describes in detail the theory and practice of protein crystallography.

This dissertation contains several studies involving the use of X-ray crystallography in the study of protein structure. The proteins I have studied are serine proteinases from the family typified by α-chymotrypsin (CHT); two of these were examined in their complexes with a protein proteinase inhibitor. In the course of this structural work, I developed some strategies and techniques for use in the molecular replacement method of crystal structure solution. In addition, I made improvements to the way in which the structure factor phases computed from a structural model are used in the calculation of electron density maps. These methods will be of general applicability in macromolecular crystallography.

A. *Streptomyces griseus* Trypsin

The major structural study described herein is that of *Streptomyces griseus* trypsin (SGT). SGT is purified from Pronase, a commercial product obtained from the extracellular culture filtrate of the K1 strain of *Streptomyces griseus*, a soil bacterium. Pronase is a complex mixture of proteinases and other products, but originally was believed to be a homogeneous proteinase of remarkably broad specificity (Nomoto and Narahashi, 1959). It was first fractionated by Hiramatsu and Ouchi (1963) into three components having proteolytic activity, and is now recognized to contain proteinases belonging to several families [see, for example, Narahashi (1972)]. The chymotrypsin serine proteinase family is represented by SGT, *Streptomyces griseus* protease A (SGPA), and *Streptomyces griseus* protease B (SGPB). They were assigned to this family by Wahlby and Engstrom (1968) on the basis of the characteristic Asp-Ser-Gly sequence at the catalytic serine that reacts with diisopropylfluorophosphate.

One of these proteinases was found to hydrolyze benzoylarginine ethyl ester and hence was termed BAEE-hydrolase (Wahlby and Engstrom, 1968). This suggested a primary specificity for lysine and arginine residues in the P_1 position of the substrate'[1], as in the case of bovine trypsin (BT). Strong amino acid sequence similarity to BT was found by Jurášek *et al.* (1969) in the sequences about the disulfide bridges. The determination of the complete amino acid sequence of the enzyme, by this time renamed *Streptomyces griseus* trypsin, confirmed the pronounced sequence similarity (Olafson *et al.*, 1975). Research has also demonstrated striking parallels in the properties of SGT and BT, including their substrate specificity and susceptibility to inhibitors (Trop and Birk, 1968, 1970; Narahashi and Fukunaga, 1969; Awad *et al.*, 1972; Awad and Ochoa, 1974; Olafson and Smillie, 1975; Yokosawa *et al.*, 1976; Mosolov *et al.*, 1978; Nishikata *et al.*, 1981; Tashiro *et al.*, 1981; Shimura and

[1] The notation of Schechter and Berger (1967) is used to facilitate discussion of the interactions between a proteinase and bound peptides. Amino acid residues of substrates are numbered P_1, P_2, etc., towards the amino-terminal direction and P_1', P_2', etc. in the carboxy-terminal direction from the scissile bond. The complementary subsites of the enzyme binding region are numbered S_1, S_2 and S_1', S_2', etc.

Kasai, 1982). Although similar, these enzymes are not identical in their properties. For example, the ovomucoid inhibitor from Japanese quail inhibits BT, but not SGT (Nagata and Yoshida, 1983).

SGT belongs to a family of serine proteinases that is extremely well-studied. [For discussions integrating chemical and structural information see the reviews by Blow (1976), Kraut (1977) and Huber and Bode (1978), and the mechanistic proposal of James *et al.* (1980).] Numerous crystal structures involving serine proteinases have been determined, and several have been refined at high resolution. In the July 1985 release of the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977), there are 30 coordinate holdings involving 8 different proteinases from this family. Eighteen of these holdings involve BT in various conditions, complexes and crystal forms. One might ask what is of interest in the structure of another serine proteinase, particularly one that bears such a strong resemblance to the well-studied BT. The study of a new member of a family will provide new insight into the organizing principles, but there is bound to be a diminishing return with each new member. Accordingly, there will be relatively little discussion in this dissertation on the implications of the structure of SGT on a mechanism for the serine proteinases. Nonetheless, there are aspects of this protein that make it of considerable interest, not just as a serine proteinase.

Since its discovery, SGT has been considered an evolutionary anomaly. Though it is produced by the same bacterium that makes SGPA and SGPB, SGT is much more homologous to the mammalian pancreatic enzymes (Jurášek *et al.*, 1976; James *et al.*, 1978). In fact, measured by amino acid sequence identity, BT and CHT are as closely related to SGPA and SGPB as is SGT. This surprising observation, among others, prompted Hartley (1970, 1979) to propose, somewhat tongue-in-cheek, that the bacterium was infected by a cow. Hewett-Emmett *et al.* (1981) have disputed Hartley's (1979) concern with the similarity between SGT and BT by providing evidence in favour of an early evolutionary origin for SGT. A genealogical tree constructed by comparing serine proteinase sequences, not including SGPA or SGPB, has SGT at the base (Hewett-Emmett *et al.*, 1981). If we reject Hartley's (1970, 1979) proposal, we are left with the implication that for some reason a tryptic specificity imposes stricter structural requirements than the chymotryptic specificity shared by CHT, SGPA and SGPB. As will be discussed, the nature of one such structural requirement has been deduced from a comparison of SGT and BT.

Another evolutionary question concerns the N-terminus of SGT. Structural studies of serine proteinase zymogens, chymotrypsinogen (Freer *et al.*, 1970) and trypsinogen (Fehlhammer *et al.*, 1977), have demonstrated that the new N-terminus, generated by the activating cleavage, becomes buried in an ion pair near the active site. The burial of the N-terminus is a key element of the zymogen activation mechanism. Despite sequence homologies between the N-termini of SGPA, SGPB and α-lytic protease, and those of the active mammalian serine proteinases, these residues of the bacterial enzymes are exposed on their surfaces (Delbaere *et al.*, 1975; Brayer *et al.*, 1978; Brayer *et al.*, 1979; Fujinaga *et al.*, 1985). On the other hand, sequence alignments (Olafson *et al.*, 1975), chemical modification experiments (Awad and Ochoa, 1974; Olafson and Smillie, 1975) and model-building studies (Jurášek *et al.*, 1976) implied that the N-terminus of SGT was buried, though Duggleby and Kaplan (1975) obtained chemical modification results contradicting this conclusion. In the crystal structure of SGT, the buried ion pair is indeed observed. Either *Streptomyces griseus* stands poised to develop a zymogen activation mechanism, or one has already evolved in this primitive organism. There are, to date, no biochemical data on this question.

SGT has been the target of two comparative model-building attempts (Jurášek *et al.*, 1976; Greer, 1981a). In comparative modelling, one attempts to construct a model of an unknown protein structure based on sequence alignments and expected homology to a protein of known three-dimensional structure. Since many proteins of biological or medical interest belong to protein families in which some members have known crystal structures, this technique is achieving some prominence. One example is renin, an aspartyl proteinase that is a potential target of drugs to control hypertension.

A model of this protein has been constructed by Blundell *et al.* (1983) based on the crystal structure of the homologous *Endothia* pepsin (Subramanian *et al.*, 1977). Another example is the blood clotting factor X_a, which Greer (1981b) modelled on the basis of the known crystal structures of other serine proteinases. It is essential to have some idea of the reliability of such comparative models. One could assess comparative modelling techniques using the known structures of two related proteins, but it would be difficult to simulate the state of ignorance that exists before a crystal structure is determined. The solution of the structure of SGT made it possible to evaluate critically the models that had been built before the answer was known.

The structure of SGT was solved primarily by the technique of molecular replacement (Rossmann, 1973), using BT and CHT as models. This technique exploits the prior structural information given by known crystal structures of the same or related proteins. Patterns in the Patterson function[2] of the observed diffraction data are matched with the comparable patterns in the model Patterson function. In this way, the orientation and position needed to place the model in the unit cell of the unknown structure are obtained. The resulting model of the crystal structure can provide structure factor phases for the calculation of electron density maps. To achieve a structure solution with

[2] The Patterson function (Patterson, 1934) is the Fourier transform of the intensities ($|F|^2$) and contains peaks at positions corresponding to the vectors between atoms.

molecular replacement, the signal indicating the correct orientation and position of the model must be distinguished from the noise. The noise level is, to a certain extent, out of the hands of the crystallographer. For example, the complexity of the Patterson function in high-symmetry space groups adds noise (Reynolds *et al*., 1985). The proper choice of some variable parameters, however, can reduce the noise level (Lifchitz, 1983). If the model is particularly good, the signal will be very strong and any choice of parameters will be adequate. This is true of many structure determinations that use molecular replacement. One example is the structure of SGPB in its complex with the third domain of the ovomucoid inhibitor from turkey (OMTKY3); with the refined structure of native SGPB as the model of SGPB in the complex, the molecular replacement problem was straightforward (Fujinaga *et al*., 1982; Read *et al*., 1983). Though BT and CHT are similar to SGT, the level of sequence identity is low compared to that found for most successful molecular replacement models. Therefore, the choice of parameters was critical to success, and the structure solution was not achieved easily. The experience gained with this structure should be of use in other difficult molecular replacement problems.

B. Turkey Ovomucoid Inhibitor Third Domain

Protein inhibitors of proteolytic enzymes are indeed proteins, so they should be substrates for proteolysis, not inhibitors. The elucidation of this paradox remains a central focus for much of the work on the structure and function of protein inhibitors of proteinases. It has been established biochemically that many such inhibitors bind productively as substrates and are cleaved at a peptide bond termed the reactive site; they bind very tightly but are cleaved very slowly (Laskowski and Kato, 1980). The role of protein crystallography has been to determine whether there is some unique feature in the structures of protein inhibitors and their complexes with proteolytic enzymes that explains why the inhibitors are not ordinary substrates.

Crystal structures were determined for two complexes of the protein proteinase inhibitor OMTKY3. This was a collaborative project among Masao Fujinaga, Anita Sielecki, Michael James, Wojciech Ardelt, Michael Laskowski, Jr. and me. Because of the close collaboration, it is difficult to assign credit to individuals for particular aspects of the work. It would be inappropriate to include here more than background information on those parts in which I played a minor role. Aspects in which I was more heavily involved are discussed in more detail.

At the time the OMKTY3 project was initiated, the only high resolution refined structures of protein inhibitors of serine proteinases involved bovine pancreatic trypsin inhibitor (PTI), which was extensively studied by R. Huber and co-workers. The available refined structures were PTI in its native form (Deisenhofer and Steigemann, 1975), in complexes with BT (Huber and Bode, 1978), anhydrotrypsin (Huber *et al.*, 1975) and trypsinogen (Huber and Bode, 1978), and in the ternary complex with trypsinogen and isoleucylvaline (Bode *et al.*, 1978). A controversial conclusion of that work was that the peptide bond at the reactive site is strongly distorted toward a tetrahedral configuration (Huber *et al.*, 1974, 1975). It was of interest to determine whether a similar distortion would be observed in complexes of the ovomucoid inhibitors. In the interim, the structures involving PTI have been further refined; the distortion is still observed, but is smaller than in earlier models (Marquart *et al.*, 1983).

The ovomucoid inhibitors are being used by M. Laskowski, Jr. in attempts to develop a sequence-reactivity algorithm (Laskowski, 1980). In this approach to the study of protein families, a large data base of amino acid sequences, reactivities and crystal structures must first be compiled. The study of this data base should reveal correlations between amino acid sequence and reactivity, in addition to structural explanations for these correlations. It is proposed that, with a large enough data base, it should be possible to predict the reactivities of new members of the family, based only on their sequences. Laskowski chose the third domains of ovomucoid inhibitors[3] because they are small and easily purified, they can be obtained in great variety from different species of birds, and they have a simple, well-defined reactivity, i.e. binding as competitive inhibitors to different serine proteinases (Laskowski, 1980; Laskowski *et al.*, 1983).

The sequence-reactivity algorithm can be greatly simplified if one can assume that the structures of the proteins involved are rigid. However the determination of two structures of OMTKY3 complexes, with SGPB and with CHT, has shown that the structure of OMTKY3 is not rigid. In fact, its flexibility around the reactive site loop appears to be quite important to extending the range of proteinases with which it can interact.

In addition to providing the structure of OMTKY3, the solution of the crystal structure of the complex with SGPB provided an opportunity to assess the differences in structure of a single protein (SGPB) in two different environments. The environment of SGPB in the native and complex crystals differs in two respects: crystal packing contacts in the two crystal forms, and the presence or absence of the binding interactions with the inhibitor.

[3] The ovomucoid inhibitors contain three domains, each of which belongs to the Kazal family of inhibitors on the basis of homology to pancreatic secretory trypsin inhibitor (Laskowski and Kato, 1980).

C. Electron Density Map Coefficients

The proper use of phase information from any source requires an accurate knowledge of the probability distribution associated with that information. Blow and Crick (1959) showed that, to minimize the root-mean-square (rms) error in the electron density map, the Fourier coefficients should be weighted by the expected value of the cosine of the phase error, a quantity they called the figure of merit (m). Though it was derived specifically for the case of phases determined by multiple isomorphous replacement (MIR), this result is general and applies equally to phases calculated from a model of the crystal structure. Therefore, accurate probability distributions are needed to provide optimum weights for maps computed using model phases. Phase information from different sources can be combined by multiplying the phase probability densities (Rossmann and Blow, 1961) or, equivalently, by adding the Hendrickson-Lattman coefficients encoding the phase information (Hendrickson and Lattman, 1970); if the phase probabilities are inaccurate, the different sources of information will not have the appropriate relative influence on the final phase of the structure factor.

Errors in calculated phases arise in part from incompleteness of the structural model. Woolfson (1956) derived partial structure phase probabilities for centric structure factors, and Sim (1959, 1960) derived probabilities for the non-centric case. Phase errors also arise from errors in the structural model. Srinivasan and co-workers extended the expressions for partial structure phase probabilities to include the effect of coordinate errors (Srinivasan and Ramachandran, 1965; Srinivasan, 1966). Nonetheless it has been common in practice to assume that the effect of coordinate errors can be ignored, or at most to add an error contribution to the estimate of the amount of missing structure (Rossmann and Blow, 1961). The work with SGT showed, however, that phase accuracy can be drastically overestimated when the effect of coordinate errors is ignored. This situation may be remedied by using the expressions of Srinivasan, but to apply these an estimate of the parameter σ_A (Srinivasan and Ramachandran, 1965; Srinivasan, 1966) is needed. With a modification of a method given by Lunin and Urzhumtsev (1984), reliable estimates of σ_A and hence of the phase probabilities for calculated phases are obtained.

It has long been known that electron density maps using model phases are biased towards the model (Luzzati, 1953). To compensate for this bias, the map coefficients $(2|\mathbf{F}_O|-|\mathbf{F}_C|)\exp(i\alpha_C)$ are commonly used. Main (1979) considered the effect of missing structure on maps computed with figure of merit weights, where the figure of merit is calculated using the results of Woolfson (1956) and Sim (1959, 1960). He concluded that the coefficients $m_C|\mathbf{F}_O|\exp(i\alpha_C)$ for centric and $(2m_C|\mathbf{F}_O|-|\mathbf{F}_C|)\exp(i\alpha_C)$ for non-centric structure factors would compensate for model bias. However, a re-examination of this work shows that Main's (1979) map coefficients must be modified when there are coordinate errors.

It has also been recognized that combined phases can lead to electron density maps with model bias, but the solution to this problem has been less clear. For the first SGT electron density map, combined phases and figure of merit weights were used in the Fourier coefficients $m_{comb}|F_O|\exp(i\alpha_{comb})$; the result was a model that looked like BT in places where SGT does not. Rice (1981) encountered similar problems in combined phase maps of phosphoglycerate kinase and achieved some improvement with the map coefficients $(2|F_O|-|F_C|)\exp(i\alpha_{comb})$. However, this does not satisfy the criterion that, in the limiting case of a single source of phase information, the combined phase coefficients should simplify to the appropriate coefficients for a single phase source. In other words, if the only phase information comes from MIR, the combined phase coefficient should be that of Blow and Crick (1959), and if the only phase information is the calculated phase, the map coefficient should be appropriate to that case [for example, the modification of Main's (1979) coefficients derived here]. One might also expect that the best results will be obtained when the coefficient for each structure factor depends on the extent to which the different sources of phase information influence the combined phase for that structure factor. Such aspects have been considered by Stuart and Artymiuk (1984). The combined phase map coefficients developed in the course of the structural work on SGT are similar_in concept to those of Stuart and Artymiuk (1984), but differ in several important details.

Bibliography

Awad, W. M. Jr., & Ochoa, M. S. (1974) *Biochem. Biophys. Res. Commun.* *59*, 527-534.

Awad, W. M. Jr., Soto, A. R., Siegel, S., Skiba, W. E., Bernstrom, G. G., & Ochoa, M. S. (1972) *J. Biol. Chem.* *247*, 4144-4154.

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

Blow, D. M. (1976) *Acc. Chem. Res.* *9*, 145-152.

Blow, D. M., & Crick, F. H. C. (1959) *Acta Cryst.* *12*, 794-802.

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

Blundell, T., Sibanda, B. L., & Pearl, L. (1983) *Nature (London)* *304*, 273-275.

Bode, W., Schwager, P., & Huber, R. (1978) *J. Mol. Biol.* *118*, 99-112.

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.

Creighton, T. E. (1983) *Proteins: Structures and Molecular Properties*, W. H. Freeman, New York.

Deisenhofer, J., & Steigemann, W. (1975) *Acta Cryst.* *B31*, 238-250.

Delbaere, L. T. J., Hutcheon, W. L. B., James, M. N. G., & Thiessen, W. E. (1975) *Nature (London)* *257*, 758-763.

Duggleby, R. G., & Kaplan, H. (1975) *Biochemistry* *14*, 5168-5175.

Fehlhammer, H., Bode, W., & Huber, R. (1977) *J. Mol. Biol.* *111*, 415-438.

Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. T., & Xuong, Ng. H. (1970) *Biochemistry* *9*, 1997-2009.

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.

Fujinaga, M., Delbaere, L. T. J., Brayer, G. D., & James, M. N. G. (1985) *J. Mol. Biol. 183*, 479-502.

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

Greer, J. (1981b) *J. Mol. Biol. 153*, 1043-1053.

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

Hartley, B. S. (1979) *Proc. Roy. Soc. Ser. B 205*, 443-452.

Hendrickson, W. A., & Lattman, E. E. (1970) *Acta Cryst. B26*, 136-143.

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

Hiramatsu, A., & Ouchi, T. (1963) *J. Biochem. 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.

Huber, R., Bode, W., Kukla, D., Kohl, U., & Ryan, C. A. (1975) *Biophys. Struct. Mechanism 1*, 189-201.

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

Jurášek, L., Fackre, D., & Smillie, L. B. (1969) *Biochem. Biophys. Res. Commun. 37*, 99-105.

Jurášek, L., Olafson, R. W., Johnson, P., & Smillie, L. B. (1976) *Miami Winter Symp. 11*, 93-123.

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

Laskowski, M. Jr. (1980) *Biochem. Pharm. 29*, 2089-2094.

Laskowski, M. Jr., & Kato, I. (1980) *Ann. Rev. Biochem. 49*, 593-626.

Laskowski, M. Jr., Tashiro, M., Empie, M. W., Park, S. J., Kato, I., Ardelt, W., & Wieczorek, M. (1983) in

Proteinase Inhibitors: Medical & Biological Aspects (Katunuma, N., Umezawa, H., & Holzer, H., Eds.) pp 55-68, Japan Scientific Societies Press, Tokyo/Springer-Verlag, Berlin.

Lifchitz, A. (1983) *Acta Cryst. A39,* 130-139.

Lunin, V. Y., & Urzhumtsev, A. G. (1984) *Acta Cryst. A40,* 269-277.

Luzzati, V. (1953) *Acta Cryst. 6,* 142-152.

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

Marquart, M., Walter, J., Deisenhofer, J., Bode, W., & Huber, R. (1983) *Acta Cryst. B39,* 480-490.

Mosolov, V. V., Fedurkina, N. V., & Valueva, T. A. (1978) *Biochim. Biophys. Acta 522,* 187-194.

Nagata, K., & Yoshida, N. (1983) *J. Biochem. 93,* 909-919.

Narahashi, Y. (1972) *Meth. Enzymol. 19,*651-664.

Narahashi, Y., & Fukunaga, J. (1969) *J. Biochem. 66,* 743-745.

Nishikata, M., Kasai, K.-I., & Ishii, S.-I. (1981) *Biochim. Biophys. Acta 660,* 256-261.

Nomoto, M., & Narahashi, Y. (1959) *J. Biochem. 46,* 1481-1487.

Olafson, R. W., & Smillie, L. B. (1975) *Biochemistry 14,* 1161-1167.

Olafson, R. W., Jurášek, L., Carpenter, M. R., & Smillie, L. B. (1975) *Biochemistry 14,* 1168-1177.

Patterson, A. L. (1934) *Phys. Rev. 46,* 372-376.

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

Reynolds, R. A., Remington, S. J., Weaver, L. H., Fisher, R. G., Anderson, W. F., Ammon, H. L., & Matthews, B. W. (1985) *Acta Cryst. B41,* 139-147.

Rice, D. W. (1981) *Acta Cryst. A37,* 491-500.

Rossmann, M. G. (1973) *The Molecular Replacement Method* International Science Review 13, Gordon & Breach, New York.

Rossmann, M. G., & Blow, D. M. (1961) *Acta Cryst. 14*, 641-647.

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

Shimura, K., & Kasai, K.-I. (1982) *J. Biochem. 92*, 1615-1622.

Sim, G. A. (1959) *Acta Cryst. 12*, 813-815.

Sim, G. A. (1960) *Acta Cryst. 13*, 511-512.

Srinivasan, R. (1966) *Acta Cryst. 20*, 143-144.

Srinivasan, R., & Ramachandran, G. N. (1965) *Acta Cryst. 19*, 1008-1014.

Stuart, D., & Artymiuk, P. (1984) *Acta Cryst. A40*, 713-716.

Subramanian, E., Swan, I. D. A., Liu, M., Davies, D. R., Jenkins, J. A., Tickle, I. J., & Blundell, T. L. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 556-559.

Tashiro, M., Sugihara, N., Maki, Z., & Kanamori, M. (1981) *Agric. Biol. Chem. 45*, 519-521.

Trop, M., & Birk, Y. (1968) *Biochem. J. 109*, 475-476.

Trop, M., & Birk, Y. (1970) *Biochem. J. 116*, 19-25.

Wählby, S., & Engström, L. (1968) *Biochim. Biophys. Acta 151*, 402-408.

Woolfson, M. M. (1956) *Acta Cryst. 9*, 804-810.

Yokosawa, H., Hanba, T., & Ishii, S.-I. (1976) *J. Biochem. 79*, 757-763.

II. *Streptomyces griseus* Trypsin[1]

Streptomyces griseus trypsin (SGT) is a serine proteinase obtained from the commercial product Pronase, an extracellular filtrate of cultures of the K1 strain of the soil bacterium *Streptomyces griseus*. Numerous similarities between SGT and bovine trypsin (BT) [summarized by Olafson and Smillie (1975)] have been attributed to their strong homology in amino acid sequence and, presumably, three-dimensional structure (Olafson *et al.*, 1975; Jurášek *et al.*, 1976). The sequence homology with BT is demonstrated in Table II.1 by a sequence alignment based on alignments of the crystal structure of SGT with those of BT (Chambers and Stroud, 1979) and α-chymotrypsin (CHT; Birktoft and Blow, 1972).

The knowledge that SGT should be similar in structure to BT and CHT was exploited in the crystal structure solution by using the molecular replacement method (Rossmann, 1973) to determine phases for the calculation of electron density maps. Some additional phase information was obtained by the method of multiple isomorphous replacement (MIR), though this played a much smaller role.

After the model of SGT was refined at 1.7Å resolution to give a good agreement between the measured and calculated diffraction data, it was examined and compared with a model of BT (Chambers and Stroud, 1979). As expected, the two

[1] A version of part C of this chapter has been published [Read, R. J., Brayer, G. D., Jurášek, L., & James, M. N. G. (1984) *Biochemistry 23*, 6570-6575].

Table II.1

Sequence Alignment of SGT with BT'

	16	17	18	19	20	21	22	23	24	25	26	27	28	29
SGT	V	V	G	G	T	R	A	A	Q	G	E	F	P	F
BT	I	V	G	G	Y	T	C	G	A	N	T	V	P	Y
	30	31	32	33	34	35	38	39	40	41	42	43	44	45
SGT	M	V	R	L	S	M	-	-	-	G	C	G	G	A
BT	Q	V	S	L	N	S	G	Y	H	F	C	G	G	S
	46	47	48	49	50	51	52	53	54	55	56	57	58	59
SGT	L	Y	A	Q	D	I	V	L	T	A	A	H	C	V
BT	L	I	N	S	Q	W	V	V	S	A	A	H	C	Y
		A	B	C	D									
	60	60	60	60	60	61	62	63	64	65	66	67	68	69
SGT	S	G	S	G	N	N	T	S	I	T	A	T	G	G
BT	K	S	-	-	-	-	-	G	I	Q	V	R	L	G
	70	71	72	73	74	75	76	77	78	79	80	81	82	83
SGT	V	V	D	L	-	Q	S	G^2	-	A^2	A	V	K	V
BT	E	D	N	I	N	V	V	E	G	N	E	Q	F	I
	84	85	86	87	88	89	90	91	92	93	94	95	96	97
SGT	R	S	T	K	V	L	Q	A	P	G	Y	N	G	-
BT	S	A	S	K	S	I	V	H	P	S	Y	N	S	N
	98	99	100	101	102	103	104	105	106	107	108	109	110	111
SGT	T	-	G	K	D	W	A	L	I	K	L	A	Q	P
BT	T	L	N	N	D	I	M	L	I	K	L	K	S	A
	112	113	114	115	116	117	118	119	120	121	122	123	124	125
SGT	I	N	-	-	-	-	Q	P	T	L	K	I	A	T
BT	A	S	L	N	S	R	V	A	S	I	S	L	P	T
	128	129	130	131	132	133	134	135	136	137	138	139	140	141
SGT	T	T	A	Y	N	Q	G	T	F	T	V	A	G	W
BT	S	C	A	S	A	G	T	Q	C	L	I	S	G	W

(Table II.1 continued)

	142	143	144	145	146	147	148	149	150	151	152	153	154	155
SGT	G	A	N	R	E	-	G	G	S	Q	Q	R	Y	L
BT	G	N	T	K	S	S	G	T	S	Y	P	D	V	L

	156	157	158	159	160	161	162	163	164	165	166	167	168	169
SGT	L	K	A	N	V	P	F	V	S	D	A	A	C	R
BT	K	C	L	K	A	P	I	L	S	D	S	S	C	K

	170	171	172	173	174	175	176	177	177A	178	179	180	181	182
SGT	S	A	Y	G	N	E	L	V	A	N	E	E	I	C
BT	S	A	Y	P	G	Q	I	T	-	S	N	M	F	C

	183	184	185	185A	185B	185C	186	187	188	189	190	191	192	193
SGT	A	G	Y	P	D	T	G	G	V	D	T	C	Q	G
BT	A	G	Y	L	-	E	G	G	K	D	S	C	Q	G

	194	195	196	197	198	199	200	201	202	203	204	204A	205	206
SGT	D	S	G	G	P	M	F	R	K	D	N	A	D	E
BT	D	S	G	G	P	V	V	C	S	-	-	-	-	-

	207	208	209	210	211	212	213	214	215	216	217	219	220	221
SGT	W	I	Q	V	G	I	V	S	W	G	Y	G	C	A
BT	G	K	L	Q	G	I	V	S	W	G	S	G	C	A

	222	223	223A	224	225	226	227	228	229	230	231	232	233	234
SGT	R	P	G	Y	P	G	V	Y	T	E	V	S	T	F
BT	Q	K	N	K	P	G	V	Y	T	K	V	C	N	Y

	235	236	237	238	239	240	241	242	243	244	245
SGT	A	S	A	I	A	S	A	A	R	T	L
BT	V	S	W	I	K	Q	T	I	A	S	N

'The sequence alignment is derived from an alignment of the structures of BT and SGT at cycle 78 of least squares refinement. [Sequence numbering, in terms of the sequence of chymotrypsinogen A (Hartley and Kauffman, 1966), is based on an alignment of SGT with CHT (Birktoft and Blow, 1972).] The structures were aligned initially using a program of W. Bennett, based on the principles of Rossmann and Argos (1975). The 190 underlined residues (solid, dashed or dotted lines) are considered to be structurally equivalent. The 122 residues with a solid underline are those for which the C^{α} atoms can be superimposed simultaneously within 1Å; the rms deviation in their positions is 0.52Å. (Unless otherwise

(Table II.1 continued)

noted, the relative orientation of BT derived from the superposition of these atoms is used for all figures showing comparisons of BT and SGT.) Dashed lines indicate the additional 63 residues in the set of 185 for which the C^α atoms can be superimposed within 1.9Å (approximately half of the distance between C^α atoms of sequential residues); the rms deviation for the 185 C^α atoms is 0.88Å. Dotted underlines indicate the five single residues with excursions greater than 1.9Å that join segments with solid or dashed underlines. Boxes outline identical residues paired by the alignment, with dashed boxes indicating those fortuitously aligned in non-homologous regions. There are 74 identical residues of the 223 in the sequence (33.2%). Of the 190 structurally equivalent residues, 70 (36.8%) are identical. With the more stringent criterion (1Å limit), 59 of 122 (48.4%) are identical.

[2]The refinement of SGT has shown that the amino acid sequence (Olafson *et al.*, 1975) must be corrected by the insertion of two residues after Ser76, currently interpreted from the electron density as Gly-Ala.

molecules are indeed quite similar, though there are some differences, especially in surface features.

The differences between these molecules are particularly relevant to assessing the success of two efforts to build a model of SGT based on its homology to BT (Jurášek *et al.*, 1976; Greer, 1981a). This technique of comparative model-building is potentially very valuable, for example, for designing drugs that will interact with a protein of medical importance but unknown structure (Blundell *et al.*, 1983; Blow, 1983). A critical evaluation of the comparative modelling attempts on SGT, however, revealed that there are serious shortcomings in the technique as it is presently applied.

A. Structure Solution and Refinement

Crystallization

SGT crystallizes in the orthorhombic space group $C222_1$. The first crystals of SGT were grown by the technique of equilibrium dialysis (Zeppezauer *et al.*, 1968). All of the buffers used in crystallization trials contained 10mM $Ca(CH_3CO_2)_2$ and 0.1mM NaN_3, and were adjusted to a pH of 6.2. SGT, purified as described by Olafson and Smillie (1975), was dissolved at a concentration of 10-15mg/ml in a crystallization buffer containing, in addition, 0.5M $(NH_4)_2SO_4$. This protein solution was then dialyzed against crystallization buffer containing 2.0M $(NH_4)_2SO_4$. Within 5 to 6 weeks, crystals grew to a size of up to 1mm in the longest dimension. Only rarely, however, were these crystals thicker than about 0.2mm.

Large crystals for high resolution data collection were grown by vapour diffusion, using seed crystals. In this case, the protein was dissolved initially at a concentration of 20mg/ml in 1.0M $(NH_4)_2SO_4$ buffer, and the concentration of precipitant was increased (to about 1.6M) by adding 2.5M $(NH_4)_2SO_4$ buffer until the solution just started to become turbid. The protein solution was centrifuged, seeded with a small fragment of a crushed crystal (<0.1mm on each edge), then placed in a sealed beaker surrounded by 1.8M $(NH_4)_2SO_4$ buffer. Using this technique, crystals of up to 3mm in the longest dimension and 0.5mm in the shortest dimension could be grown within about one week. These crystals were significantly thicker than crystals grown by dialysis, but they were too large to fit into the X-ray beam. Using a razor blade, it was possible to cut them fairly cleanly along planes parallel to the major crystal faces so that the maximum dimension would be no greater than about 1mm.

Data Collection

Information on the native and isomorphous derivative data sets that were used in the SGT structure solution and refinement is summarized in Table II.2. The earlier data sets were collected with a Picker FACS-1 diffractometer, using the diffractometer computing and controlling system of Lenhert (1975). Reflections were measured by omega scans, using a scan width of 0.6° and a scan speed of 2°/minute. Backgrounds were measured for 4 seconds on either side of each reflection, typically 0.8° in the 2θ direction from the center of the reflection. The later data sets, GTU2 and GTN3, were collected with a Nonius CAD4 diffractometer. The reflections were again measured by omega scans, 0.53° in width with a scan speed of 1°/minute. To measure backgrounds, the peak scans were extended by 1/4 of the scan width in each direction.

For all data sets, the reflection backgrounds were estimated by averaging the individual measurements as a function of θ and ϕ in reciprocal space. The average background values should be more accurate than the individual values;

Table II.2

SGT Data Collection

Data set (Code)	Native (GTN2)	Uranyl acetate (GTUA)[1]	Mersalyl (GTM2)[1]	Platinum chloride (GTP1)[1]	Uranyl acetate (GTU2)	Native (GTN3)
Diffractometer	Picker	Picker	Picker	Picker	CAD4	CAD4
Resolution(Å)	2.8	2.8	2.8	2.8	2.8	1.7
Cell dimensions(Å)						
a	72.04	72.05	71.71	71.54	72.52	72.29
b	50.86	51.06	51.11	50.91	51.12	50.98
c	120.42	120.54	120.42	120.53	120.21	120.09
No. of reflections measured	13253	8231	7738	8272	13451	25580
No. of matches with unique native data	5718	5718	5706	5688	5717	24878
No. of anomalous differences	0	1613	1272	1739	4780	0
Maximum absorption correction factor	1.34	1.56[2]	1.30	1.66	1.26	1.60
Maximum decay correction factor	1.29	1.26[2]	1.09[3]	1.08[3]	1.56	1.70

[1] Data collected by G. D. Brayer
[2] Correction derived by comparison with corrected GTN2 data
[3] For these data sets, decay was monitored and corrected as a function of time by a set of standard reflections

these were measured for a relatively short period of time, especially on the Picker diffractometer.

Absorption effects were corrected by the method of North *et al*. (1968). In the case of the GTUA data, a crack in the crystal caused the empirical absorption curve to be unreliable. Therefore, an approximate absorption curve was derived by comparing the derivative data to the absorption-corrected GTN2 data as a function of ϕ.

Radiation decay was evaluated as a function of θ and time for four of the data sets, then corrected by the method of Hendrickson (1976). In this procedure, decay is usually evaluated by collecting a set of reflections at the beginning and end of data collection. For GTUA, however, the decay standards had not been collected. Instead, the GTUA data were compared to the decay-corrected GTN2 data as a function of θ and time. The improvement in the heavy-atom differences was not judged to be sufficient to warrant applying this procedure to the GTM2 and GTP1 data; these data were corrected for decay as a function of time only, as monitored by several standard reflections.

Since the final quality of a protein model depends on the quality of the data against which it is refined, it is useful to discuss these data in somewhat more detail. The GTN3 data were collected from a single crystal measuring approximately 1.0mm in the longest dimension and 0.45mm in the shortest dimension. The diffraction limit was estimated by performing quick scans of reflections in several resolution ranges; it was determined that relatively few intensities would be observed at greater than about 1.7Å resolution.

To minimize the effect of radiation damage on the relatively weak high resolution reflections, the data were collected in two shells. First, about 1300 reflections covering the overall resolution limits from 22.1 to 1.7Å were collected as decay standards. Next, the data in the 2.0 to 1.7Å high resolution shell were collected. After the decay standards had been collected again, the 22.1 to 1.99Å data were collected, allowing an overlap of about 200 reflections with the high resolution shell. Finally, the decay standards were collected for a third time.

The decay standards were used to estimate parameters for the radiation damage model of Hendrickson (1976). The maximum decay correction factor applied in the high resolution shell was 1.25 (at 1.7Å resolution), corresponding to a 20% loss in intensity. For the low resolution shell, the maximum correction factor was 1.70 (41% loss in intensity) at 1.99Å resolution. When this much decay occurs, one must be concerned about the adequacy of the radiation damage model. An indication of the quality of the decay correction can be gained from the agreement between separate measurements of the intensities. $R_{\mathrm{merge}}(=\Sigma(I-\bar{I})/\Sigma I)$ had a value of 5.5% for the three sets of decay standards.

Finally, the GTN3 data were placed on an absolute scale using the program ORESTES (Thiessen and Levy, 1973), which found a scale of 17.6 and a mean isotropic temperature factor (*B*) of 17.3Å^2.

Figure II.1 shows, as a function of resolution, the fraction of reflections classified as observed by the criterion that I is greater than either one or two times its standard deviation, σ(I). From the trend in Figure II.1, one would expect a diminishing return in the amount of additional information to be obtained by measuring data at higher than 1.7Å resolution. Still, it is apparent that the crystal diffracts somewhat beyond 1.7Å resolution. The 2.0Å boundary between the two resolution shells in which data were collected shows up in this histogram as a slight

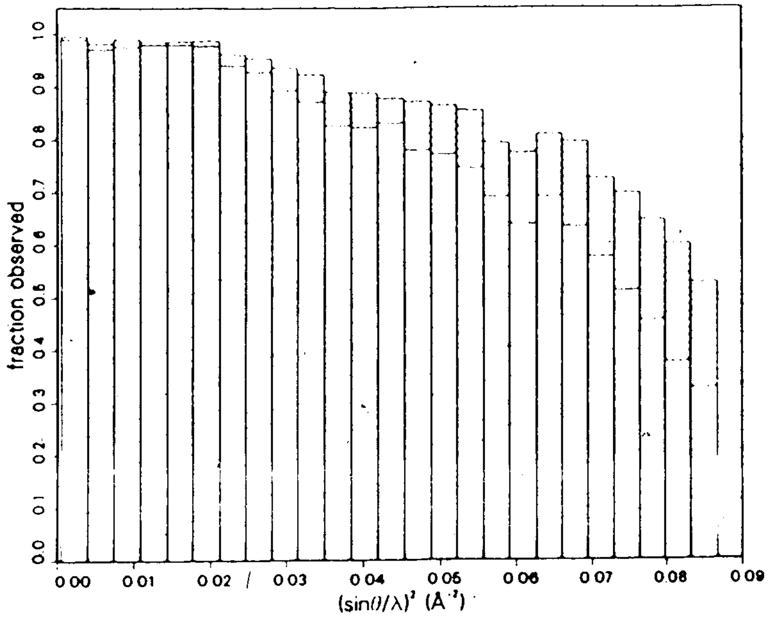


Figure II.1. Observed Reflections in SGT High Resolution Data. The histograms indicate, as a function of resolution, the fraction of reflections in the SGT high resolution data set classified as observed by the criterion that either I>σ(I) (upper histogram) or I>2σ(I) (lower histogram).

discontinuity at the corresponding value of $(\sin\theta/\lambda)^2$, i.e., $0.0625Å^{-2}$. Had the data not been collected in shells, there would have been a greater loss in the number of reflections with significant intensities at high resolution.

Molecular Replacement

The structure of SGT was solved primarily by the molecular replacement method (Rossmann, 1973), with some additional phase information from isomorphous replacement. In the molecular replacement method, one requires a search model that is expected to be fairly similar to the unknown structure. Two reasonable models for SGT are BT [the model of Chambers and Stroud (1979) was used], which has 33% sequence identity based on structural alignments, and CHT (Birktoft and Blow, 1972), which has 32% sequence identity. Coordinates for BT and CHT were obtained from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977).

The first task in the molecular replacement method is to find the rotational parameters that place the model structure in the same orientation as the homologous structure in the crystal. The rotation function (Rossmann and Blow, 1962) that solves this problem can be understood in terms of the Patterson function. Intramolecular vectors, i.e., peaks corresponding to vectors between atoms within the same molecule, will be concentrated near the origin of the Patterson map. A similar set of ~~vectors, but~~ in a different orientation, will be found near the origin of the model Patterson map. The orientation of the model Patterson map that maximizes the agreement near the origin with the observed Patterson map will be the orientation required to orient the model like the unknown structure.

This work used Crowther's (1973) fast rotation function, which is similar in concept to that of Rossmann and Blow (1962). Normalized structure factor amplitudes ($|\mathbf{E}|$s) were used. For the observed GTN2 data, $|\mathbf{E}_o|$s were determined using the program ORESTES (Thiessen and Levy, 1973), which found an absolute scale of 19.0 and a mean *B*-factor of 17.2Å^2. $|\mathbf{E}_c|$s were computed for molecular replacement models placed in a cubic cell of symmetry P1, with a unit cell edge of 60Å. BT and CHT have dimensions of approximately 40×40×45Å; with an upper integration limit[2] of about 21Å, one should expect little contribution of intermolecular vectors to the rotation function (Lifchitz, 1983). The rotation function was calculated initially on a 5° grid for an asymmetric unit of rotation function space (Rao *et al.*, 1980); the Euler angles α, β, and γ defining the orientation are those given by Crowther (1973). Promising peaks were evaluated with a 1° interval in the Euler angle β (the program does not allow finer than 5° sampling in the angles α and γ) and the best interpolated peak position was estimated by eye. The interpolation was to the nearest degree in α and γ, and to the nearest tenth of a degree in β. Some

[2] The integration limits correspond to the distances from the origin over which the Patterson maps are compared.

rotation function results are summarised in Table II.3. Most of the work with the rotation function involved BT, so the discussion will center on the use of that model.

As shown in Table II.3, there are a number of variable parameters in the use of the rotation function. Judging from the results obtained when all of BT was used as the model, the most important variables are the resolution limits and the number of strong reflections accepted. If both accuracy and signal-to-noise are considered, the best results were obtained when medium resolution (d_{min}=3.5Å) data were used. (Note that in the second run using 5.0Å data, though the orientation obtained from the rotation function peak was very accurate, the peak was quite broad and was difficult to interpolate precisely.) The lack of accuracy of the run using 2.8Å data might be attributed in part to the structural differences that become more significant at this resolution. Another factor could be the low value of the upper integration limit; in the version of the fast rotation function program that was used, this integration limit must be no greater than about 6×d_{min}. According to Lifchitz(1983), the upper integration limit should be the mean value of the radii for the principal axes of an ellipse that approximates the shape of the molecule. On this basis, 21Å would be a good choice. However, the correct answer was obtained with the 5.0Å data only when this limit was 27Å. The results in Table II.3 indicate little sensitivity to the lower integration radius, but only values near d_{min} were

Table II.3

Rotation Function Results

Search model	Interpolated peak position (height)[1]	Orientation error(°)[2]	Resolution limits(Å)	Integration limits(Å)	No. of $\|E_o\|$ ($\|E_c\|$) accepted[3]
BT					
all	33,60.4,225 (3.85)	109.2	10.0-5.0	6.0-27.0	73 (1312)
	69,41.0,96 (3.85)	2.4	10.0-5.0	6.0-27.0	402 (1312)
	114,49.8,122 (3.60)	67.1	10.0-5.0	6.0-21.0	402 (1312)
	67,44.1,94 (5.54)	3.5	10.0-3.5	5.0-21.0	595 (1433)
	68,44.7,95 (5.97)	2.7	10.0-3.5	5.0-21.0	1722 (2498)
	67,43.9,94 (5.62)	3.4	10.0-3.5	3.0-21.0	595 (1433)
	72,46.0,89 (5.32)	6.9	10.0-2.8	4.0-16.8	1136 (2018)
Jurášek model[4]	67,43.2,95 (5.13)	2.2	10.0-3.5	5.0-21.0	595 (1480)
conservative model[4]	12,22.3,68 (4.16)	83.1	10.0-3.5	5.0-21.0	595 (1449)
CHT					
all	26,87.5,-4 (5.80)	2.5	10.0-3.5	3.0-21.0	595 (1920)

[1] Peak position is expressed in the Euler angles defined by Crowther (1973), given in degrees. Peak height is measured in number of standard deviations above the mean.

[2] Orientation error is measured by the angular deviation from the final molecular replacement orientation.

[3] $|E|$s were accepted when they were larger than a minimum cutoff value. Due to program limits, no more than 2500 reflections could be accepted.

[4] To construct the Jurášek and conservative models of BT, some non-homologous parts were removed. The criteria, which differed in their stringency, are explained in the text.

used.

Reflections within the resolution limits are screened further on the basis of a minimum intensity value. When too few reflections were accepted for the low resolution data, the correct orientation was indicated by only a small peak, with a height of 3.04 standard deviations above the mean. For the 3.5Å data, increasing the number of reflections accepted improved both the height and the accuracy of the rotation function solution.

It was expected that, if the model of BT were edited to remove parts that differ from SGT, a better molecular replacement model would result. For the rotation function, at least, this expectation was not entirely realized. Two edited models were used. The first, which will be called the Jurášek model, was produced by removing segments that Jurášek *et al*. (1976) judged would take up a different conformation in SGT than in BT. For this model, the following 25 residues, containing 174 atoms, were removed: Asn34-Tyr39, Asn97-Thr98, Ser110-Ile121, Gln175-Asn179 and Ser202-Gly207. Since BT contains 1629 non-hydrogen atoms, the Jurášek model comprises about 89% of BT. The second model, called the conservative model, was made by retaining only those main-chain and side-chain atoms of BT that were considered most likely to be conserved in SGT. The choice of conserved atoms was based on the sequence alignments and assignments of conserved structure in James *et al*. (1978). In this model there were 898 atoms, about 55% of BT. The rotation function calculated with the Jurášek model had a slightly less favourable signal-to-noise ratio, but the position of the peak was more accurate than in the comparable run using all of BT (Table II.3). With the conservative model, the correct peak became the third highest, with a height of only 3.39 standard deviations above the mean.

For the rotation function calculated with CHT as a model, the parameters were chosen according to the experience with BT, and an unambiguous peak resulted. When the orientation of CHT was compared with that of BT, it was found that the two orientations agreed within a few degrees. Combined with the height of the peaks for the two models, this provided convincing evidence that the orientation was correct.

The second task in the molecular replacement solution of a crystal structure is to find the translation vector which places the oriented model in the correct place in the unit cell. This translation part of the molecular replacement problem was solved using a brute-force technique. The program BRUTE, written by M. Fujinaga, calculates structure factors for an oriented model translated over an array of possible positions in the unit cell. To save computing time, the structure factor calculation is factored into molecular transform and translational components. The possible translation vectors are evaluated by the agreement between the observed and calculated structure factors, measured either by *R*-factors[3] or by correlation coefficients.[4] For this work, the correlation between $|\mathbf{F}_O|$ and $|\mathbf{F}_C|$ was used.

In some tests of BRUTE performed since this work was done, M. Fujinaga (personal communication) has observed that the best results, measured by signal-to-noise criteria, are generally obtained by using the correlation coefficient between $|\mathbf{F}_O|^2$ and $|\mathbf{F}_C|^2$ instead. This observation can be rationalized as follows. Generally we have used a fairly narrow resolution range of data for these computations, in which case this correlation coefficient is virtually the same as the correlation between $|\mathbf{E}_O|^2$ and $|\mathbf{E}_C|^2$. As discussed in Chapter IV, the correlation coefficient on $|\mathbf{E}|^2$ is an estimate of $\sigma_A{}^2$ (Hauptman, 1982), which in turn varies monotonically with the mean figure of merit of the model phases (Srinivasan and Chandrasekaran, 1966). Alternatively, noting that the mean value of a map will be zero when the reciprocal lattice origin term is zero and applying the convolution theorem, this correlation is equivalent to the correlation between the two origin-removed Patterson maps (M. Fujinaga, personal communication). Thus the posi-

[3] The *R*-factor is a measure of agreement between observed and calculated structure factors defined as $R=\Sigma||\mathbf{F}_O|-|\mathbf{F}_C||/\Sigma|\mathbf{F}_O|$.

[4] The coefficient of correlation between two variables x and y is given by $r=\Sigma(x-\bar{x})(y-\bar{y})/[\Sigma(x-\bar{x})^2\Sigma(y-\bar{y})^2]^{1/2}$.

tion of the model that maximizes the correlation between $|F_o|^2$ and $|F_c|^2$ will be the one that minimizes the phase error and maximizes the agreement with the observed origin-removed Patterson map. This may be contrasted with the traditional translation functions, e.g., that of Tollin (1966), in which only those parts of the Patterson that arise from a particular symmetry axis are compared.

In order not to miss the correlation peak one must use a sufficiently fine translation grid, which depends on the resolution of the data used. Some numerical experiments suggested that a grid spacing of $d_{min}/4$ would be adequate, so this was used in most translation experiments. Rabinovitch and Shakked (1984) have also suggested a value of $d_{min}/4$, although they were using *R*-factors to judge the translations. Subsequent experience indicates that it is best to repeat the search with the translation origin offset by 1/2 grid unit in each of x, y and z. Though one does not entirely miss a peak when the grid spacing is $d_{min}/4$, the correlation can be reduced sufficiently that it is difficult to distinguish the peak from noise, especially if the molecular replacement model is not particularly good.

Since all possible choices of unit cell origin are equally valid, it is not necessary to search the entire unit cell. The volume to be searched is reduced to the unique set of vectors relative to a possible choice of origin, which is not necessarily equivalent to the crystallographic asymmetric unit. For $C222_1$, as for other orthorhombic space groups, only the vectors from 0 to 1/2 in each of the unit cell axes *a*, *b* and *c* need be considered.

The first attempts to solve the translation problem used BT oriented according to the 2.8Å rotation function result. In retrospect, this was the worst possible choice since this orientation was in error by 6.9° (Table II.3). At the time, however, it seemed that the use of higher resolution data should lead to a sharper and more accurate rotation function peak. Also, the rotation function with CHT had not yet been calculated, so the use of CHT was not considered at this point.

In different attempts to solve the translation problem, various shells of data were used: 4-5Å, 8-10Å or 10-20Å. No consistent answer emerged from these attempts, and in no case was a peak observed much above the background noise. An attempt was made to use packing restrictions to reduce the ambiguity. Hendrickson and Ward (1976) measure the amount of molecular overlap with a packing function defined as the volume taken up in the cell by the whole set of symmetry-related molecules, divided by the volume that would be occupied if there were no molecular overlap. This packing function takes a maximum value of one when the symmetry-related molecules do not interpenetrate. Even with packing information, however, it was not possible to resolve the ambiguity. In hindsight, this failure is not surprising; the correct peak was never the highest in any of the initial attempts to solve the translation problem.

Several observations suggested that the failure was caused by inaccuracy of the orientational parameters. The several runs of the rotation function gave peaks that differed by several degrees. Also, experience with BRUTE in this laboratory showed that correlation coefficients of 0.3 or greater would be found for 4-5Å data at the correct translation; using the orientation of BT found with the 2.8Å rotation function, the highest correlations had been of the order of 0.18. At this point, the orientation of BT was verified by calculating the rotation function with CHT. Since CHT gave a higher rotation function peak than BT (Table II.3), it was used in the next stage of the molecular replacement work.

The new approach was to perform a brute-force 6-dimensional search for the correct orientation and position of the molecular replacement model. This approach is quite computationally intensive, both because of the increased number of dimensions in the problem and because the molecular transform must be recomputed for each new orientation. Therefore, it was necessary to limit the search to render it practical.

The search of the orientational parameters was limited to values near the rotation function peak, and this search was performed in terms of the orientational space θ_+, β, θ_- ($\theta_+=\alpha+\gamma$, $\theta_-=\alpha-\gamma$). As discussed by Lattman (1973), who defined similar variables using an alternate Euler angle convention, these variables are more nearly locally orthogonal than α, β, γ, so searches in this space are more efficient. The model was placed with its center of mass at the origin so that reorientations would not affect the translational component. For a model with a radius of about 20Å, a rotation of 3° will shift the atoms a maximum of about 1Å, so the increments in θ_+, β and θ_- were chosen to be equivalent to rotations of about 3° initially. These increments can be calculated using an equation of Lattman (1973); the rotational difference between two orientations, χ_d, is approximated by

$$\chi_d{}^2 \simeq \Delta\theta_+{}^2\cos^2(\beta/2) + \Delta\beta^2 + \Delta\theta_-{}^2\sin^2(\beta/2)$$

(In the current version of BRUTE6D, the rotational search is performed by orienting the model, then applying rotations about the new x, y and z axes. These rotational variables are also locally orthogonal, but they are more easily understood, and β values of 0 or 90° do not lead to special cases.)

The search of the translational parameters was limited to parts of the unit cell where the packing function (Hendrickson and Ward, 1976) had values of about 0.9 or greater. Two rectangular regions, centered on the two peaks in the packing function, were chosen: x=5.5-29.5Å, y=1.5-15.0Å, z=8.5-19.0Å and x=10.5-28.5Å, y=7.0-25.0Å, z=40.0-50.5Å. These regions contain only about 12% of the volume that would be searched if packing were not considered. In a search on a 1Å grid, it is possible to miss the center of a peak by as much as 0.87Å. The same coverage can be attained much more efficiently using a body-centered cubic grid. With two runs on a 1.5Å grid, the second displaced from the first by 0.75Å in each of x, y and z, the maximum distance by which a peak can be missed is only 0.84Å, but only 0.59 times as many points need be evaluated as with the 1Å grid.

Even with these limits, the search using 4-5Å data (925 reflections) with 8 possible orientations centered on the CHT rotation function result took about 6.5 hours on an FPS 190L array processor. Therefore, it was not practical to repeat the search using data from several alternative resolution ranges. The range of 4-5Å was chosen in part because other structures had been solved in this laboratory using data from that range. It is desirable to use low resolution data in order to maximize the translation grid spacing, and thereby minimize the number of points to be evaluated. Also, the effect of differences between the crystal structure and the molecular replacement model will be less pronounced at lower resolution. However, molecular replacement models do not account for the disordered solvent present in the crystal. Since the contribution of disordered solvent will be minimal at higher than 5Å resolution, the choice of 4-5Å data is reasonable.

Using this 6-dimensional search procedure, an unambiguous orientation and position was found for CHT, with a correlation coefficient of 0.208. Successively finer searches around this solution increased the correlation to 0.274.

(Some results of 6-dimensional molecular replacement trials are summarized in Table II.4.) The model at this point included all of CHT, even those parts (such as the N-terminal 8 residues) that could not have any counterpart in SGT. In addition, an analysis of crystal packing indicated some regions of the CHT model where alterations must be considered. Accordingly, a new molecular replacement model was generated by removing 56 residues, or 24% of the structure. When this model was adjusted with BRUTE6D, the correlation increased to 0.288. The model was then edited further by removing all atoms further than 19Å from the center of mass, leaving 1089

Table II.4

Six-dimensional Molecular Replacement Results

Search model	Correlation coefficient[1]	Euler angles(°)	Center of mass(Å)
CHT			
all	0.274	28.0,87.4,-6.0	19.4,22.1,42.5
close contacts deleted	0.288	28.0,87.4,-6.0	18.5,22.2,43.0
spherically truncated[2]	0.322	27.5,87.8,-6.0	18.2,22.7,42.7
BT			
Jurášek model	0.344	66.1,-42.4,-81.9	19.4,21.8,41.2
close contacts deleted[2]	0.339	66.3,-42.6,-82.4	18.4,21.4,42.6

[1] All correlation coefficients were calculated with the 925 reflections in the 4-5Å resolution shell.
[2] Final molecular replacement model.

atoms (63% of CHT). This was motivated by the knowledge that the cores of homologous proteins are more similar than their peripheries. Adjustment of this new edited model with BRUTE6D increased the correlation further to 0.322.

The molecular replacement solution was verified by phasing the heavy atom differences for GTUA with the edited CHT model of SGT. A single peak in the electron density was found. This peak was consistent with the single site determined from the difference Patterson map for this derivative (see below).

The molecular replacement with BT could have been performed by positioning BT at the same site as CHT. Instead, the 6-dimensional search technique was tested further by repeating the limited search using BT. Because of the experience that an edited model was superior for the 6-dimensional search, the Jurášek model of BT described above was used. To make comparisons with CHT easier, the symmetry-related rotation function peak corresponding to the CHT orientation was used for the starting orientation. A single solution, consistent with the CHT solution, was found. After finer adjustment of the parameters, this model gave a correlation of 0.344. A further 21 amino acid residues were removed after an analysis of crystal packing contacts, but in contrast to the CHT case, the correlation after further adjustment dropped to 0.339. Nonetheless, because of the relief of packing contacts, this was chosen as the final molecular replacement model based on BT.

In the course of using the program BRUTE6D, it was observed that as the correlation peak became higher, so did the level of the background. Though the correlation was calculated on $|F|$, this is easier to understand in terms of the interpretation of the correlation on $|F|^2$ as a correlation between origin-removed Patterson maps. The background, or the translation-independent part of the correlation, must arise from the intramolecular vectors in the Patterson map. To test this interpretation, the symmetry-related molecules were ignored by working in the space group C1. Native data in the 4.2-4.5Å resolution shell were expanded to C1; using BRUTE6D, correlation coefficients were calculated for a number of orientations of BT centered on its final orientation. The highest correlation, 0.092, was obtained at the final molecular replacement orientation, while a correlation of only 0.023 was obtained for the orientation from the 2.8Å rotation function. With the orientation from the first 3.5Å rotation result in Table II.3, the correlation was 0.083. This suggests that part of the improvement over the 2.8Å rotation function may come from neglecting the higher resolution data. A second experiment, still in the space group C1 but using unexpanded 4-5Å data, gave similar results but the correlation did not fall off quite so cleanly as a function of angular deviation from the final orientation. This approach may be of use for refining the orientations obtained from rotation functions without increasing the dimensionality of the molecular replacement problem.

The phases from BT and CHT were combined by the method of Hendrickson and Lattman (1970) and were used to phase heavy atom differences in order to solve the isomorphous derivatives. At this point in the work, phase probabilities were estimated using the equations of Woolfson (1956) and Sim (1959,1960), with the parameter Σ_Q estimated as suggested by Blundell and Johnson (1976, p. 418). As discussed in Chapter IV, this leads to a drastic overestimation of model phase accuracy.

Multiple Isomorphous Replacement

Several sets of data for isomorphous derivatives had been collected by G. D. Brayer. Of these, three proved to be suitable for further analysis (Table II.2). The native protein data set GTN2 was used to determine the isomorphous differences.

In the derivative data sets, Friedel pairs had not been collected throughout data collection; instead, after a number sufficient to judge the quality of the anomalous signal had been collected, only the positive θ member of the pair was measured. Since the data were collected with the reciprocal lattice index h varying least rapidly, the Friedel pairs all have low h indices, typically less than 8.

For the calculation of both isomorphous and anomalous differences, the local scaling procedure of Smith and Hendrickson (1982) was used to minimize systematic errors in the differences. (The programs used to calculate and apply the local scale factors were kindly supplied by S. Sheriff.) Isomorphous differences are sufficiently large as a rule that this procedure is probably unnecessary, but the use of local scales has been shown to improve dramatically the quality of anomalous differences (Hendrickson and Teeter, 1981; Smith and Hendrickson, 1982).

The only derivative for which the difference Patterson map was interpretable was GTUA. The solution of this derivative in terms of a single site was straightforward, especially when data to only 5Å resolution were used in the calculation of the Patterson map. When data to higher resolution were included, the height of the peaks dropped substantially. Both the isomorphous and anomalous difference Patterson maps could be interpreted; the quality of these maps is demonstrated in Figure II.2, in which the Harker section at w=1/2 is shown for each map. The GTUA derivative was the most useful, even though the crystal from which these data were collected was cracked. Also, though the anomalous signal was strong enough that the anomalous difference Patterson was interpretable, only a subset of the anomalous differences were available from the GTUA data. Accordingly, a second uranyl acetate derivative data set, GTU2, was collected and a full set of anomalous differences was obtained. However, it became evident that the uranyl site in the new crystal had a relatively low occupancy, so the GTUA data were not discarded.

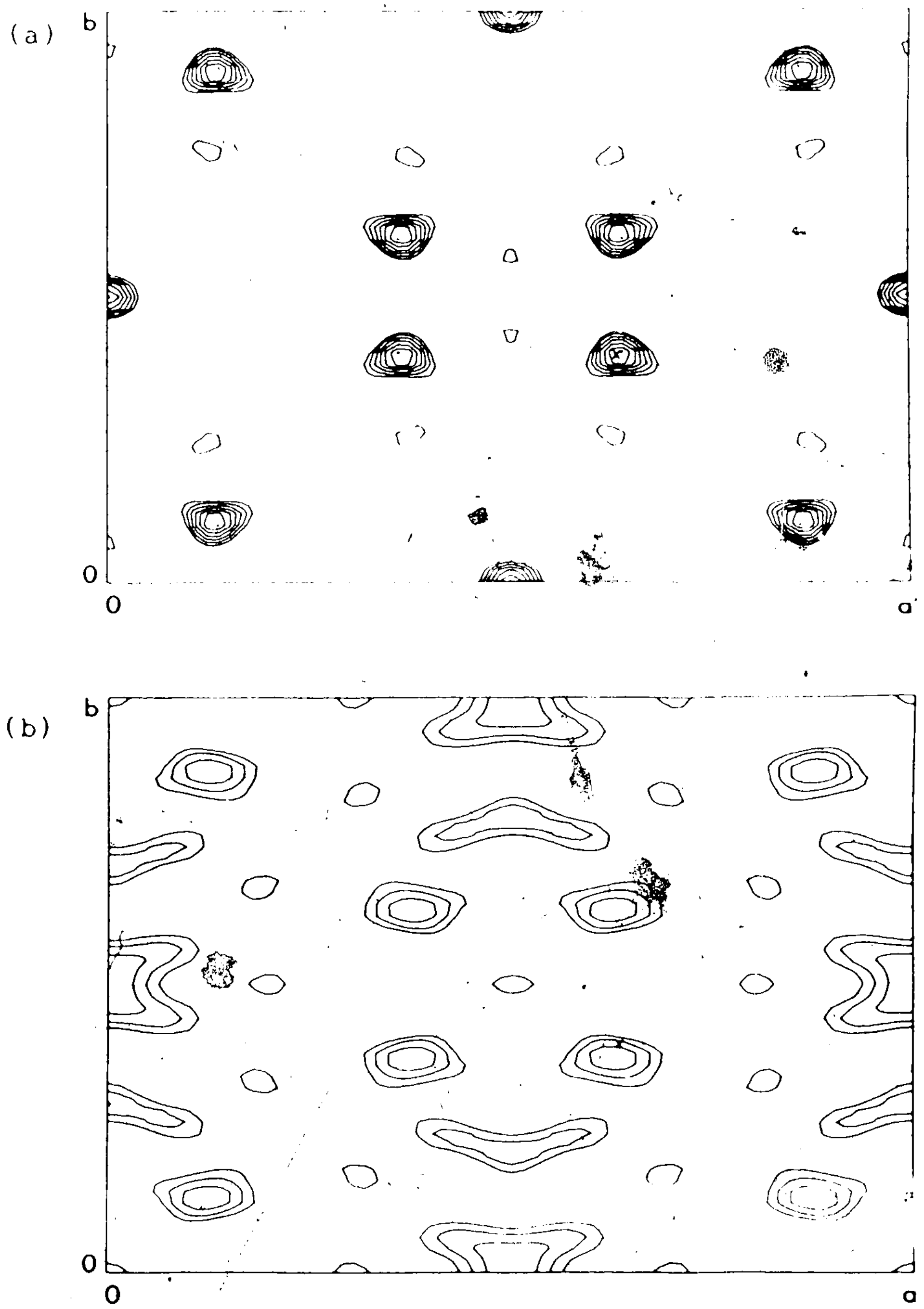


Figure II.2. GTUA Isomorphous and Anomalous Difference Patterson Maps. Harker sections at w=1/2 for GTUA uranyl acetate derivative from: (a) isomorphous difference Patterson map (b) anomalous difference Patterson map. Both maps use data to 5.0Å resolution. The 8 peaks in general positions are all symmetry-related; the cross in each map indicates 2x, 2y for the refined uranyl site.

When the molecular replacement solution was achieved, the other derivatives were solved using the combined phases from the final BT and CHT molecular replacement models to phase the isomorphous heavy-atom differences. This procedure was also used to find the correct hand and origin for the uranyl acetate derivative. For each derivative, the major site was close to the site determined for GTUA; for GTP1, two additional minor sites were chosen and for GTM2 a single additional site was found.

The heavy atom sites were refined using a program of Adams *et al*. (1969). During this refinement, the two minor sites for GTP1 were discarded. Refinement of heavy atom parameters was not stable when data beyond 4Å resolution or non-centric data were included. Therefore, only centric data from 10 to 4Å resolution were used for the final cycles of refinement. Refinement of thermal parameters did not behave well with this limited resolution range, so the *B*-factors were set to reasonable values, and only the occupancies were varied. The final set of heavy atom sites is given in Table II.5.

The quality of the MIR phases and the contribution of each derivative can be judged by several criteria. These are summarized in Table II.6. Figure II.3 shows the phasing power of each derivative and the mean figure of merit as a function of resolution. (The overall mean figure of merit for data to 2.8Å resolution is 0.559.) From the results in Tables II.5 and II.6 and in Figure II.3, one can see that

Table II.5

Heavy Atom Binding Sites

Derivative	Occupancy'	x/*a*	y/*b*	z/*c*	B(Å^2)
GTUA	41.8	0.3167	0.1980	0.2277	15.0
GTM2	18.2	0.3303	0.1993	0.2239	10.0
	3.2	0.4026	0.8556	0.0125	10.0
GTP1	15.1	0.3314	0.2001	0.2244	15.0
GTU2	22.4	0.3148	0.1937	0.2274	20.0

'Occupancy is measured in units approximately of electrons.

the MIR phasing of SGT was not particularly successful. This was probably due to the lack of isomorphism indicated by changes in cell dimensions by as much as 0.7% (*a* in GTP1 *vs. a* in GTN2, Table II.2). Nonetheless, it might have been possible to make improvements. Judging from its occupancy (Table II.5), the minor site of GTM2 should probably have been deleted. The data in Table II.6 indicate that GTM2 and GTP1 have major errors. In addition, other minor sites might have been found for the derivatives. However, the phases obtained from molecular replacement seemed to be so much more accurate that it was not deemed worthwhile to pursue the MIR phasing any further; the mean figure of merit for the combined BT and CHT phases was 0.780. Though it turned out that the accuracy of the model phases had been greatly overestimated (Chapter IV), the electron density map computed using combined phases was sufficiently good that the structure of SGT could be developed.

Table II.6

Heavy Atom Refinement Statistics

	R-factors (numbers of reflections used)				
Derivative	R_{iso}[1]	R_{ano}[1]	R_C[2]	rms F_H[3]	rms E_H[3]
GTUA	0.264 (5718)	0.100 (1613)	0.754 (870)	118.4	148.9
GTM2	0.122 (5706)	0.046 (1272)	0.921 (868)	41.7	78.1
GTP1	0.138 (5688)	0.058 (1739)	0.888 (866)	42.9	87.7
GTU2	0.129 (5717)	0.089 (4780)	0.792 (870)	58.6	69.0

[1] $R_{iso}=\Sigma||F_{PH}|-|F_P||/\Sigma|F_P|$ and $R_{ano}=\Sigma||F_{PH^+}|-|F_{PH^-}||/\Sigma|F_{PH}|$ where $|F_{PH}|=(|F_{PH^+}|+|F_{PH^-}|)/2$ for the reflections for which both measurements are available. These *R*-factors were calculated after application of local scales.

[2] $R_C=\Sigma|||F_{PH}|\pm|F_P||-|F_H||/\Sigma||F_{PH}|\pm|F_P||$, where the sums are taken over centric data (Cullis *et al.*, 1961).

[3] F_H is the scattering contribution from the heavy atom model and E_H is the lack-of-closure error (Blow and Crick, 1959).

SGT Model-building and Refinement

Molecular replacement and MIR phases were combined by the method of Hendrickson and Lattman (1970). Model phase probabilities were estimated, as discussed above, with the equations of Woolfson (1956) and Sim (1959, 1960), and the estimate of Σ_Q given by Blundell and Johnson (1976). The electron density map used in the initial model-building was computed with the Fourier coefficients $[m_{comb}|F_o|\exp(i\alpha_{comb})]$. The starting model for SGT was a "mutated" BT, generated by substituting the side-chains of non-conserved amino acids according to the sequence

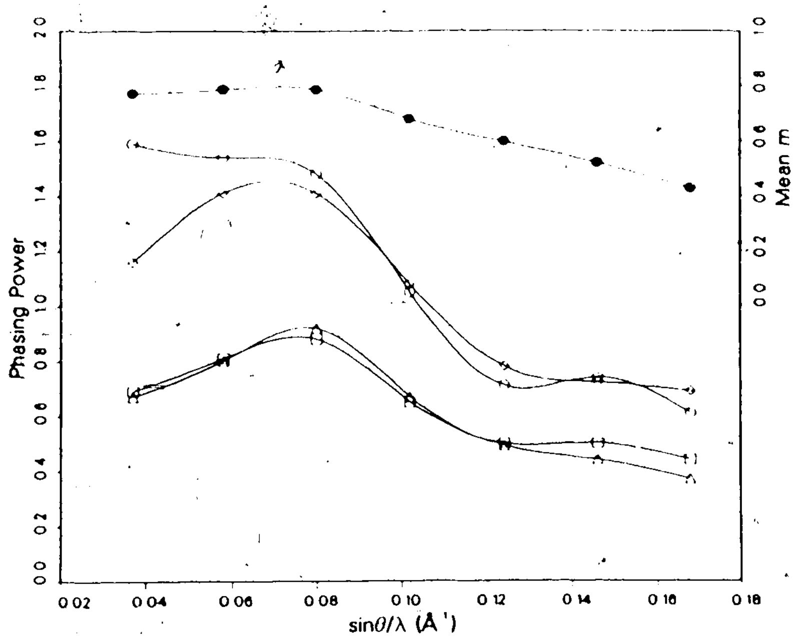


Figure II.3. MIR Phasing Power and Figures of Merit. Variation of phasing power[=(rms F_H)/(rms E_H)] and mean figure of merit as a function of resolution. The lower curves represent the phasing power of GTUA (open circles), GTM2 (squares), GTP1 (triangles) and GTU2 (diamonds). The upper curve (filled circles) and the scale to the right shows the mean figure of merit as a function of resolution.

alignment of Jurášek *et al.* (1976). The structure of BT was that of Chambers and Stroud (1979), obtained from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977). The mutated model was fit into the electron density, where possible, and amino acid residues were deleted from the model in regions where the map was unclear.

This first model, which contained 204 residues, was refined for 7 cycles against the GTN2 data from 6.0-2.8Å resolution having I>3σ(I), using the reciprocal space restrained-parameter least-squares refinement program of

Hendrickson and Konnert (1980). (The course of the structure refinement is summarized in Table II.7.) The *R*-factor on these data dropped from 0.487 to 0.425. Maps were computed using the following expressions for Fourier coefficients: $[(2|\mathbf{F}_o|-|\mathbf{F}_c|)\exp(i\alpha_c)]$, $[(2m_c|\mathbf{F}_o|-|\mathbf{F}_c|)\exp(i\alpha_c)]$, and $[m_{comb}|\mathbf{F}_o|\exp(i\alpha_{comb})]$. (Model phase probabilities were determined as for the initial combined phase map.) Examination of these maps revealed that there was a serious model bias in the combined phase maps and that parts of BT included in the first model of SGT should have been omitted. This observation prompted the work on phase probabilities and map coefficients that culminated in the results reported in Chapter IV.

The approach described in Chapter IV evolved in parallel with the structure of SGT. The exact expression for map coefficients varied through the structure refinement, but between cycles 7 and 35 the philosophy was to use coefficients that reduce model bias in the combined phase map, where the expression to reduce model bias varies according to the extent to which the model determines the phase for an individual reflection. After cycle 35, the MIR phases had very little influence on the combined phases, so their use was discontinued.

At each point of manual intervention in the refinement process, the model was examined and evaluated on an MMS-X interactive graphics device (Barry *et al.*, 1976), using the macromolecular modelling program M3 written by C. Broughton

Table II.7

Course of Least-squares Refinement

Cycle number	Data used(Å)	Number of reflections	R-factor	Comments
1	6.0-2.8	4756	0.487	Start refinement with model containing 204 amino acid residues. Use GTN2 reflections having I>3σ(I).
7	6.0-2.8	4756	0.425	Rebuild model 3 times, calculating map with new combined phases each time. Correct sequence alignment at Ser60B-Thr67.
8	6.0-2.8	4756	0.390	Restart with 200 residues.
15	6.0-2.8	4756	0.351	Last refinement cycle using GTN2 data.
16	6.0-2.8	4826	0.361	Use GTN3 reflections having I>2σ(I). In following cycles, add higher resolution data slowly.
35	6.0-1.7	17052	0.341	Stop using combined phases in maps. Adjust protein model and add first solvent molecules, interpreted as water.
36	6.0-2.5	6733	0.318	Restart with 205 residues and 11 waters. Start individual B-factor refinement.
45	6.0-1.7	17052	0.284	Adjust protein and solvent model.
46	6.0-2.5	6733	0.274	Restart with 212 residues and 20 waters.
57	6.0-1.7	17052	0.223	Adjust model twice, with map calculation between. Reassign solvent site as Ca^{2+} ion.
58	6.0-2.5	6733	0.220	Restart with 220 residues, 44 waters and 1 Ca^{2+}.
67	6.0-1.7	17052	0.189	Adjust model. Find 2 residue insertion relative to published sequence; include these and flanking residues as alanines.
68	6.0-2.5	6733	0.193	Restart with completely traced polypeptide (223 residues, some still left as alanines), 144 waters and 1 Ca^{2+}.
78	6.0-1.7	17052	0.159	Adjust model and deduce sequence from density at insertion.
79	6.0-1.7	[illegible]	0.167	Restart with model that includes 178 waters.

(Table II.7 continued)

87	6.0-1.7	17052	0.152	Adjust model. Reinterpret residue 79 as alanine.
88	6.0-1.7	17052	0.154	Restart with model that includes 211 waters.
96	6.0-1.7	17052	0.141	Last cycle before increasing resolution range of data.
97	8.0-1.7	17396	0.148	Add 8.0-6.0Å data.
100	8.0-1.7	17396	0.147	Adjust model.
101	8.0-1.7	17396	0.160	Restart with model including 208 waters. All side chains of the protein are now present.
104	8.0-1.7	17396	0.142	Last cycle before lowering data acceptance criterion.
105	8.0-1.7	20046	0.162	Add reflections having I between 1 and 2 times $\sigma(I)$.
108	8.0-1.7	20046	0.158	Adjust model.
109	8.0-1.7	20046	0.163	Restart with model including 203 waters. Tighten restraints on model geometry.
114	8.0-1.7	20046	0.161	Perform minor adjustments of a few side chains. Delete some questionable solvent sites.
115	8.0-1.7	20046	0.163	Restart with model including 192 waters.
119	8.0-1.7	20046	0.161	End of refinement. No significant changes since cycle 114.

(Sielecki *et al.*, 1982). Incorrectly positioned amino acid residues were adjusted, if possible, or removed if the electron density was not yet clear enough, and missing residues were added as they became visible. In some cases, the main-chain was clearly visible, while the side-chain was not; these residues were temporarily replaced by alanines in the hope that further refinement would clarify the details sufficiently.

In the early stages of refinement, extensive intervention was necessary. For example, at cycle 7 the protein model was adjusted three times before refinement was restarted. The first time, almost the entire model was rebuilt; 41 residues were removed and 7 were added. This new model was used to calculate phases that were then combined with MIR phases to produce a new map. Using the new map, reasonable positions could be found for 27 residues that had not been in the model, though 3 more residues were deleted. In this map, an incorrect sequence alignment involving the segment Ser60B-Thr67 was detected. The second new model was used to produce another new map, from which the positions of 6 additional residues could be deduced. Without the benefit of refinement, this extensive model rebuilding lowered the *R*-factor from 0.425 to 0.390, even though the new model contained 4 fewer residues. As the refinement continued, progressively less manual intervention was required at each stage.

Not until cycle 67 was the polypeptide chain tracing completed. At this point, it was recognized that the published sequence of SGT (Olafson *et al.*, 1975) should be corrected by the insertion of 2 residues near position 76. Four residues, Gln75-Ser79 in the current sequence, were included as alanines for the next set of refinement cycles. In the electron density map calculated at cycle 78, these residues were interpreted as Gln-Ser-Gly-Ser. Since then, Ser79 has been reinterpreted as an alanine, and it has become apparent that residue 77 is not a glycine. Electron density in this region of the final map, shown in Figure II.4, is consistent with a disordered side-chain at position 77. The side-chain density for this residue is very weak compared to the main-chain density, which persists to a contour level of 0.70e/Å^3.[5] Though it could be interpreted as a threonine (as shown in Figure II.4) or a valine, the weakness of the density suggests that this residue might have a longer side-chain that is disordered about the χ_1 rotation. A longer side-chain could be accomodated, as it would project into the solvent. Because its identity is ambiguous, residue 77 has been left as a glycine. Since the side-chain is quite disordered, its omission should have minimal effect on the phasing model.

[5] All electron density maps were calculated omitting the contribution of the F_{000} term and thus have a mean electron density of zero. Contour levels therefore do not refer to the actual electron density but rather to the deviation above the mean density of the map.

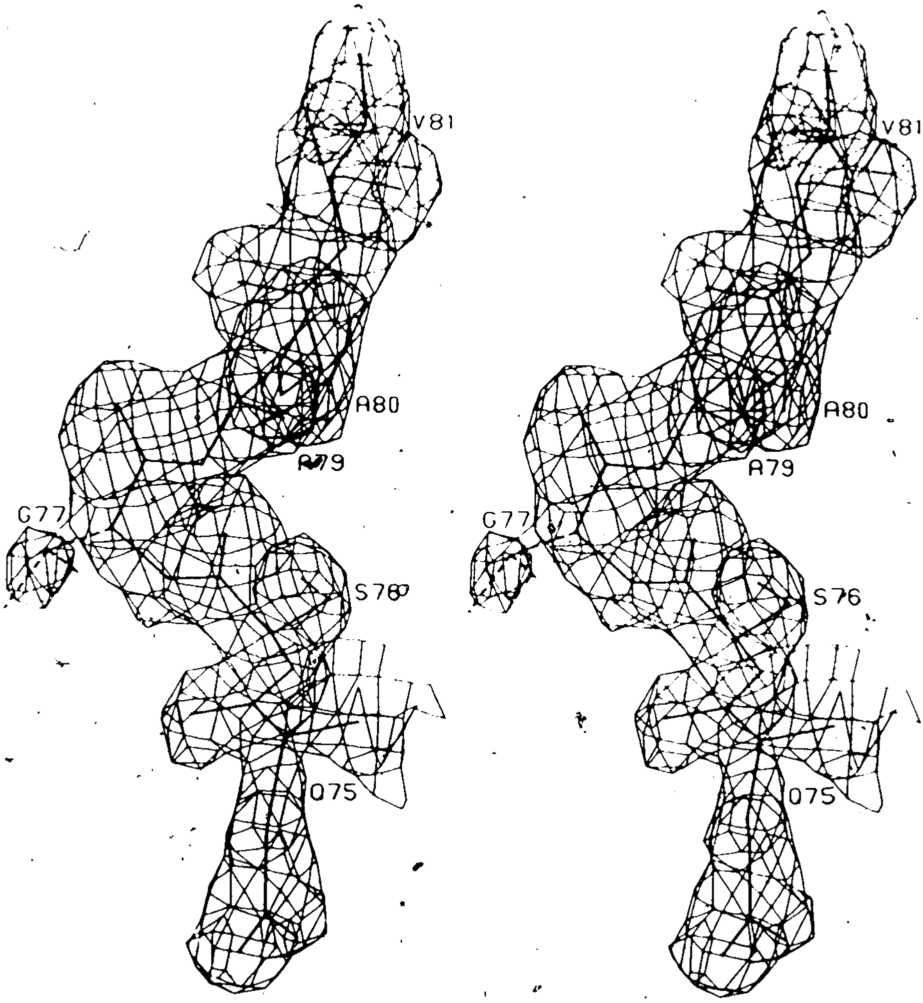


Figure II.4. Electron Density in Region of Sequence Correction. The current interpretation of SGT in the segment Gln75-Val81 is shown in solid lines with electron density from the final map at refinement cycle 119. Dashed lines show the position a threonine side-chain could take at residue 77. The map, computed with Fourier coefficients $[(2m_c|F_o|-D|F_c|)\exp(i\alpha_c)]$ (see Chapter IV), is contoured at a density of $0.30e/Å^3$; contours further than 1.5Å from atoms in the figure are omitted for clarity.

Some residues were left in the model as alanines until very late in the refinement. The last 3 side-chains were added at cycle 100, completing the protein model. Even now, the position of some side-chains is somewhat arbitrary, presumably because of thermal motion or static disorder in their positions. Table II.8 summarizes the residues for which the electron density is not entirely satisfactory. In Table II.8, when the density is referred to as noisy, there are often indications of static disorder about some of the rotatable bonds, but the conformation chosen for the structural model is likely the predominant one. The noise can appear as an extra lump in an alternate direction to which the side-chain could point, but complete density for an alternate conformation is not generally seen. Only for the

Table II.8

Residues with Poor Electron Density in Final Map

Residue	Comment
Thr20	Possible static disorder with χ_1 rotated by -120°.
Arg21	Poor density for guanidinium group.
Thr65	Possible static disorder with χ_1 rotated by +120°.
Lys82	Good density up to N^{ζ}, which is not visible.
Arg84	Noisy density beyond C^{δ}.
Lys87	Very weak density beyond C^{δ}
Thr98	Possible static disorder with χ_1 rotated by +120°.
Gln110	Noisy side-chain density.
Lys122	Weak density beyond C^{γ}
Gln133	Orientation of side-chain amide (i.e., χ_3 angle) difficult to establish from density.
Arg145	Density of side-chain, which projects into solvent, is noisy and ambiguous.
Asn174	Little density for $O^{\delta 1}$ and $N^{\delta 2}$.
Gln192	Density noisy and weak beyond C^{β}
Asn204	Noisy density, possibly due to static disorder about χ_1.
Asp205	Very weak density for C^{β}; carboxyl group density is noisy.
Ser236	Possible static disorder about χ_1 rotation.
Arg243	Density is ve y noisy; side-chain conformation past C^{γ} is essentially arbitrary.

shorter side-chains, such as threonine or serine, is it sometimes possible to assign two alternative positions. For these cases, the refinement might have benefited from having a version of the refinement program that allows for static disorder (Smith *et al.*, 1984). Figure II.5 shows electron density corresponding to a better and a poorer region of the final map. In Figure II.5(b) there is an example of noisy density that can be interpreted as side-chain disorder. An extra lump of density for the side-chain of Asn204 could arise from a conformation in which χ_1 of this residue is approximately -60°.

After cycle 35 of refinement, difference electron density maps [coefficients $(m_C|\mathbf{F}_O|-|\mathbf{F}_C|)\exp(i\alpha_C)$] were examined to determine potential positions of solvent molecules, which were interpreted as water. Once waters were included in the model, they were examined in the maps used to evaluate the protein model, and were deleted if the electron density was weak or if the peaks were not convex. One water molecule, outstanding because indicated parameter shifts would have given it a negative *B*-factor, was reinterpreted as a Ca^{2+} ion at cycle 57. Toward the end of refinement, when it was noted that few added solvent sites were satisfactory in the next electron density map, new solvent sites were added much less liberally. Because water molecules continued to be deleted, the number of accepted water positions went down to 192 from a high of 211.

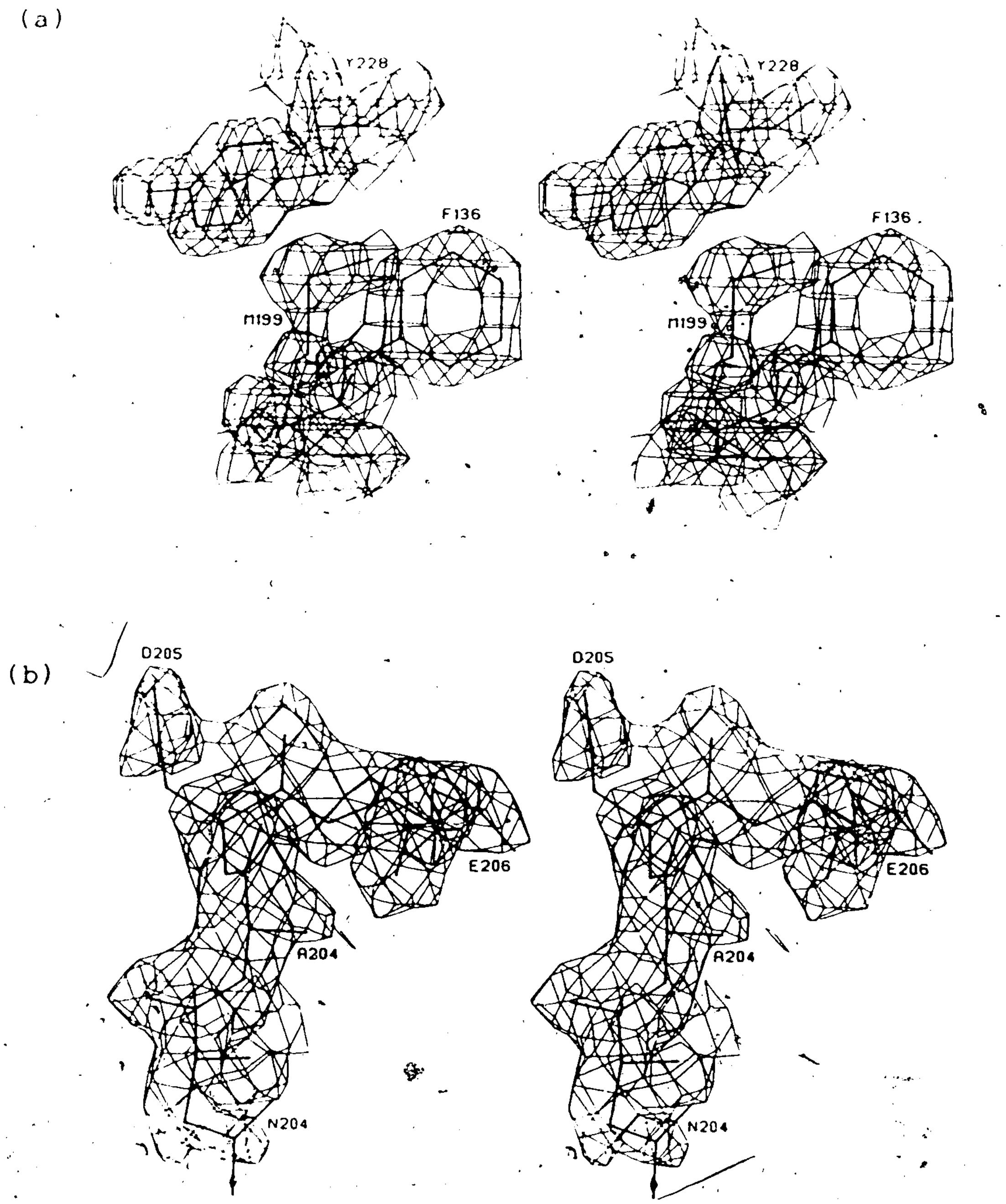


Figure II.5. Good and Bad Regions of Electron Density. Portions of the final electron density map are shown for two parts of the SGT model. For clarity, contours further than 1.5Å from atoms in the figures are omitted. (a)Electron density in this well-ordered part of SGT is contoured at 0.60e/Å^3. (b)Electron density in this more poorly-ordered part of SGT is contoured at 0.35e/Å^3.

In the assignment of water positions, it is difficult to be certain that peaks in the difference map are not just noise. As a result, incorrect positions are sometimes chosen for water molecules. To a certain extent, these are effectively removed from the phasing model when they refine to low occupancies and high *B*-factors. Even when this does not occur, the map coefficients $[(2m_C|F_O|-D|F_C|)(\exp(i\alpha_C)]$ lead to minimal model bias, as discussed in Chapter IV, so that incorrect waters can be detected. Nonetheless, some water positions in the final model are still likely to be in error. Fujinaga *et al*. (1985) removed and then redetermined the solvent structure of α-lytic protease and found that 24 of 153 water molecules did not reappear. These presumably incorrect waters all had low values of an empirical quality factor, defined as occupancy^2/*B* (James and Sielecki, 1983). By ordering the water molecules in decreasing order of quality factor, one can use the sequence number as a guide to relative reliability. Rather than being ordered by quality factor, however, the SGT water molecules are numbered in increasing order of estimated positional standard deviation, calculated as described below. The standard deviation in position is more easily rationalized as a measure of reliability, but in any event the order of water molecules derived using the two measures is very similar. The water molecules with sequence numbers near 1, then, should be very reliable, while those nearer 192 are much less reliable.

Quality of Final Model of SGT

There are two major aspects to consider when evaluating a crystal structure determination. First is the question of whether the structure refinement is complete. In a completed refinement, there should be few indications that further change to the model is necessary, either in the parameter shifts at the final least squares cycle or in discrepancies between the atomic model and the final elecuron density map. There should be a good agreement between the observed and calculated structure factors, but this should not be achieved at the expense of reasonable stereochemistry. The second aspect is the accuracy of the model. The accuracy that can be attained is limited by the accuracy and resolution of the diffraction data, which in turn is limited by the quality of the crystal. Even in a properly refined structure, therefore, there will be coordinate errors. Error estimates can be essential in evaluating mechanistic proposals made on the basis of enzyme models, where the validity of a proposal can depend on distances of the order of 0.1Å.

Very little changed in the model of SGT after cycle 100 of refinement. The few adjustments to the protein model involved atoms with high *B*-factors, and only water molecules with high *B*-factors and low occupancies were added or deleted. The indicated parameter shifts for cycle 119, the final cycle, were very small; the rms shift of coordinate positions was only 0.010Å.

Agreement between observed and calculated structure factors is usually measured by the *R*-factor. Though this measure has little basis in statistical theory, it has the decided advantage of being familiar to crystallographers. On the other hand, crystallographers apply varying criteria to the acceptance of reflections used to calculate the *R*-factor, so some of the benefit of familiarity is illusory. For SGT, the *R*-factor is 0.161 for the 20046 reflections from 8.0-1.7Å having $I > \sigma(I)$. These constitute 80.6% of the measured reflections. Using all of the data (with $|\mathbf{F}_o|$ of reflections having negative net intensities set to zero), the *R*-factor is 0.230.

An alternative measure of agreement is the coefficient of correlation between $|\mathbf{F}_o|$ and $|\mathbf{F}_c|$. This measure is occasionally cited in the crystallographic literature (e.g. Birktoft and Blow, 1972), but much less often than the *R*-factor. Unlike the *R*-factor, the correlation coefficient is not affected by overall scaling errors, though it is affected by errors in overall *B*-factors. In addition, it is a common and well-studied statistic. The correlation coefficient takes the value of zero for unrelated variables and one for linearly related variables. Unfortunately, a correlation on $|\mathbf{F}|$ will be greater than zero even for unrelated structures because the average values of $|\mathbf{F}_o|$ and $|\mathbf{F}_c|$ will vary similarly with resolution. Correlations calculated on $|\mathbf{F}|^2$ are less useful because the effect of the variation with resolution is amplified.‘ For SGT the correlation on $|\mathbf{F}|$, calculated using all measured data, is 0.908.

The *R*-factor and the correlation on $|\mathbf{F}|$ are both simple to compute. But for both of them, though the value expected for a perfect structure is known (0 for the *R*-factor and 1 for the correlation), it is difficult to interpret the value obtained. For the correlation coefficient, normalizing the structure factors would remove the component of the correlation that arises from the variation with resolution, so that a correlation of zero would be expected for an unrelated structure. In addition, a correlation between normalized structure factors would be insensitive to errors in the overall *B*-factor. If one is using normalized variables, however, it is better to calculate the correlation between $|\mathbf{E}_o|^2$ and $|\mathbf{E}_c|^2$. As long as no limits are applied to the magnitude of the accepted structure factors (see Chapter IV; resolution limits can still be applied), this correlation can be interpreted with simple physical models. As noted above and discussed in Chapter IV, the correlation on $|\mathbf{E}|^2$ should vary monotonically with the mean figure of merit of

‘The correlation coefficient calculated with intensities can also be dominated by the alteration in relative scale of $|\mathbf{F}_o|$ and $|\mathbf{F}_c|$ at very low resolution that results from the omission of disordered solvent in the model. For example, the correlation on $|\mathbf{F}|^2$ increases from 0.670 to 0.969 when the 270 reflections at lower than 8.0Å resolution are ignored. At the same time, the correlation on $|\mathbf{F}|$ increases only from 0.908 to 0.955. Note that these considerations do not apply to the molecular replacement work, in which the correlations were calculated over very narrow resolution shells.

the calculated phases. Its square root is an estimate of σ_A (Hauptman, 1982), which is a combined measure of the completeness and the accuracy of a structural model (Srinivasan and Ramachandran, 1965). When there are no coordinate errors, $\sigma_A{}^2$ is equal to Σ_P/Σ_N, the fraction of the total scattering matter contained in the partial structure. Incompleteness of the structural model thus sets an upper bound on the correlation coefficient; deviations from this value reflect the influence of errors in the coordinates and in the measurement of the structure factors. Finally, an overall value of σ_A will be approximately equal to the correlation coefficient between the correct and model E maps (Chapter IV). For SGT, the data were normalized in 25 equal ranges of $(\sin\theta/\lambda)^2$, and the correlation on $|\mathbf{E}|^2$, using all of the data, was 0.863.

The good agreement between observed and calculated structure factors for SGT was not achieved at the expense of model stereochemistry. Since variations in geometry are observed even when comparing similar groups in accurately determined small molecule structures, some deviations from ideal geometry should be expected. The parameters listed in Table II.9 demonstrate that the deviations from ideal geometry found in SGT are in the range of those found in small molecule structures (Marsh and Donohue, 1967; James *et al.*, 1980). Another indication of the quality of the model geometry is the ϕ-ψ plot (Ramakrishnan and Ramachandran, 1965) shown in Figure II.6. No non-glycine residues have ϕ-ψ

Table II.9

Final Refinement Parameters and Results

No. of protein atoms[1]	1621
No. of solvent atoms	192
No. of variable parameters[2]	7442
Rms deviations from ideal values[3]	
distance restraints(Å)	
bond distance	0.019(0.014)
angle distance	0.038(0.027)
planar 1-4 distance	0.041(0.027)
plane restraint(Å)	0.017(0.016)
chiral-center restraint($Å^3$)	0.208(0.130)
non-bonded contact restraints(Å)	
single torsion contact	0.279(0.350)
multiple torsion contact	0.130(0.350)
possible hydrogen bond	0.177(0.350)
conformational torsion angle restraint(°)	
planar(ω)	3.1(2.5)
isotropic thermal factor restraints($Å^2$)	
main-chain bond	1.816(1.500)
main-chain angle	2.535(2.000)
side-chain bond	5.729(3.500)
side-chain angle	8.466(5.000)

[1] Including Ca^{2+} ion.
[2] Positional parameters for the solvent molecule O18, which lies on a crystallographic 2-fold axis, were not varied.
[3] The values of σ, in parentheses, are the input estimated standard deviations that determine the relative weights of the corresponding restraints [see Hendrickson and Konnert (1980)].

values significantly outside their allowed conformational regions.

There is no entirely satisfactory method for estimating coordinate errors in structures refined by restrained-parameter least-squares refinement. Nonetheless, the several available methods give results that agree reasonably well among themselves (Read *et al.*, 1983), as well as agreeing roughly with the size of the coordinate differences that are observed when the same structure is determined twice

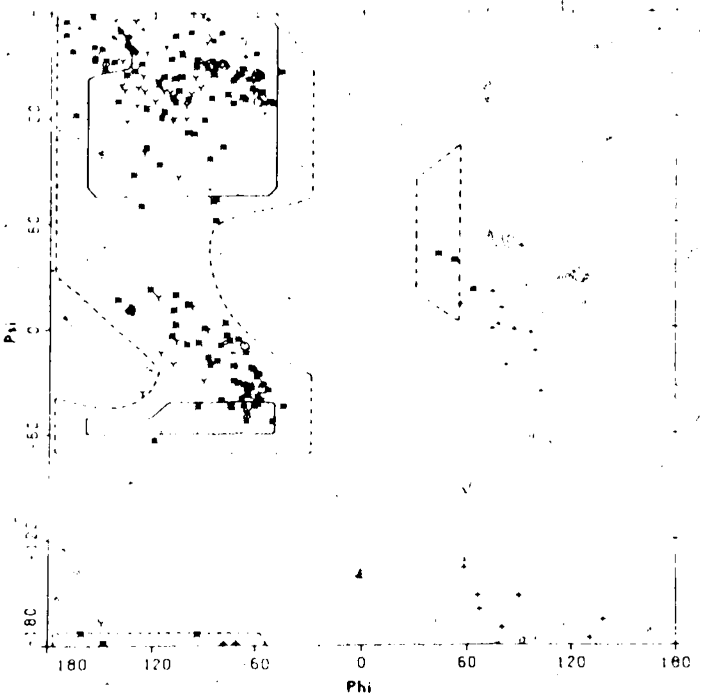


Figure II.6. ϕ-ψ Plot of Main-chain Conformational Angles. The symbols correspond to the following residue types: (O) proline; (Y) β-branched amino acids; (+) glycine; (*) others. The continuous lines show the fully allowed conformational regions for $\tau(C^{\alpha})=110°$, and the broken lines show the regions obtained by relaxing the van der Waals' contact constraints and setting $\tau(C^{\alpha})=115°$ (Ramakrishnan and Ramachandran, 1965).

(Chambers and Stroud, 1979; but see their comments on Cruickshank's method summarized below).

An estimate of the mean coordinate error is useful for comparing the overall accuracy of protein structures, though it is not very useful when one is interested in the accuracy of a particular atom. Overall coordinate error has commonly been estimated using the Luzzati plot (Luzzati, 1952). The σ_A plot, an alternate method built on the same theoretical foundation, is developed in Chapter IV. As discussed there, the σ_A plot shares some limitations with the Luzzati plot, but circumvents some of the problems in the application of that method. A σ_A plot for SGT is shown in Figure II.7. It is interesting to note that the rms coordinate error deduced from this plot, 0.244Å, is slightly larger than that obtained at cycle 78, 0.231Å, even though the model has doubtless improved. Also, the intercept in the σ_A plot has increased from zero, the maximum theoretical value, to a small positive value. Both effects can be explained by noting that the relative weight on low and high resolution data was

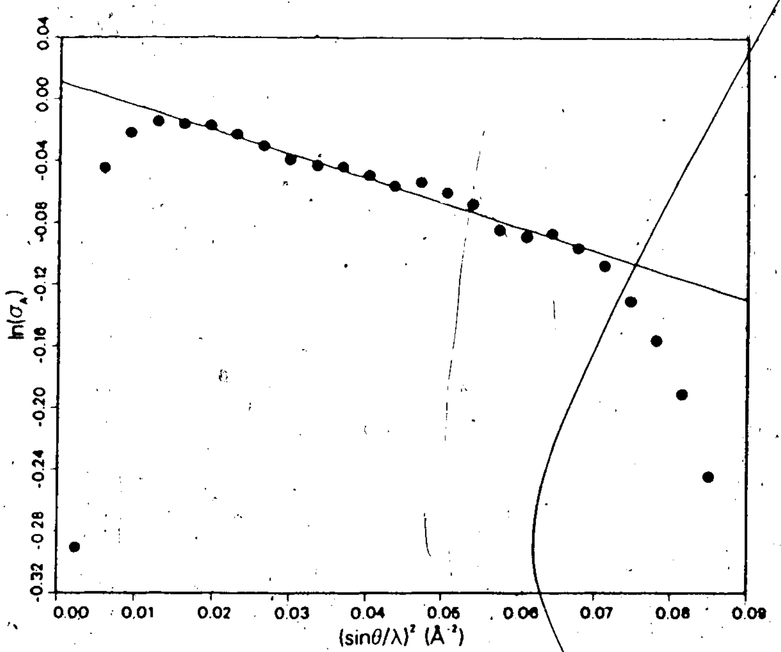


Figure II.7. Overall Coordinate Error by σ_A Plot. The values of σ_A, represented by circles, were determined as discussed in Chapter IV. The line is a least squares fit to all the points excluding the first 3 and the last 4. From the intercept (0.0112), (Σ_P/Σ_N)=1.023, and from the slope (-1.568Å^2), the rms coordinate error is 0.244Å.

changed between cycles 78 and 119. The structure factor terms are weighted in refinement using an input standard deviation given by the expression $[C_1+C_2(\sin\theta/\lambda-1/6)]$. Between cycles 78 and 119 of refinement, C_1 was changed from 35.0 to 28.0 and C_2 from -150.0 to -50.0. Because of the increased relative weight on the agreement of lower resolution reflections, refinement increased σ_A more for the lower resolution ranges than for the higher resolution ranges. This demonstrates that it is possible to change the indicated overall coordinate error by changing the structure factor weight as a function of resolution.

Estimates of coordinate error for individual atoms can be obtained from a modification of the formula given by Cruickshank (1949, 1967). For centered space groups, the expression for coordinate error must be divided by the number of centering translations. This can be deduced by considering that the estimated coordinate error should be invariant with respect to allowed choices of space group lattice. For example, if one transformed a structure from space group P1 to C1, the error obtained using an unmodified form of Cruickshank's (1967) equation would double. With that in mind, for SGT the estimated standard deviation in the x coordinate for atom i is given by

$$\sigma(x_i) = \frac{a}{4\pi} \times \frac{[\Sigma h^2(|\mathbf{F}_o|-|\mathbf{F}_c|)^2]^{1/2}}{\Sigma(m/2)h^2 f_{0i}\exp[-B_i(\sin\theta/\lambda)^2]}$$

where a is the axial length, f_{0i} is the atomic scattering factor and m=2 or 1, depending on whether the reflection is centric or non-centric, respectively. The sums must be taken over the limiting sphere of the reciprocal lattice or, alternatively, corrections must be made for reflection multiplicity. This formula gives the standard deviation as a function of atom type, *B*-factor and, implicitly, occupancy. (The occupancy is a constant in the atomic scattering factor, hence a constant in the denominator of the expression.) The expressions for the error in the y and z coordinates are defined analogously. When $\sigma(x_i)=\sigma(y_i)=\sigma(z_i)$, which is very nearly true for SGT, the radial error in atomic position is given by $\sigma(r_i)=\sqrt{3}\sigma(x_i)$. In this work, the almost equivalent expression $\sigma(r_i)=[\sigma(x_i)^2+\sigma(y_i)^2+\sigma(z_i)^2]^{1/2}$ was used. Estimated radial errors as a function of *B*-factor are given in Figure II.8 for carbon, nitrogen, oxygen and sulfur atoms. Figure II.9, which shows the variation in *B*-factor along the polypeptide chain, can be used in conjunction with Figure II.8 to gain an impression of the relative accuracy of different parts of the structure.

To obtain estimates of coordinate errors for water molecules from this plot, the standard deviations must be divided by the water occupancies. A problem arises in the case of the solvent O18, which lies in a special position on a crystallographic 2-fold axis parallel to the y axis. The occupancy (0.5) of this molecule must be doubled to find the appropriate error in the y coordinate; because of symmetry constraints there will be no error in the x and z coordinates.

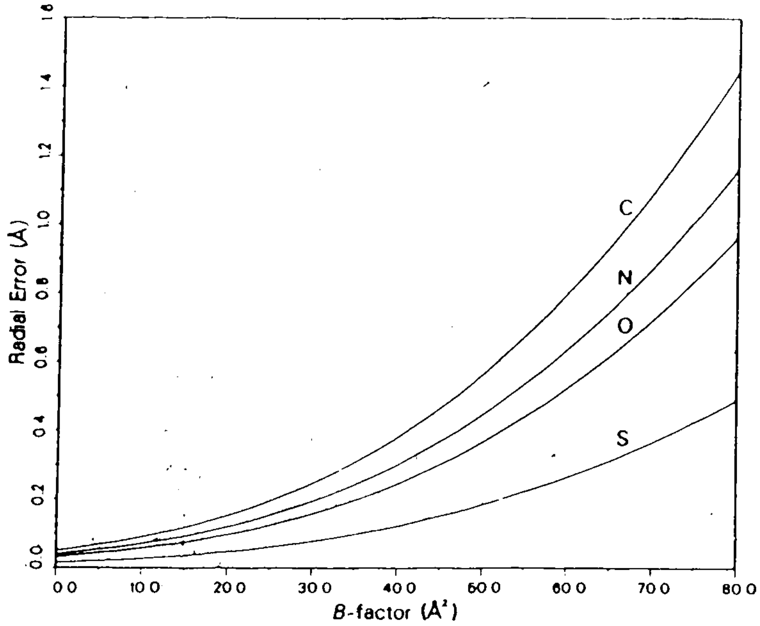


Figure II.8. Atomic Coordinate Error by Method of Cruickshank. The estimated radial standard deviations in atomic position are given as a function of B-factor. The 4 curves, from top to bottom, are for carbon, nitrogen, oxygen and sulfur atoms in the final refined structure of SGT. No curve is shown for Ca^{2+}; the single Ca^{2+} ion has an estimated standard deviation of 0.025Å.

An overall estimate of the coordinate error can be obtained from the errors for individual atoms by computing the rms value of the estimated radial standard deviations for all of the atoms that make up the SGT structure. The value of the rms coordinate error, 0.245Å, agrees remarkably well with the value obtained from the σ_A plot, though this close agreement may in part be fortuitous. Chambers and Stroud (1979), in contrast, found that the errors they obtained by the method of Cruickshank (1967) were only about half as large as those obtained either by the method of Luzzati (1952) or by comparing their structure of BT with that of

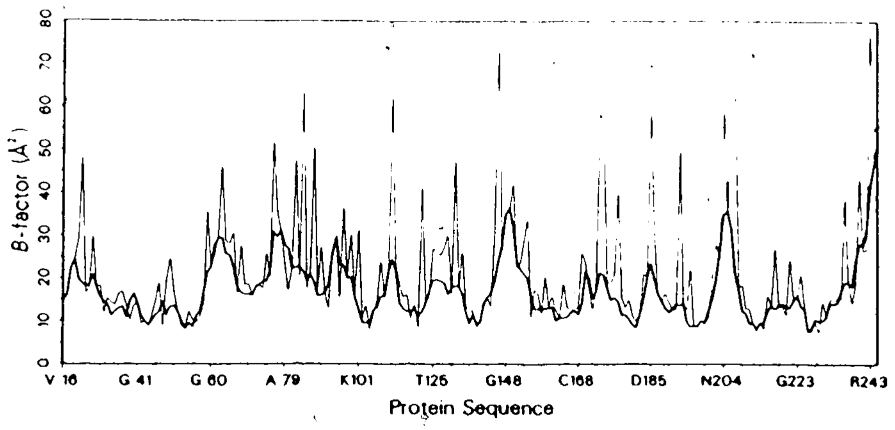


Figure II.9. Variation in *B*-factor along the Polypeptide Chain. The thick lines show the mean *B* of the main-chain atoms while the thin lines show the mean *B* of the side-chain atoms. Every 20th amino acid in the sequence is labelled.

Bode and Schwager (1975).

With an estimated rms coordinate error of 0.24Å, the structure of SGT is not quite as accurate as some comparable structures for which similar analyses of error have been performed. For example, the complex between SGPB and OMTKY3, refined at 1.8Å resolution to an *R*-factor of 0.145 on 90% of the data, has an estimated rms coordinate error of 0.14Å (Read *et al*., 1983). For α-lytic protease (1.7Å resolution; *R*-factor of 0.131 on 66% of the data) the estimated error is about 0.12Å (Fujinaga *et al*., 1985). The larger estimated error for SGT can probably be attributed to the greater amount of thermal motion in this structure. For all of the atoms in the structure, the mean *B*-factor is 21.3Å^2. Including only protein atoms, the mean *B*-factor is 20.0Å^2. The relatively high degree of thermal motion will limit the accuracy that can be attained with this crystal.

In summary, the structure of SGT has been solved at 1.7Å resolution by a combination of molecular replacement and MIR. It has been refined to convergence, and the resulting structure agrees well with the observed structure factors, while acceptable geometry has been maintained. The estimated accuracy of the coordinates, while not the best that has been achieved for structures refined at similar resolution, is sufficient that the structure of SGT can legitimately be used in detailed studies of the structure, function and evolution of the serine proteinases.

B. The Structure of SGT

SGT is very similar in structure to BT, CHT and other mammalian serine proteinases. The structures of these serine proteinases, especially BT, are extremely well-studied. A structural comparison with BT, which will allow one to relate to SGT the wealth of data on this protein family, will be of more use than a detailed description of SGT alone. Therefore, after a short description of some aspects of the crystal structure of SGT, the discussion will concentrate on a comparison with the structure of BT.

Two different views of all of the atoms comprising SGT are given in Figure II.10. These representations give a good impression of the overall molecular structure. In addition, they provide a structural context for the examination of isolated details.

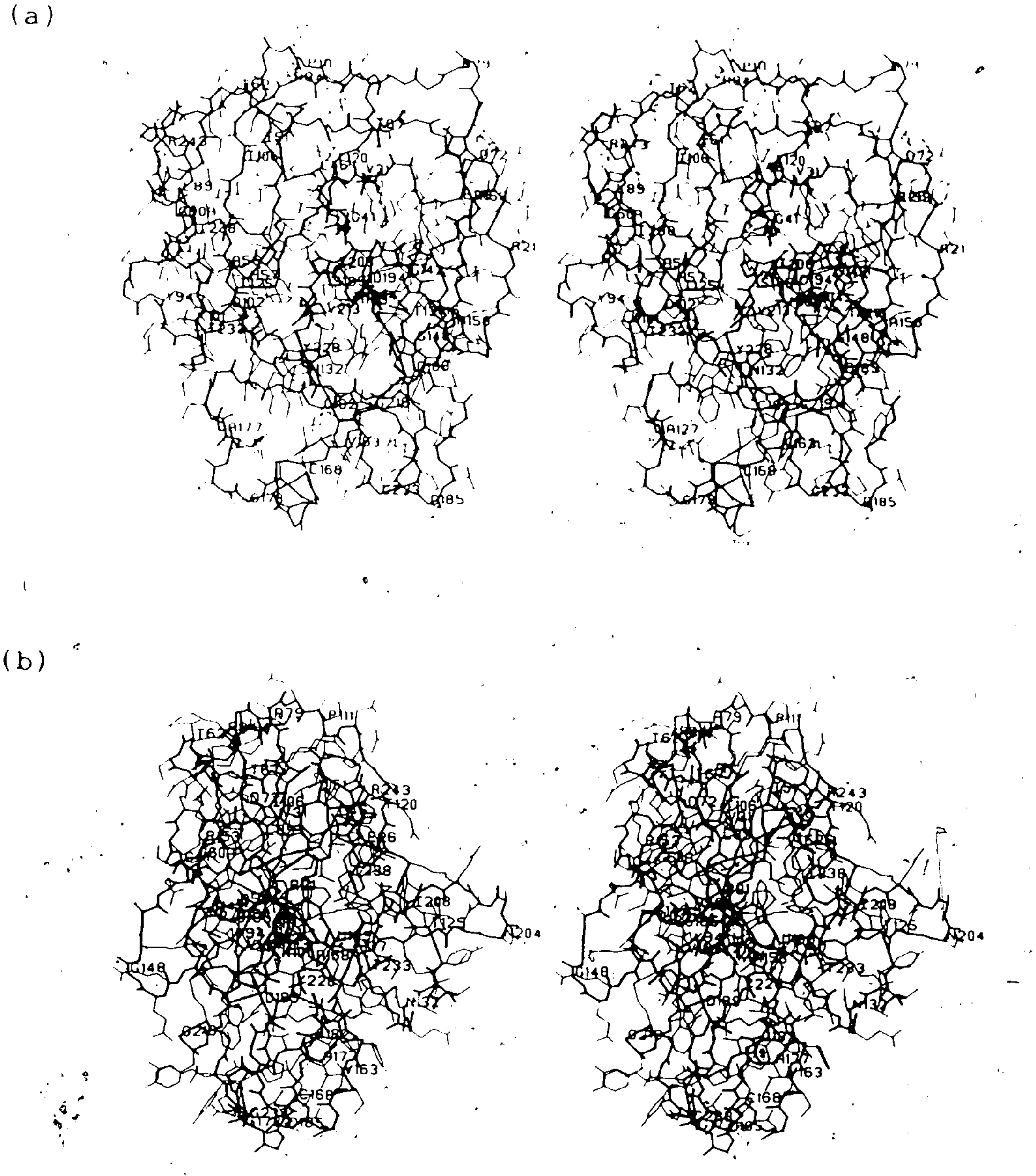


Figure II.10. Two Views of SGT Showing All Atoms. In both views, the main-chain is shown in thick lines and the side-chains in thin lines. Every fifth amino acid residue is labelled. (a) SGT in the "standard" orientation with the active site at the front. Similar orientations are used for most figures showing the entire structure. (b) SGT rotated by 90° from the orientation shown in (a), so that the active site is to the left.

Secondary Structure

A more schematic view of SGT is given in Figure II.11. This representation of the main-chain atoms and their hydrogen bonds gives an impression of the secondary structure that provides the framework for SGT.

The hydrogen bonds were assigned using the electrostatic criteria proposed by Kabsch and Sander (1983) and, as those authors note, should more properly be called polar interactions. Lifson *et al.* (1979) argue that the major features of hydrogen bonds are consistent with a simple electrostatic model, and that charge transfer effects are minor. On this basis, the energy of interaction between dipoles is a reasonable measure of the strength of a hydrogen bond.

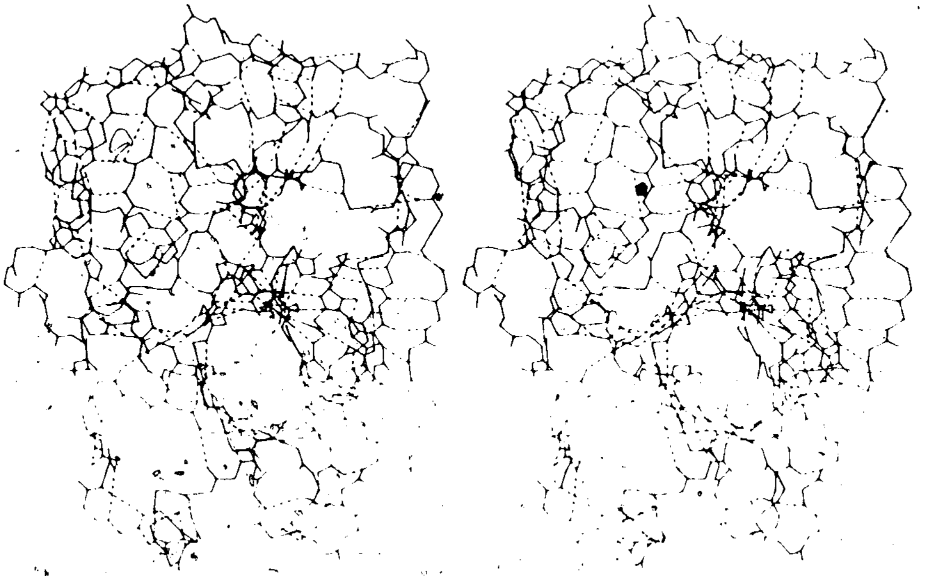


Figure II.11. Main-chain Hydrogen Bonds. Solid lines represent main-chain atoms of SGT, and dashed lines represent hydrogen bonds determined by the criteria of Kabsch and Sander (1983).

Not all of the interactions in Figure II.1 accord with the traditional picture of a hydrogen bond. For example, a number of interactions are of the type $(C=O)_i \cdot\cdot (H-N)_{i+2}$ in segments having the conformation termed C_7 by Avignon *et al.* (1969). These are found mostly in strands of β-sheet (Kabsch and Sander, 1983) and have calculated interaction energies of the order of -0.5 to -1.0 kcal/mole, which is weak but similar to that found for the 3_{10}-helical interactions. However, one must be aware that the relative strength of different interactions can be assessed only roughly due to the crudeness of the model of the polar interaction (J. Moult, personal communication). The two dipoles are approximated by partial charges at the positions of the N, H, C and O atoms, other atoms with partial charges (such as C^α) are ignored, and no allowance is made for polarization. Nonetheless, even though the model of the hydrogen bond used by Kabsch and Sander (1983) is not rigorous, the single energy criterion they suggest provides a more objective method for assigning hydrogen bonds than the customary *ad hoc* set of geometrical criteria.

Elements of secondary structure in SGT were determined using the program DSSP (Kabsch and Sander, 1983), and are summarized in Table II.10. As is evident from this table, the predominant secondary structure in SGT is β-sheet. Apart from some short stretches of α- and 3_{10}-helix, the only major helical secondary structural element is the characteristic C-terminal helix that is observed even in the

Table II.10

Secondary Structural Elements in SGT

Segment	Secondary structure'	Segment	Secondary structure'
Thr20-Arg21	β(A)	Thr135-Gly140	β(J)
Met30-Leu33	β(B)	Leu156-Val163	β(K)
Cys42-Ala48	β(C)	Asp165-Tyr172	α
Ile51-Thr54	β(D)	Gly173-Glu175	3_{10}
Ala56-Cys58	3_{10}	Glu180-Ala183	β(L)
Gly60C-Asn60D	β(E)	Pro198-Lys202	β(M)
Thr65-Gly68	β(F)	Trp207-Trp215	β(N)
Val81-Gln90	β(G)	Gly226-Glu230	β(O)
Ala104-Leu108	β(H)	Val231-Arg243	α
Lys122-Ile123	β(I)	Ala242-Thr244	3_{10}

'Elements of secondary structure are indicated by α for α-helix, β-for strand of β-sheet (letters in parentheses give the strand designation), and 3_{10} for 3_{10}-helix.

structurally less-related bacterial serine proteinases SGPA, SGPB and α-lytic protease (James *et al.*, 1978; Fujinaga *et al.*, 1985).

Like CHT (Birktoft *et al.*, 1970; Birktoft and Blow, 1972) and elastase (Shotton and Watson, 1970), SGT can be divided into two domains of similar topology. This internal symmetry has been taken as evidence for a gene duplication event in the evolution of the serine proteinases (Birktoft and Blow, 1972; McLachlan, 1979). Each domain is organized around a folding unit that was originally interpreted as a 6-stranded cylinder of antiparallel β-structure, termed a Greek key β-barrel by Richardson (1981). Chothia and Janin (1982) have reinterpreted the cylinder as a pair of orthogonally packed β-sheets, joined at 2 corners by bent strands. In the N-terminal domain of SGT, this folding unit is constructed from the antiparallel β-sheet B·C·D·H·G·F; bends in strands C and G allow an antiparallel interaction between strands B and F. The C-terminal domain is similarly constructed from the sheet J·K·L·O·N·M. Additional β-strands interact at the edges of the orthogonally packed sheets: the short strand E forms two antiparallel hydrogen bonds to strand G, strand A is antiparallel to strand K, and strand I pairs with strand N in the only parallel β interaction in the protein.

Water Structure

The crystal structure of SGT consists not only of protein atoms, but also of solvent. Figure II.12 shows the water molecules that surround SGT in the final model of the crystal structure. Though a detailed examination of the water structure will not be attempted here, a few observations can be made. A number of water molecules occupy internal positions in the protein structure. These water molecules are generally very well-ordered and hence have low sequence numbers in the structural model. The solvent coverage of the protein surface is fairly even, with a few exceptions. Some areas of the protein surface are involved in crystal packing contacts and thus are not accessible to solvent. This explains, for example, the lack of water molecules in the upper area of the active site (Figure II.12); a crystal packing contact involves the following residues in

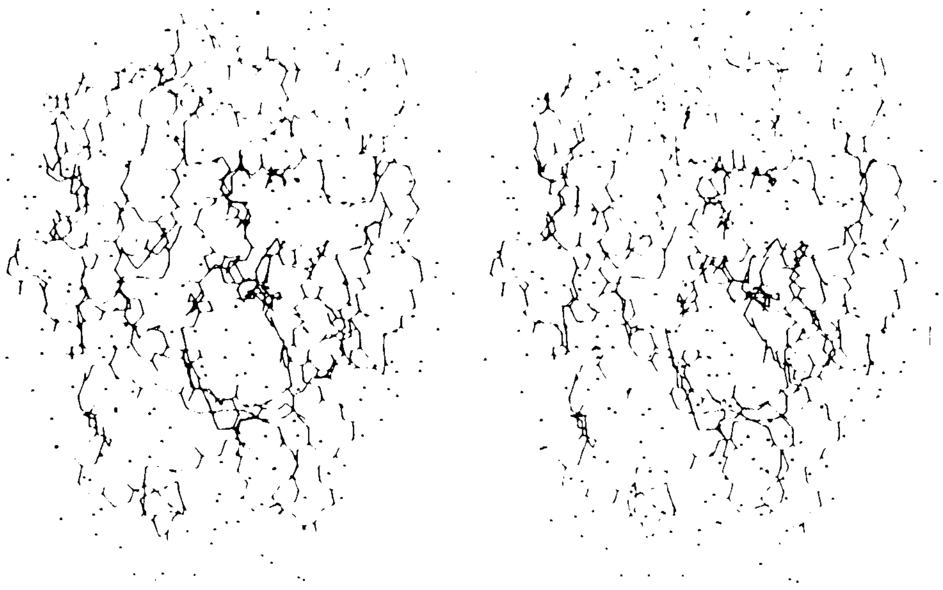


Figure II.12. Water Structure Around SGT. Water molecules are indicated by small circles. For clarity, only the main-chain atoms of SGT are shown. The water molecules in this figure include all of the solvent atoms in the model of the SGT crystal structure, plus all symmetry-related solvent atoms that lie within 4Å of any protein atom.

this region: Met35, Cys42, His57 through Ser60, and Gln192. (This packing contact at the active site might frustrate attempts to soak substrates or inhibitors into crystals of SGT.) In other cases, the lack of observed solvent molecules can be attributed to the high thermal motion of the associated region of the protein. For example, the surface loop Lys202-Glu206 has some of the highest *B*-factors in the molecule, which is understandable considering the extent to which it projects from the protein surface [seen most clearly in Figure II.10(b)], as well as the fact that it is involved in no crystal packing contacts. Any water molecules associated with this loop must have sufficient thermal motion to render them invisible in the electron density.

Calcium Binding Site

A number of serine proteinases possess Ca^{2+} ion binding sites that confer stability against thermal or chemical denaturation, or proteolytic degradation [see reviews by Kretsinger (1976) and Martin (1984), and references therein]. Kretsinger (1976) has suggested that the requirement for calcium ensures that these enzymes will only be active extracellularly, since intracellular calcium concentrations are extremely low.

Among the serine proteinases with calcium sites are BT and SGT. The site of calcium binding in BT has been identified (Bode and Schwager, 1975) and involves the residues Glu70 and Glu80. Calcium binding inhibits autolysis in BT (Bier and Nord, 1951), as well as in SGT (Russin *et al*., 1974; Olafson and Smillie, 1975). However, the two glutamate residues at positions 70 and 80 of BT are not conserved in SGT, so its binding site must be located elsewhere. Jurášek *et al*. (1976) suggested that the site could be found in the cluster of acidic residues Asp203, Asp205 and Glu206.

Using lanthanide ions as probes in NMR (Abbott *et al*., 1975) and fluorescence energy-transfer (Darnall *et al*., 1976) experiments, Darnall and co-workers concluded that the Ca^{2+} binding site in BT in solution was not the same as that found in the crystal structure. These spectroscopic experiments allow the estimation of distances from the lanthanide ions to atoms of an inhibitor bound in the active site. On the basis of these distances, as well as sequence alignments with several serine proteinases including SGT, it was suggested that the binding site is located between Asp194 and Ser190 (Thr190 in SGT). Epstein *et al.* (1977), in turn, used several spectroscopic techniques to show that there are two lanthanide binding sites on BT. They argued that the lower affinity site, which does not bind Ca^{2+}, is the one observed by Darnall and co-workers, and that the higher affinity site is the one observed crystallographically.

The Ca^{2+} ion site found in the crystal structure of SGT is not the one suggested by Darnall and co-workers (Abbott *et al.*, 1975; Darnall *et al.*, 1976), which casts further doubt on their proposal. Neither is it located in the cluster of acidic residues noted by Jurášek *et al.* (1976). The site found in the crystal structure, shown in Figure II.13, is composed of the side-chains of Asp165 (bidentate coordination), and Glu230 (unidentate), the main-chain carbonyl groups of Ala177A and Glu180, and the two well-ordered water molecules O13 and O15. Judging from the survey of crystal structures of calcium complexes carried out by Einspahr and Bugg (1980, 1981, 1984), the calcium site in SGT is quite typical, both in coordination number (7) and in geometry. As noted by Einspahr and Bugg (1981, 1984) for other structures, the Ca^{2+} ion is near to the planes of the carboxyl and carbonyl groups that make up the protein ligands (Figure II.13). The coordination bond lengths and Ca-O-C angles,

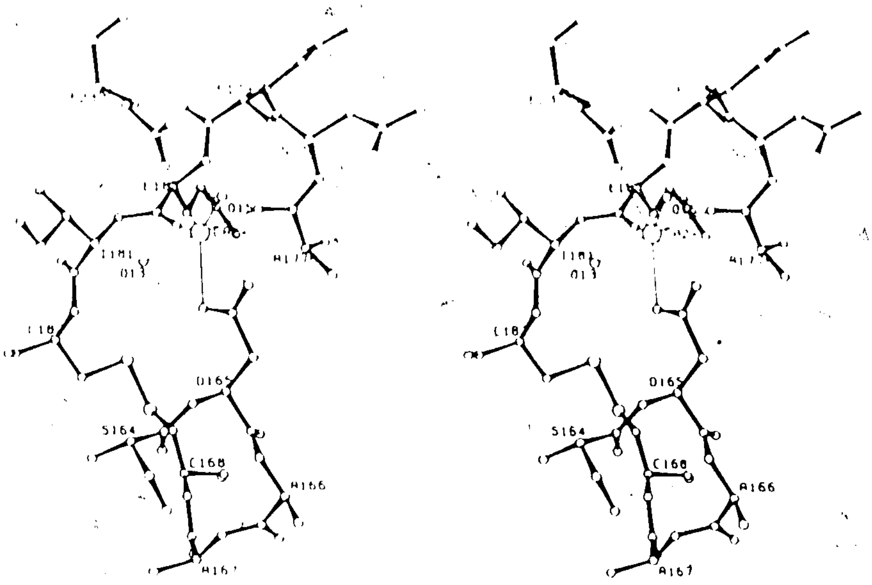


Figure II.13. Coordination Sphere of Ca^{2+} in SGT. Residues of SGT involved in calcium binding are connected by thick bonds, water molecules are shown as small circles and coordination bonds are indicated by thin lines. A number of ordered water molecules are found in this region of the structure but, for clarity, only those directly coordinated to the calcium ion are shown. This figure, and Figure II.14, were prepared with the program PLUTO, written by W. D. S. Motherwell.

summarized in Table II.11, are within the ranges observed in small molecule crystal structures that have similar interactions. [The distance of 2.29Å to O13 is not significantly lower than the minimum water coordination distance of 2.3Å observed by Einspahr and Bugg (1980).] In all proteins examined, Einspahr and Bugg (1984) found that most of the ligands in a calcium binding site come from a stretch of no more than 12 residues in the sequence. SGT is clearly an exception to this rule. However, because of the disulfide bridge 168-182, all of the protein ligands but one are connected by a continuous chain that is similar in length to a

Table II.11

Geometry of Calcium Binding Site

Oxygen atom		Coordination bond length(Å)	Ca-O-C angle(°)
Asp165	$O^{\delta 1}$	2.52	88.6
Asp165	$O^{\delta 2}$	2.43	92.1
Ala177A	O	2.24	159.5
Glu180	O	2.26	159.1
Glu230	$O^{\epsilon 2}$	2.43	140.4
O13	O	2.29	—
O15	O	2.41	—

sequence of 12 residues.

Comparison with BT

SGT and BT display significant homology in sequence and structure (summarized in Table II.1), in addition to similar substrate specificity. An overall structural comparison is shown in Figure II.14, from which it can be seen that the basic structural framework is quite highly conserved, though some of the surface loops differ markedly in length and conformation.

A space-filling colour representation of SGT and BT is shown in Figure II.15. In this representation, the differences in surface features tend to dominate the visual impression, and the relationship between the proteins is thus somewhat obscured. However, close examination reveals the similarity between these molecules, especially in the region around the active site.

In comparing related proteins, one expects that the parts most highly conserved in sequence and structure will

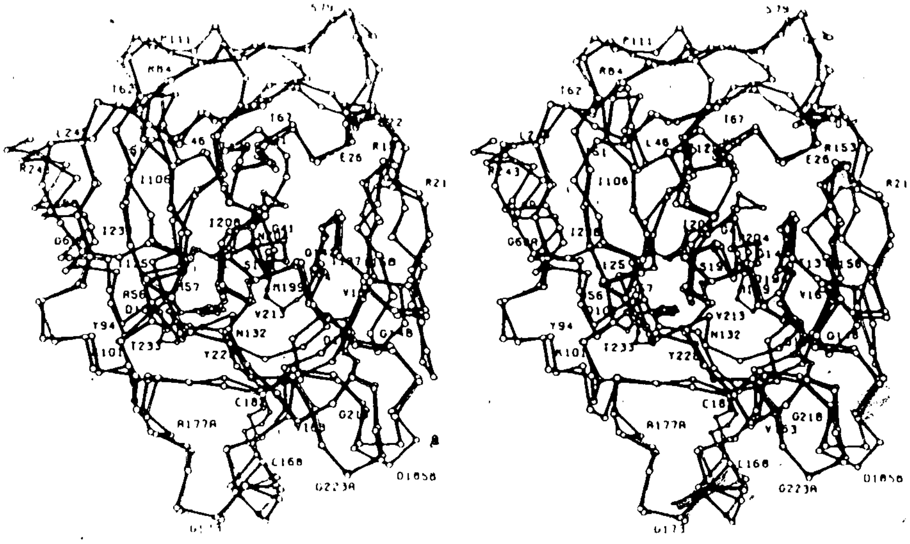


Figure II.14. C^{α}-atom Representation of Superposed SGT and BT. Filled bonds denote SGT and open bonds denote BT. Every fifth amino acid residue in SGT is labelled with the residue type and sequence number. Disulfide bridges and side-chain atoms of His57, Asp102 and Ser195 of both proteins are also shown.

include the hydrophobic core of the protein. To a certain extent, that is true in the case of SGT and BT. Figure II.16 shows the residues of SGT that are identical in sequence and structurally homologous to residues in BT. These conserved residues are indeed more likely to be found in the hydrophobic cores of the two domains than on the protein surface. However, conserved residues are much more heavily concentrated in the active site and in the parts that are involved in the binding of substrate, particularly the primary specificity pocket or S_1 subsite (slightly to the right of and below the active site in the standard orientation used for Figure II.16). These residues are of direct

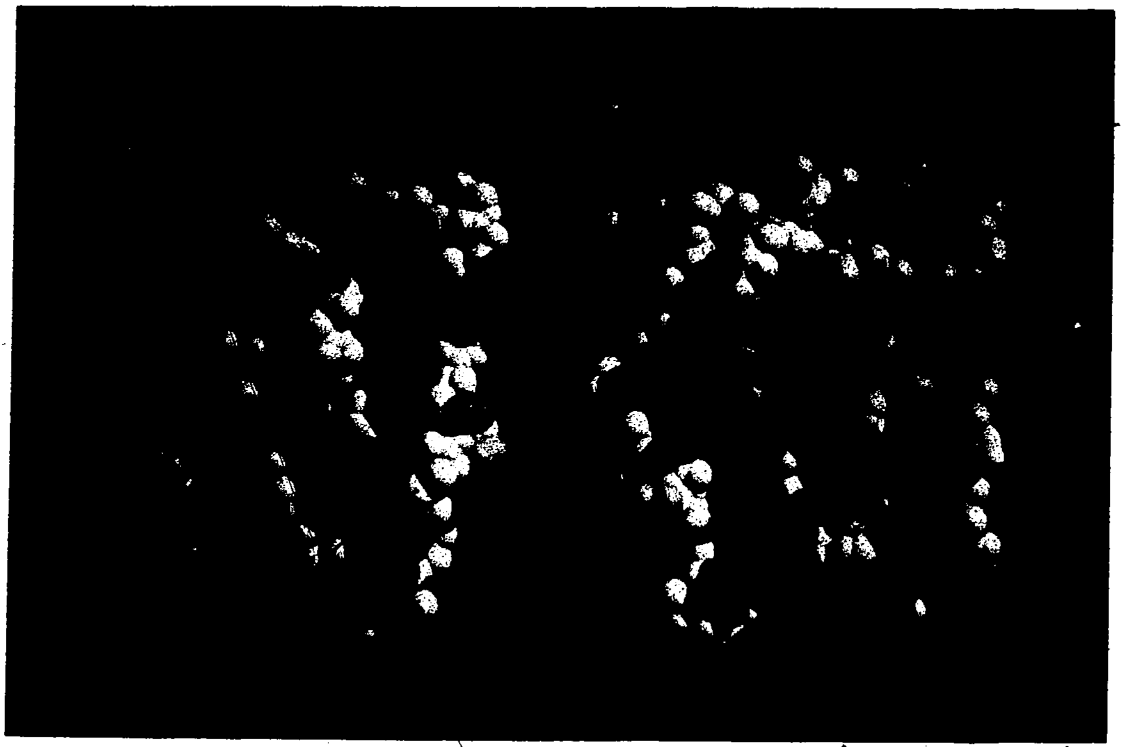


Figure II.15. Space-filling Representation of SGT and BT. SGT is on the left and BT is on the right, both in the "standard" orientation of Figure II.10(a). Gly and main-chain are shown in off-white, and side-chains are coloured according to residue type as follows: dark gray for Pro; increasingly deeper shades of green for Ala, Val, Leu and Ile; shades of brown for Tyr, Phe and Trp; shades of yellow for Cys and Met; shades of pink for Asn and Gln; shades of orange for Ser and Thr; shades of red for Asp and Glu; shades of blue for His, Lys and Arg. This figure was generated using a computer program written by David Bacon.

importance to the function of the protein, i.e., recognizing side-chains of Lys and Arg, positioning a substrate correctly and catalyzing peptide bond hydrolysis. Presumably there is a greater constraint on these parts of the protein than on the parts that are involved only in stabilizing its structure.

To compare the substrate binding regions of SGT and BT, it is useful to have a model of a substrate bound to the

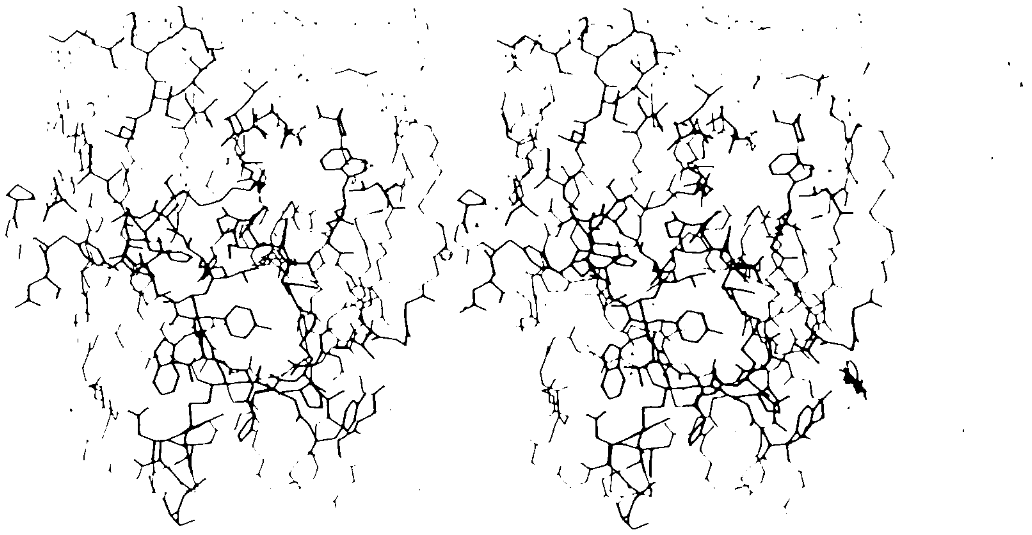


Figure II.16. Residues Conserved in Sequence and Structure Between SGT and BT. Thin lines show the main-chain of SGT, and thick lines show the 70 residues that are identical in the sequence alignment and homologous in the structural alignment with BT. These are the residues outlined by solid boxes in Table II.1.

active site. The protein proteinase inhibitor PTI, in its complex with BT (Rühlmann *et al.*, 1973; Huber *et al.*, 1974; Marquart *et al.*, 1983), provides a good model, especially since it also has been found to inhibit SGT (Trop and Birk, 1968). This complex was superimposed on the structure of SGT to produce a model of the interaction between SGT and PTI. For the superposition, residues in the active site and substrate-binding region were chosen (residues 55-58, 101-103, 189-198, 213-220 and 225-228) so that the modelled interaction would be as similar as possible to the structurally observed enzyme:inhibitor interaction. The rms coordinate difference between the 112 main-chain atoms in these residues was 0.31Å after the least-squares superposition. [For the same comparison with the native BT model of Chambers and Stroud (1979), the rms difference is 0.27Å.] The result is illustrated in Figure II.17. Based on this model, the interactions near to the scissile bond will be very similar to those of BT. On the P_n' side of the scissile bond, the first significant differences occur at S_2', where the loop from Glu146 to Arg153 is one residue shorter and lies closer to the active site in SGT. Binding of PTI to SGT would require shifts of at least the side-chains of Gln151 and Arg17I (P_2'). Interactions further toward the

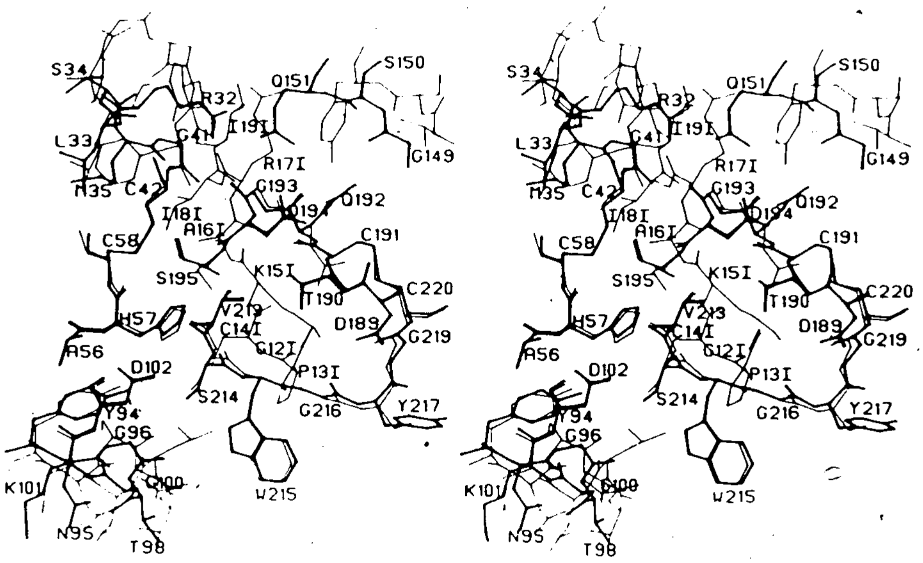


Figure II.17. Modelling a Complex of SGT with PTI. Thick lines indicate residues in the active site and substrate binding region of SGT. Thin lines show the comparable residues in the structure of BT observed in its complex with PTI (Huber *et al.*, 1974; Marquart *et al.*, 1983) as well as residues of PTI near the scissile bond (Lys15I-Ala16I). Residues of SGT and of PTI are labelled; an I after the sequence number indicates an inhibitor residue.

C-terminus of the substrate could also be affected by the conformational differences in the loop from Arg32 to Gly41. On the P_n side, interactions further than P_2 toward the N-terminus could be affected by the altered conformation of the loop from Tyr94 to Lys101; the contacts between PTI and Leu99 of BT would not occur in a complex with SGT. The effect of these conformational differences can also be seen in the space-filling pictures of Figure II.15. On the S_n side, the substrate-binding cleft is more open in SGT than in BT, while it is less open in SGT on the S_n' side.

The general similarity of the substrate-binding regions of SGT and BT, especially near the scissile bond, is consistent with the similarity of their interactions with substrates and inhibitors. SGT cleaves the oxidized B chain of insulin in the same manner as BT (Jurášek *et al.*, 1969; Olafson and Smillie, 1975), as well as the synthetic substrate N-α-benzoyl-L-arginine ethyl ester (Wählby, 1968; Olafson and Smillie, 1975). It can be purified on the same affinity columns as BT [tryptic digest of salmine (Yokosawa *et al.*, 1976), double-headed protein proteinase inhibitor from kidney bean (Mosolov *et al.*, 1978), glycylglycylarginal (Nishikata *et al.*, 1981), and rice bran trypsin inhibitor (Tashiro *et al.*, 1981)]. Finally, it is inhibited by the same protein proteinase inhibitors [soybean trypsin inhibitor and PTI (Trop and Birk, 1968), chicken ovomucoid (Trop and Birk, 1968; Nagata and Yoshida, 1983), and pancreatic secretory trypsin inhibitor (unpublished results)].

SGT and BT are not, however, identical in their susceptibility to inhibitors. Nagata and Yoshida (1983) tested two parts of the ovomucoid inhibitor from Japanese quail, which has three homologous inhibitory domains. The first part contains the domains I and II, and the second part contains domain III. Both strongly inhibit BT, but neither inhibits SGT.

In Figure II.17, it is apparent that the relative disposition of the side-chains of His57 and Ser195 is different in SGT and in BT complexed with PTI. The interaction between His57 $N^{\epsilon 2}$ and Ser195 O^{γ} has been a matter of some controversy. Matthews *et al.* (1977) point out that, in serine proteinases of both the chymotrypsin and subtilisin families, the relative disposition of the side-chains of the active site histidine and serine is consistent with only a weak hydrogen bond. Huber and Bode (1978) suggest that, though the interaction between these atoms is weak in native serine proteinases, in an enzyme:substrate complex the hydrogen bond becomes stronger, establishing the proton transfer pathway and helping to activate Ser195 O^{γ} as a nucleophile. In the formation of the complex between BT and PTI, for example, the distance between $N^{\epsilon 2}$ and O^{γ} decreases from 3.0 to 2.6Å (Marquart *et al.*, 1983). Most workers agree that this interaction is long and poorly oriented in native enzymes [SGPA (James *et al.*, 1980), BT (Chambers and Stroud, 1979; Marquart *et al.*, 1983); α-lytic protease (Fujinaga *et al.*, 1985) and CHT (Blevins and Tulinsky, 1985)], but

Tsukada and Blow (1985) see a strong hydrogen bond in native CHT. In SGT, the interaction between His57 and Ser195 is very similar to that seen in native BT (Figure II.14); the atoms of the catalytic triad (His57, Asp102 and Ser195) in the two enzymes can be superimposed with an rms coordinate difference of 0.24Å, and the distance from $N^{\epsilon 2}$ to O^{γ} in SGT is 3.0Å. This interaction can be seen more clearly in Figure II.18, which shows the active site residues of SGT and their associated electron density in the final map. The fit of the model to the density suggests that there is little error in the positions of these atoms; from the results illustrated in Figure II.8, the estimated coordinate errors are 0.08Å for His57 $N^{\epsilon 2}$ and 0.14Å for Ser195 O^{γ}.

There has been disagreement among researchers about the position of the amino-terminus in SGT. Because of sequence

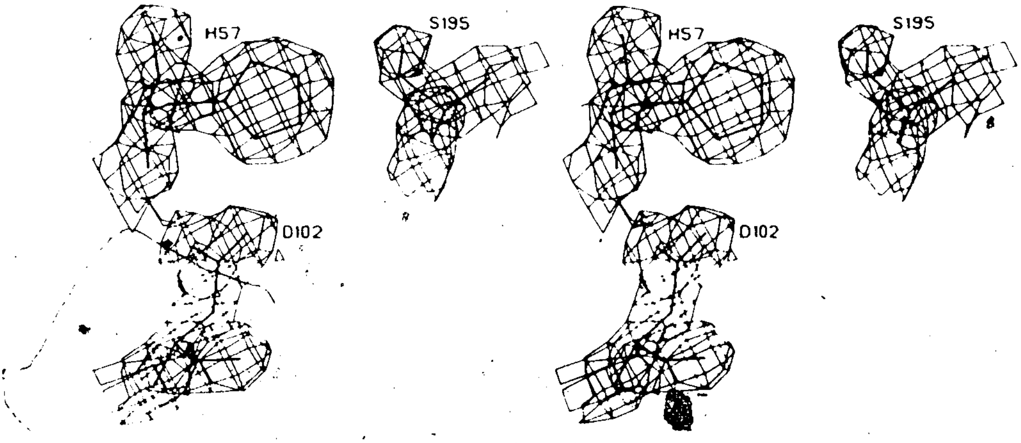


Figure II.18. SGT Active Site in Electron Density. The active site residues His57, Asp102 and Ser195 are shown in their associated electron density from the final map, contoured at a level of 0.60e/Å^3. For clarity, contours further than 1.5Å from atoms in the figure are omitted.

similarities to BT and the resistance of the amino group of Val16 in SGT to acetylation and carbamylation, it was suggested that the N-terminus is involved in a buried ion pair with Asp194, as in BT and other mammalian serine proteinases (Awad and Ochoa, 1974; Olafson and Smillie, 1975). In addition, Jurášek *et al.* (1976) found no difficulties in constructing a model of SGT in this region on the basis of its homology with BT. On the other hand, Duggleby and Kaplan (1975) found that the N-terminus of SGT reacts with 1-fluoro-2,4-dinitrobenzene, and concluded that this group is exposed; however, Jurášek *et al.* (1976) point out that the control experiment with BT was not performed and might very well have given the same result. As is apparent from Figure II.14, the conformations of BT and SGT are in fact very similar around the N-terminus. Figure II.19 shows a more detailed comparison of BT and SGT in this region. In BT, the N-terminal residue is isoleucine, while in SGT it is valine. These are the only amino acids found in this position in the serine proteinase sequences summarized by de Haen *et al.* (1975), Greer (1981a) and Hewett-Emmett *et al.* (1981); isoleucine is the more common choice among these sequences. One may note in Figure II.19 the additional small differences in sequence and conformation that compensate for the presence or absence of the methyl group by which these amino acids differ. The most obvious change is at residue 190, where the extra methyl group of the threonine in SGT compared to the serine in BT provides the

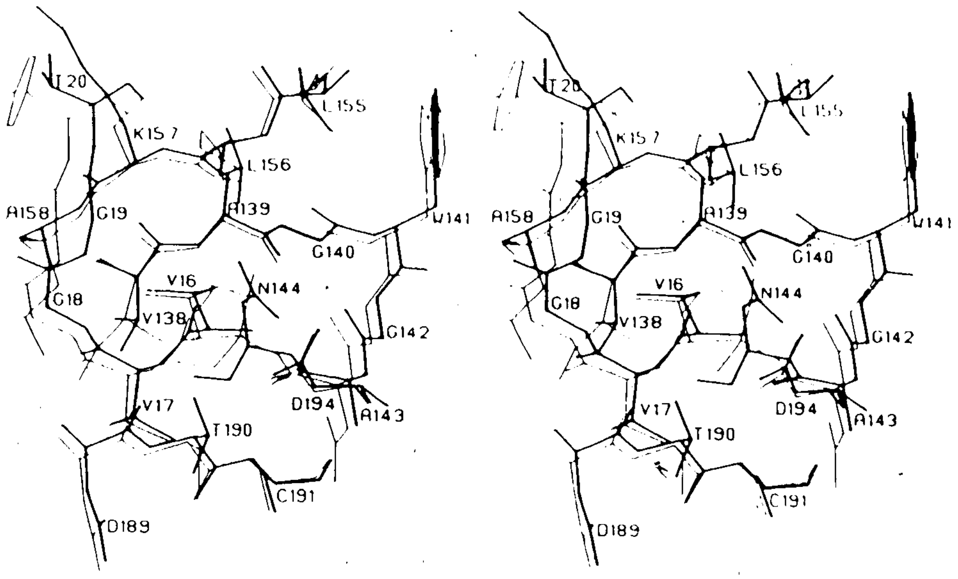


Figure II.19. Comparison of N-termini of SGT and BT. Thick lines and labels indicate residues in the vicinity of the N-terminus in SGT. Thin lines show the corresponding residues in BT.

exact compensation in volume that is needed. In the superposition shown in Figure II.19, Thr190 $C^{\gamma 2}$ of SGT is only 3.35Å from Ile16 $C^{\delta 1}$ of BT. Another difference that might fill the hole left by the removal of the methyl group is the altered conformation of Val138 in SGT compared to Ile138 of BT. Val138 $C^{\gamma 2}$ of SGT is 3.65Å from Ile16 $C^{\delta 1}$ of BT. The existence of these small compensating differences emphasizes the complementarity of the N-termini and the pockets into which they fit.

As is apparent in Figure II.14, a number of regions of SGT have considerably larger conformational differences than the subtle changes that occur at the N-terminus. These differences are of particular relevance to the task of building a comparative model of SGT from the structure of BT. Rather than discussing them in the context of structure comparison, selected examples will be used in the next section to illustrate the difficulties inherent in comparative modelling.

C. Critical Evaluation of Comparative Models of SGT

The general protein folding problem, that of deducing the minimum free energy conformation from just an amino acid sequence, is far from being solved. Since proteins seem to fall into a reasonably small number of structural families (Dayhoff, 1972), we often know that a structure will be similar to that of a homologous protein. This provides a powerful set of constraints that makes comparative model-building much more tractable than prediction from sequence alone.

One of the earliest uses of comparative model-building was the prediction of the structure of BT from that of the homologous serine proteinase CHT (Hartley, 1970). The specificity of BT for Arg or Lys in the P_1 position of the substrate was successfully explained by the substitution of Asp189 for Ser in a region that is extremely similar in the two proteins.

However, most of the interesting questions addressed by comparative model-building involve features that are unique to the unknown structure, for example non-homologous parts of an enzyme involved in extended substrate specificity (Furie *et al*., 1982; Strassburger *et al*., 1983). Recently, it has been proposed that comparative models be used for the rational design of highly specific drugs (Blundell *et al.*, 1983; Blow, 1983). The drug industry is reported to have shown considerable interest in the atomic coordinates for human renin, based on crystal structures of other aspartyl proteinases (The Economist, 1984).

To use such comparative models, it is essential to have an idea of the probable errors in both the homologous and the non-homologous parts. Yet few of the protein structures that have been modelled have subsequently been determined by X-ray crystallography. The structure of BT has been known for some time (Stroud *et al.*, 1972; Chambers and Stroud, 1979; Bode and Schwager, 1975), but the overall accuracy of the model based on CHT has not been assessed. Delbaere *et al.* (1979) evaluated the model of α-lytic protease built by McLachlan and Shotton (1971) from the structure of elastase, but the large errors they found might not be typical of models based on more closely homologous structures. The sequences of elastase and α-lytic protease are identical in only 18% of the amino acids, whereas BT and CHT have 45% sequence identity (James *et al.*, 1978).

SGT has been modelled twice on the basis of its homology to BT (Jurášek *et al.*, 1976; Greer, 1981a). Following the solution of the structure of SGT, it became possible to examine the types and sizes of errors in comparative models, and the effects these errors may have on the usefulness of such models. The analysis of the comparative models was performed after cycle 78 of the least-squares refinement of SGT, but the subsequent refinement does not significantly affect the conclusions.

Comparative Models of SGT

Jurášek *et al.* (1976) built a model of SGT from Watson-Kendrew protein components, based on preliminary coordinates for BT provided by R. M. Stroud. The path of the polypeptide was altered only where necessary to accommodate insertions and deletions. As far as possible, side-chain conformations of similar or identical residues were retained in the model of SGT. This model will be referred to as LJ-SGT.

It might be argued that LJ-SGT does not represent the state of the art of comparative model-building, because it was not constructed using computer graphics. However, as long as computers are used only as a tool for manually adjusting torsion angles, albeit a more efficient and somewhat more accurate tool, there is no significant conceptual difference between a physical model and a computer-built model. Few comparative model-building studies have used computer techniques for the automatic adjustment of conformation. Furie *et al.* (1982) used a program to adjust side-chain (but not main-chain) torsion angles to minimize unfavourable close contacts in models of blood coagulation factors based on the pancreatic serine proteinases. Warme *et al.* (1974) built a model of α-lactalbumin from the structure of hen egg-white lysozyme, then refined it by energy minimization, which included some exploration of alternative conformations. Unfortunately, none of the structures of the proteins treated in these two studies has yet been determined.

Greer's (1981a) approach to modelling SGT was somewhat different, leading not to an actual model, but rather to a descriptive outline of a model. The structures of 3 serine proteinases (BT, CHT and elastase) were aligned, as were the sequences of these and 8 related proteinases of unknown structure, including SGT. This revealed structurally conserved regions (SCRs) and variable regions (VRs). The sequence alignment provided the core of SCRs for the model of each unknown proteinase. Instead of constructing the VRs, Greer suggested which known structure would give the best starting model for each VR. The model Greer outlined for SGT will be called JG-SGT.

Evaluation of Sequence Alignments

To build a comparative model, one must correctly align the amino acid sequences for the homologous parts; alignments of non-homologous parts are essentially arbitrary for structural purposes. Insofar as the sequence alignment is incorrect, the model is guaranteed to be wrong. Table II.1 shows the results of the structural alignment of SGT and BT (see Figure II.14), and indicates the homologous segments. We can compare this structurally derived sequence alignment with those of the two SGT models, and with those from three studies more concerned with evolutionary relationships (de

Haën *et al*., 1975; Hewett-Emmett *et al*., 1981; Titani *et al*., 1983). Of the 190 homologous residues, de Haën *et al*. (1975) correctly matched 151 (79%), Hewett-Emmett *et al*. (1981) 156 (82%), Titani *et al*. (1983) 149 (78%), Greer (1981a) 173 (90%) and Jurášek *et al*. (1976) 174 (91%). This supports Greer's assertion that alignment by maximizing sequence equivalence (minimizing evolutionary distance) is inadequate for model-building (Greer, 1981a), though it may be appropriate for tracing evolutionary pathways. Even using structural information, major alignment errors occur. For example, the segment 63-73 (which includes Greer's SCR 63-71) is structurally homologous in SGT and BT, but virtually undetectable sequence homology caused all predicted alignments to be wrong. Table II.12 shows correct and incorrect alignments of this segment. Because of the misalignments, neither LJ-SGT nor JG-SGT include the short extra β-strand E (Table II.10) that results from the insertion, relative to BT, at position 60A.

Constructing Homologous Parts

In comparative model-building, it is generally assumed that side-chains in similar environments will adopt similar conformations. This is a useful generalization, supported by the fact that all 475 atoms of the 70 identical, structurally equivalent residues of SGT and BT superimpose with an rms deviation of 1.01Å. But there are exceptions. Figure II.20 shows a region in which strong sequence

Table II.12

Incorrect Sequence Alignments of SGT with BT

	55		57			60														70					75
BT	A	A	H	C	Y	K	S	-	-	-	-	-	G	I	Q	V	R	L	G	E	D	N	I	N	V
SGT	A	A	H	C	V	S	G	S	G	N	N	T	S	I	T	A	T	G	G	V	V	D	L	-	Q
LJ-SGT	A	A	H	C	V	S	G	-	-	-	-	-	S	G	N	N	T	S	I	T	A	T	G	G	V
JG-SGT	A	A	H	C	V	S	G	S	-	-	-	-	G	N	N	T	S	-	I	T	A	T	G	G	V

						80					84
BT	-	V	E	G	N	E	-	Q	F	I	S
SGT	-	S	G	-	A	A	-	V	K	V	R
LJ-SGT	-	V	D	L	Q	S'	A	V	K	V	R
JG-SGT	V	D	L	Q	S'	A	-	V	K	V	R

'As noted above, the sequence must be corrected by the insertion of two residues, currently interpreted from the electron density as Gly-Ala.

similarity gives rise to varying degrees of structural similarity.

An inspection of Figures II.14 and II.20 shows that there is an additional complication in modelling homologous parts. Local structure is conserved much more strongly than global structure. Individual homologous segments have very similar conformations, but differ slightly in their orientation and position relative to other homologous segments. The data in Table II.13 demonstrate the improvement in alignment obtained by considering the segments individually instead of globally. A particularly clear example is the

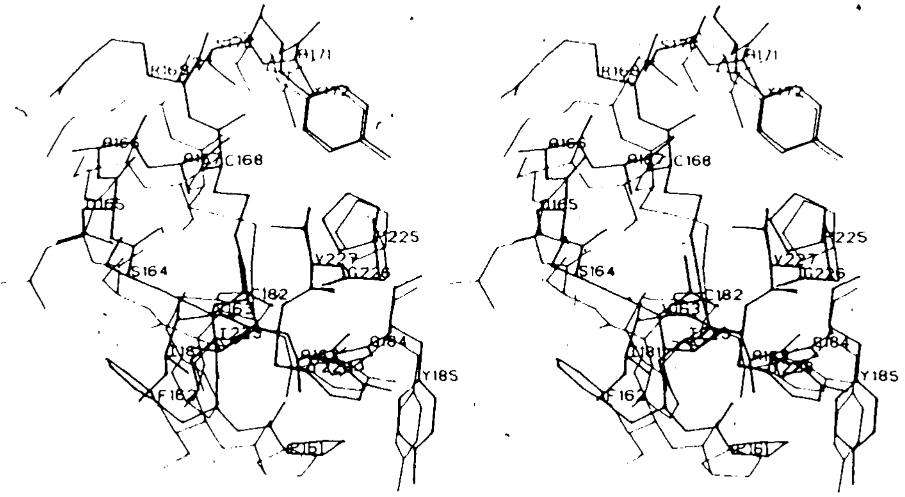


Figure II.20. Comparison in Region of Similar Sequence. Shown are SGT (thick lines) and BT (thin lines) in the region composed of the segments 161-172, 181-185 and 225-229. The sequences and, for the most part, the structures are very similar in this region. However, there are larger differences in the segment 161-171, including a sizeable shift of the main-chain. The side-chain of Phe162 does not occupy the same volume as that of Ile162 in BT. In part, this may compensate for the replacement of Phe181 by Ile and of Cys136 by Phe (see below) in SGT. Finally, the side-chains of Asp165 and Ser170 and the disulfide bridge 168-182 take on different conformations. The differences between the two structures in this region are probably related to the large differences in the segment 129-134 (see below).

C-terminal α-helix. Figure II.14 shows that this helix is packed more tightly against the β-sheet in SGT than in BT. Closer packing might have been predicted from the less bulky side-chains in SGT between these elements of secondary structure. In BT, for example, the side-chains of His91, Trp237, Thr241 and Ile242 are found in this interface. All of these residues are alanines in SGT. At the C-terminus, Asn245 is replaced by Leu in SGT. A substitution of Ile51 for Trp51 opens a hydrophobic pocket for Leu245 in SGT,

Table II.13

Global and Local Conservation of Structure

SGT segment[1]	BT segment[1]	rms (global) (Å)[2]	rms (local) (Å)[2]	angle (°)[3]	distance (Å)[3]
V16-Q24	I16-A24	1.25	0.81	9.0	0.49
G25-S34	N25-N34	1.24	0.93	9.8	0.15
G41-Q49	F41-S49	0.84	0.77	2.5	0.21
D50-V59	Q50-Y59	0.81	0.56	5.9	0.04
S63-L73	G63-I73	1.19	0.93	3.9	0.09
V81-K87	Q81-K87	0.80	0.35	3.1	0.35
V88-N95	S88-N95	1.21	0.87	10.2	0.05
G100-I106	N100-I106	0.50	0.29	7.5	0.14
K107-N113	K107-S113	1.51	0.61	9.2	0.56
P119-T129	A119-C129	1.54	0.76	28.5	0.52
G134-R145	T134-K145	0.92	0.50	6.4	0.02
R153-S164	D153-S164	0.99	0.55	6.3	0.00
D165-V177	D165-T177	1.24	0.83	4.4	0.76
N178-P185A	S178-L185A	0.72	0.50	7.2	0.17
T185C-G193	E185C-G193	0.65	0.45	5.2	0.09
D194-K202	D194-S202	0.95	0.79	5.6	0.05
W207-G216	G207-G216	0.68	0.54	9.6	0.04
Y217-P225	S217-P225	0.38	0.23	-4.4	0.11
G226-A235	G226-V235	0.80	0.46	2.4	0.27
S236-L245	S236-N245	1.53	0.59	4.7	0.37

[1] The structurally equivalent residues were divided into segments of about 10 residues, broken where possible at turns.
[2] Rms (global) and rms (local) refer to the rms deviation between main-chain atoms (N, C^{α}, C, O) of the ~~segments~~, calculated from the global superposition of BT on SGT, and after a local superposition by least squares of the main-chain atoms of the segment.
[3] Angle and distance are a measure of the amount of reorientation involved in the local superposition, and refer to the angle of rotation and distance of translation along the rotation axis.

allowing the end of the helix to approach more closely to the sheet. Similar observations of relative shifts of secondary structural elements have been made in other protein families by Lesk and Chothia (1980; 1982; Chothia and Lesk, 1982).

The segment 217-225 has the best fit between SGT and BT in both the global and the local superpositions (Table II.13). Yet it is not obvious from the sequence alignment (Table II.1) that this should be so. However, there is at least one functional reason for this strong similarity. As noted above, the charged side-chain of Asp189 is necessary to the P_1 specificity of BT and SGT for Arg or Lys. J. Moult (personal communication) has speculated that the burial of this charge in the specificity pocket of BT would lead to structural instability if the protein environment did not provide electrostatic stabilization. He has examined electrostatic interactions that stabilize the buried charge of Asp189 in BT, and finds that the most important interactions include the dipoles of the peptide bonds 220-221, 221-222 and 225-226. The strong conservation of the main-chain conformation in this part of SGT provides additional evidence of the functional importance of the orientation of these dipoles. This is quite a subtle point compared to the normal considerations in comparative modelling, but it is an example of the type of reasoning that might usefully be incorporated as the technique becomes more sophisticated.

Some errors arise from expecting the two structures to be similar where they are not. In both LJ-SGT and JG-SGT, the segment 129-134 was expected to be homologous to BT. In fact, the two proteins differ markedly here (Figure II.21). Several features of SGT are incompatible with the

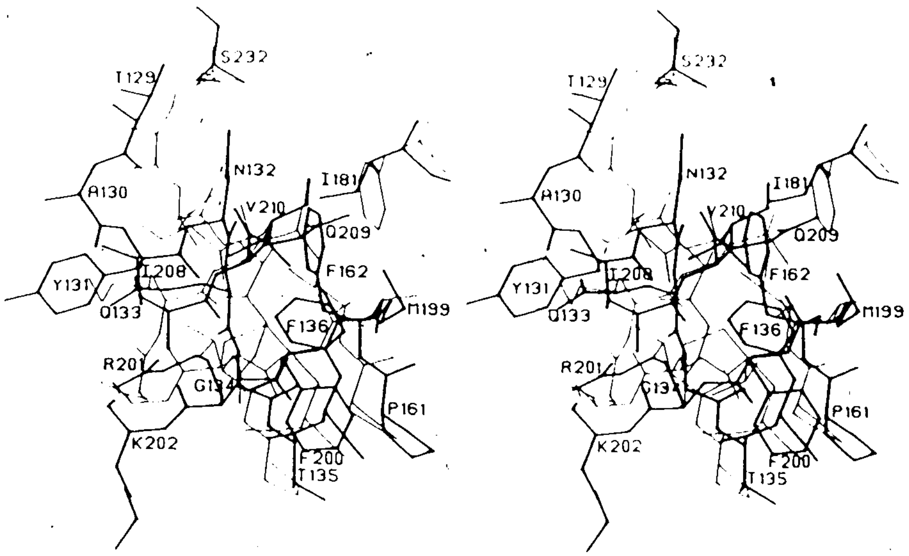


Figure II.21. Conformational Difference in Residues 129-134. A comparison of SGT (thick lines and residue labels) and BT (thin lines) is shown in the region of the segment 129-134.

conformation in BT; which of these cause the difference and which are compensating readjustments cannot be determined unambiguously. An examination of Figure II.21 provides one possible interpretation. The disulfide bridge 136-201 in BT is replaced by Phe136 and Arg201 in SGT; the volume is taken up by the Phe136 side-chain and part of the Arg201 side-chain. Phe136 and, as noted above, Ile181 exclude the side-chain of Phe162 from the volume occupied by Ile162 in BT. The other favoured conformations of Phe162, including the one observed, are incompatible with the main-chain conformation of residues 129-134 in BT. In addition, Gly133 in BT is in a left-handed helical conformation, which would be highly unusual for a threonine residue. The absence in SGT of the constraints from the disulfide bridges 129-232 and

136-201 in BT might also be relevant. A reasonable goal for more sophisticated methods of comparative model-building would be the ability to recognize the necessity for conformational change in cases such as this.

Constructing Non-Homologous Parts

The most difficult task in comparative model-building is constructing the non-homologous parts. This problem is a small-scale, somewhat constrained, version of the general folding problem. The model of α-lactalbumin was constructed in the regions where it is not homologous to hen egg-white lysozyme by exploring possible conformations of dipeptides, combining dipeptides with low energy conformations, then refining by energy minimization (Warme *et al.*, 1974).[7] Exhaustive exploration of conformations for whole regions would probably be superior, though computationally demanding, and the use of molecular dynamics might obviate the need for manual intervention to escape local energy minima. Nonetheless, this work represents an exception to the usual procedures, and it will be interesting to see how close this model is to the true structure. In most comparative models, non-homologous segments have been constructed by intuition,

[7] It should be noted that Novotný *et al.* (1984) question the validity of using the total energy calculated from empirical energy functions to judge comparative models. They find comparable energies, after energy refinement, for correct structures and for incorrect structures built by substituting completely unrelated sequences. However, they suggest that the inclusion of adequate models for solvent effects would alleviate this problem.

combined with a fairly limited manual exploration of the possible conformations (Furie *et al.*, 1982; Strassburger *et al.*, 1983; Blundell *et al.*, 1983; Jurášek *et al.*, 1976; Greer, 1981b). Two examples from LJ-SGT demonstrate the inadequacy of this approach.

In SGT, the loop 34-41 has a deletion of 3 residues relative to that in BT. In contrast, the loop 200-209 in SGT requires the insertion of 5 residues into the structure of BT. Neither the deletion nor the insertion was modelled correctly in LJ-SGT [Figures II.22 and II.23]. For JG-SGT, it was suggested that good models would be provided by BT for the loop 34-41, and CHT for the loop 200-209. As seen in LJ-SGT, a minimal perturbation of BT does not provide an adequate model of the first loop (Figure II.22). On the other hand, CHT, which lacks only one residue in the loop 200-209 compared to SGT, is indeed quite similar to SGT in this region (Figure II.24).

Effects of Errors

At the present level of sophistication, comparative model-building has a number of sources of error. The seriousness of the errors will depend on the accuracy required for the intended use of the model. For use in molecular replacement, large errors in the model are evidently tolerated; even BT was a sufficiently good model of SGT, though barely.

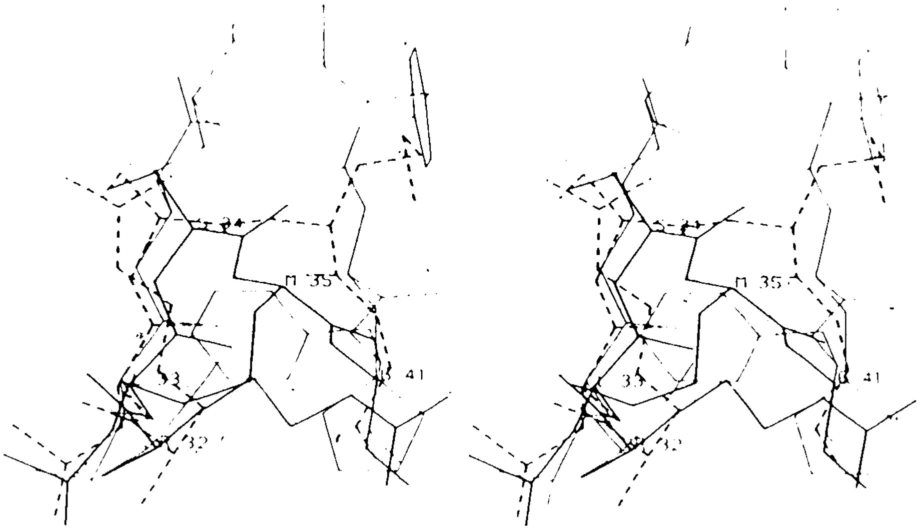


Figure II.22. Errors in Model-building of Deletion. Comparison of the segment 32-41 in SGT (thick lines), BT (thin lines) and LJ-SGT (dashed lines). Coordinates for LJ-SGT used in this and in Figure II.23 were measured from the model, refined against restraints for ideal geometry (Hendrickson and Konnert, 1980) and oriented relative to BT by least squares superposition of those atoms judged during model-building (Jurášek *et al.*, 1976) to be identical in SGT and BT. Some differences in conformation (e.g., the side-chain of Leu33) probably result from the fact that the BT coordinates available for model-building in 1976 were unrefined, preliminary coordinates supplied by R. M. Stroud. The loop 32-41 is 3 residues shorter in SGT than in BT. LJ-SGT accommodates this deletion with minimal perturbation of the main-chain from BT. In this model, the side-chain of Arg32 projects into the solvent. In fact, the main-chain and side-chain of Met35 occupy the space vacated by the change of Phe41 to Gly in SGT. The side-chain of Arg32 crosses the loop, placing the guanidinium group in the position occupied by the side-chain of His40 in BT. His40 has been implicated in the stabilization of the zymogen forms of chymotrypsin (Freer *et al.*, 1970) and trypsin (Fehlhammer *et al.*, 1977). It is evidently not essential to the zymogen mechanism, however, because in kallikrein, Phe40 has the same position and conformation (Bode *et al.*, 1983). Whether a zymogen exists for SGT is not known; this unexpected conservation of a positive charge may occur for some other reason.

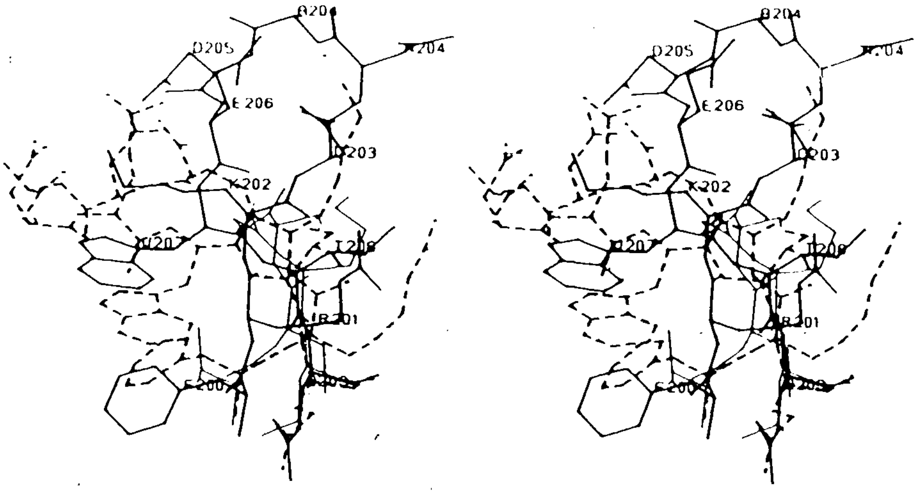


Figure II.23. Errors in Model-building of Insertion. Comparison of the segment 200-209 in SGT (thick lines), BT (thin lines) and LJ-SGT (dashed lines). This loop is longer by 5 residues in SGT than in BT. In LJ-SGT, the extension is folded back over the side-chain of Trp207, covering this potential hydrophobic surface and making additional contacts to the rest of the protein. In fact, the extension projects out from the surrounding surface of the protein (compare with Figure II.10).

One must also recognize that the accuracy of homologous and non-homologous parts of comparative models differs quite widely. Predictions of conserved structure involving conserved sequence are almost certain to be correct, as was the prediction for SGT that the N-terminus would form an ion pair with the side-chain of Asp194 (Jurášek *et al.*, 1976). It is also reasonable to make predictions involving non-conserved amino acids in homologous regions [e.g., the role in substrate specificity of Asp189 in trypsin (Hartley, 1970)], though these are more susceptible to errors in side-chain conformation, or to misalignments of sequence.

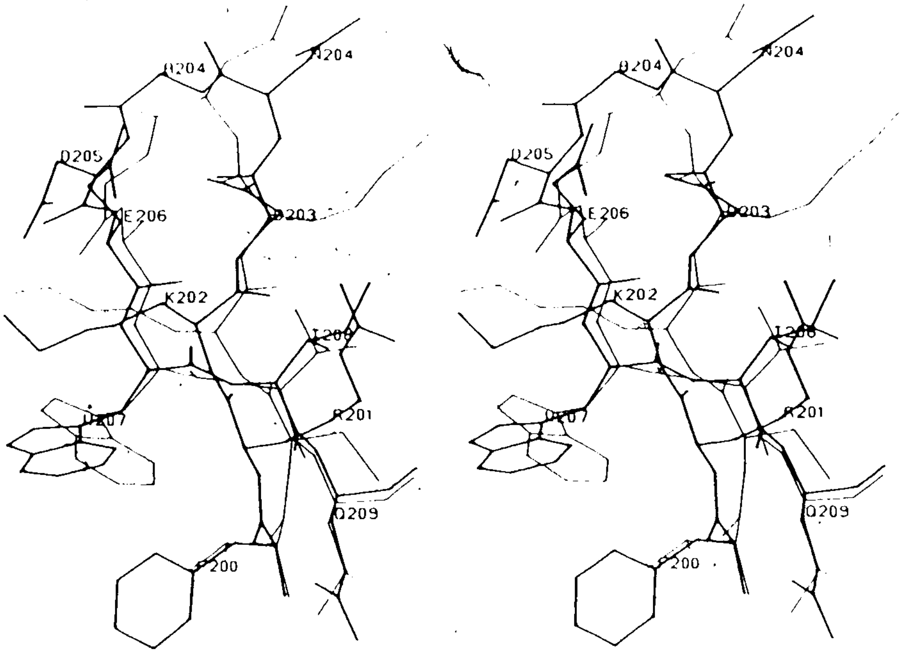


Figure II.24. Similar Loop in SGT and CHT. Comparison of SGT (thick lines) and CHT (thin lines) in the same region shown in Figure II.23. As predicted by Greer (1981a), SGT is very similar here to CHT, differing mainly by a single residue insertion at the end of the loop.

Predictions involving models of non-homologous regions are generally of more interest but will be much less reliable. On a gross level, these models can suggest which parts of an enzyme contribute to unique aspects of substrate specificity (Furie *et al.*, 1982; Strassburger *et al.*, 1983), or help to organize experimental data [e.g., data on protection from proteolysis when haptoglobin binds to hemoglobin (Lustbader *et al.*, 1983)]. The details are, however, quite likely to be inaccurate. To attain the accuracy and precision needed to design highly specific drugs (Blundell *et al.*, 1983; Blow, 1983) will require considerably more sophisticated techniques.

Bibliography

Abbott, F., Gomez, J. E., Birnbaum, E. R., & Darnall, D. W. (1975) *Biochemistry 14*, 4935-4943.

Adams, M. J., Haaŝ, D. J., Jeffery, B. A., McPherson, A. Jr., Mermall, H. L., Rossmann, M. G., Schevitz, R. W., & Wonacott, A. J. (1969) *J. Mol. Biol. 41*, 159-188.

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

Awad, W. M. Jr., & Ochoa, M. S. (1974) *Biochem. Biophys. Res. Commun. 59*, 527-534.

Barry, C. D., Molnar, C. E., & Rosenberger, F. U. (1976) *Technical Memo No. 229*, Computer Systems Lab, Washington University, St. Louis, MO.

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

Bier, M., & Nord, F. F. (1951) *Arch. Biochem. Biophys. 33*, 320-332.

Birktoft, J. J., & Blow, D. M. (1972) *J. Mol. Biol. 68*, 187-240.

Birktoft, J. J., Blow, D. M., Henderson, R., & Steitz, T. A. (1970) *Phil. Trans. Roy. Soc. Ser. B 257*, 67-76.

Blevins, R. A., & Tulinsky, A. (1985) *J. Biol. Chem. 260*, 4264-4275.

Blow, D. (1983) *Nature (London) 304*, 213-214.

Blow, D. M., & Crick, F. H. C. (1959) *Acta Cryst. 12*, 794-802.

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

Blundell, T., Sibanda, B. L., & Pearl, L. (1983) *Nature (London) 304*, 273-275.

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.

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

Chothia, C., & Janin, J. (1982) *Biochemistry 21*, 3955-3965.

Chothia, C., & Lesk, A. M. (1982) *J. Mol. Biol. 160*, 309-323.

Crowther, R. A. (1973) in *The Molecular Replacement Method* (Rossmann, M. G., Ed.) International Science Review 13, pp 173-178, Gordon & Breach, New York.

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

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

Cullis, A. F., Muirhead, H., Perutz, M. F., Rossmann, M. G., & North, A. C. T. (1961) *Proc. Roy. Soc. Ser. A 265*, 15-38.

Darnall, D. W., Abbott, F., Gomez, J. E., & Birnbaum, E. R. (1976) *Biochemistry 15*, 5017-5023.

Dayhoff, M. O. (1972) *Atlas of Protein Sequence and Structure*, Vol. 5, National Biomedical Research Foundation, Washington, D. C.

de Haen, C., Neurath, H., & Teller, D. C. (1975) *J. Mol. Biol. 92*, 225-259.

Delbaere, L. T. J., Brayer, G. D., & James, M. N. G. (1979) *Nature (London) 279*, 165-168.

Duggleby, R. G., & Kaplan, H. (1975) *Biochemistry 14*, 5168-5175.

Einspahr, H., & Bugg, C. E. (1980) *Acta Cryst. B36*, 264-271.

Einspahr, H., & Bugg, C. E. (1981) *Acta Cryst. B37*, 1044-1052.

Einspahr, H., & Bugg, C. E. (1984) in *Metal Ions in Biological Systems* (Sigel, H., Ed.) Vol. 17, pp 51-97, Marcel Dekker, New York.

Epstein, M., Reuben, J., & Levitzki, A. (1977) *Biochemistry 16*, 2449-2457.

Fehlhammer, H., Bode, W., & Huber, R. (1977) *J. Mol. Biol. 111*, 415-438.

Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. T., & Xuong, Ng. H. (1970) *Biochemistry 9*, 1997-2009.

Fujinaga, M., Delbaere, L. T. J., Brayer, G. D., & James, M. N. G. (1985) *J. Mol. Biol. 183*, 479-502.

Furie, B., Bing, D. H., Feldmann, R. J., Robison, D. J., Burnier, J. P., & Furie, B. C. (1982) *J. Biol. Chem. 257*, 3875-3882.

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

Greer, J. (1981b) *J. Mol. Biol. 153*, 1043-1053.

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

Hartley, B. S., & Kauffman, D. L. (1966) *Biochem. J. 101*, 229-231.

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

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

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

Hendrickson, W. A., & Lattman, E. E. (1970) *Acta Cryst. B26*, 136-143.

Hendrickson, W. A., & Teeter, M. M. (1981) *Nature (London) 290*, 107-113.

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

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

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.

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

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

Jurášek, L., Olafson, R. W., Johnson, P., & Smillie, L. B. (1976) *Miami Winter Symp. 11*, 93-123.

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

Kretsinger, R. H. (1976) *Ann. Rev. Biochem. 45*, 239-266.

Lattman, E. E. (1973) in *The Molecular Replacement Method* (Rossmann, M. G., Ed.) International Science Review 13, pp 179-185, Gordon & Breach, New York.

Lenhert, P. G. (1975) *J. Applied Crystallogr. 8*, 568-570.

Lesk, A. M., & Chothia, C. (1980) *J. Mol. Biol. 136*, 225-270.

Lesk, A. M., & Chothia, C. (1982) *J. Mol. Biol. 160*, 325-342.

Lifchitz, A. (1983) *Acta Cryst. A39*, 130-139.

Lifson, S., Hagler, A. T., & Dauber, P. (1979) *J. Amer. Chem. Soc. 101*, 5111-5121.

Lustbader, J. W., Arcoleo, J. P., Birken, S., & Greer, J. (1983) *J. Biol. Chem. 258*, 1227-1234.

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

Marquart, M., Walter, J., Deisenhofer, J., Bode, W., & Huber, R. (1983) *Acta Cryst. B39*, 480-490.

Marsh, R. E., & Donohue, J. (1967) *Adv. Protein Chem. 22*, 235-256.

Martin, R. B. (1984) in *Metal Ions in Biological Systems* (Sigel, H., Ed.) Vol. 17, pp 1-49, Marcel Dekker, New York.

Matthews, D. A., Alden, R. A., Birktoft, J. J., Freer, S. T., & Kraut, J. (1977) *J. Biol. Chem. 252*, 8875-8883.

McLachlan, A. D. (1979) *J. Mol. Biol. 128*, 49-79.

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

Mosolov, V. V., Fedurkina, N. V., & Valueva, T. A. (1978) *Biochim. Biophys. Acta 522*, 187-194.

Nagata, K., & Yoshida, N. (1983) *J. Biochem. 93*, 909-919.

Nishikata, M., Kasai, K.-I., & Ishii, S.-I. (1981) *Biochim. Biophys. Acta 660*, 256-261.

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

Novotný, J., Bruccoleri, R., & Karplus, M. (1984) *J. Mol. Biol. 177*, 787-818.

Olafson, R. W., & Smillie, L. B. (1975) *Biochemistry 14*, 1161-1167.

Olafson, R. W., Jurášek, L., Carpenter, M. R., & Smillie, L. B. (1975) *Biochemistry 14*, 1168-1177.

Rabinovitch, D., & Shakked, Z. (1984) *Acta Cryst. A40*, 195-200.

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

Rao, S. N., Jih, J.-H., & Hartsuck, J. A. (1980) *Acta Cryst. A36*, 878-884.

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

Richardson, J. S. (1981) *Adv. Prot. Chem. 34*, 167-339.

Rossmann, M. G. (1973) *The Molecular Replacement Method* International Science Review 13, Gordon & Breach, 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.

Rühlmann, A., Kukla, D., Schwager, P., Bartels, K., & Huber, R. (1973) *J. Mol. Biol. 77*, 417-436.

Russin, D. J., Floyd, B. F., Toomey, T. P., Brady, A. H., & Awad, W. M. Jr. (1974) *J. Biol. Chem. 249*, 6144-6148.

Shotton, D. M., & Watson, H. C. (1970) *Nature (London) 225*, 811-816.

Sielecki, A. R., James, M. N. G., & Broughton, C. G. (1982) in *Computational Crystallography* (Sayre, D., Ed.) pp 409-419, Oxford University Press, Oxford.

Sim, G. A. (1959) *Acta Cryst. 12*, 813-815.

Sim, G. A. (1960) *Acta Cryst. 13*, 511-512.

Smith, J. L., & Hendrickson, W. A. (1982) in *Computational Crystallography* (Sayre, D., Ed.) pp 209-222, Oxford University Press, Oxford.

Smith, J. L., Hendrickson, W. A., Honzatko, R. B., & Sheriff, S. (1984) *Acta Cryst. A40*, C51.

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

Srinivasan, R., & Ramachandran, G. N. (1965) *Acta Cryst. 19*, 1008-1014.

Strassburger, W., Wollmer, A., Pitts, J. E., Glover, I. D., Tickle, I. J., Blundell, T. L., Steffens, G. J., Günzler, W. A., Ötting, F., & Flohé, L. (1983) *FEBS Lett. 157*, 219-223.

Stroud, R. M., Kay, L. M., & Dickerson, R. E. (1972) *Cold Spring Harbor Symp. Quant. Biol. 36*, 125-140.

Tashiro, M., Sugihara, N., Maki, Z., & Kanamori, M. (1981) *Agric. Biol. Chem. 45*, 519-521.

The Economist (January 7, 1984) *290* (Number 7323), 71-74.

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

Titani, K., Sasagawa, T., Woodbury, R. G., Ericsson, L. H., Dörsam, H., Kraemer, M., Neurath, H., & Zwilling, R. (1983) *Biochemistry 22*, 1459-1465.

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

Trop, M., & Birk, Y. (1968) *Biochem. J. 109*, 475-476.

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

Wählby, S. (1968) *Biochim. Biophys. Acta 151*, 394-401.

Warme, P. K., Momany, F. A., Rumball, S. V., Tuttle, R. W., & Scheraga, H. A. (1974) *Biochemistry 13*, 768-782.

Woolfson, M. M. (1956) *Acta Cryst. 9*, 804-810.

Yokosawa, H., Hanba, T., & Ishii, S.-I. (1976) *J. Biochem. 79*, 757-763.

Zeppezauer, M., Eklund, H., & Zeppezauer, E. S. (1968) *Arch. Biochem. Biophys. 126*, 564-573.

III. Turkey Ovomucoid Inhibitor Third Domain'

Protein inhibitors of serine proteinases are well studied, and much is known about the interactions between an inhibitor and its cognate enzyme (Laskowski and Kato, 1980). Most of these inhibitors act by a common mechanism; they bind very tightly to the enzyme (low K_M) but are hydrolyzed very slowly, if at all (low k_{cat}). The inhibitors have at least one peptide bond called the reactive site. This is the bond that interacts with the enzyme's catalytic site and is the one that is cleaved if and when hydrolysis occurs.

There are several families of homologous serine proteinase inhibitors; ovomucoid inhibitors belong to the pancreatic secretory trypsin inhibitor (Kazal) family (Laskowski and Kato, 1980). The structure of the third domain of the ovomucoid inhibitor of Japanese quail (OMJPQ3) was first reported at 2.8Å resolution (Weber *et al.*, 1981) and has subsequently been refined at 1.8Å resolution (Papamokos *et al.*, 1982). Also, the structure of the third domain of the silver pheasant ovomucoid (OMSVP3), has been refined at 1.5Å resolution (Bode *et al.*, 1985). The homologous third domain of the ovomucoid inhibitor of turkey (OMTKY3), used in this study, differs from OMJPQ3 at 5 residues and from OMSVP3 at only the P_1 position (Kato *et al.*,

'Versions of parts of this chapter have been published [Read, R. J., Fujinaga, M., Sielecki, A. R., & James, M. N. G. (1983) *Biochemistry 22*, 4420-4433] or accepted for publication [Read, R. J., & James, M. N. G. (in press) in *Proteinase Inhibitors* (Barrett, A. J., & Salvesen, G. S., Eds.) Elsevier Science Publishers, Amsterdam].

1978). The numbering origin for OMTKY3 corresponds to residue 131 of the complete ovomucoid inhibitor. An I follows the sequence numbers of the residues of the inhibitor in order to distinguish them from those of the enzyme.

α-Chymotrypsin (CHT) is in some respects the archetype of the family of serine proteinases to which it belongs. Its structure was determined in 1967 (Matthews *et al.*, 1967) and has recently been refined at about 1.7Å resolution by two independent groups (Blevins and Tulinsky, 1985; Tsukada and Blow, 1985). CHT is strongly inhibited by OMTKY3 (Empie and Laskowski, 1982).

Streptomyces griseus Protease B (SGPB), a serine proteinase from the same family as CHT, is obtained from the extracellular culture filtrate, Pronase. Its structure has been described at 2.8Å resolution (Delbaere *et al.*, 1975) and is now refined at 1.7Å resolution (L. Sawyer, A. R. Sielecki and M. N . G. James, unpublished). The three-dimensional structure has been compared with those of other serine proteinases, including CHT, and shows a high degree of topological equivalence (James *et al.*, 1978). Kinetic studies have also been carried out to determine the specificity of its binding sites (Bauer, 1978; James *et al.*, 1980a). Like CHT, SGPB is strongly inhibited by OMTKY3 (Laskowski *et al.*, 1983). The numbering of the SGPB molecule according to the sequence of chymotrypsinogen (Hartley and Kauffman, 1966) follows a structural alignment with CHT (James *et al.*, 1978).

Aside from this study, the structures of several other complexes between a serine proteinase and a protein inhibitor have been determined. Most of the work has been done with the pancreatic trypsin inhibitor (PTI). The complexes of this inhibitor with bovine trypsin (PTI:BT, Huber *et al.*, 1974), anhydrotrypsin (PTI:BTan, Huber *et al.*, 1975) and trypsinogen (PTI:BTn), as well as the ternary complex with trypsinogen and Ile-Val (PTI:BTn:IV, Bode *et al.*, 1978) have been refined at high resolution (Marquart *et al.*, 1983). The refined, high-resolution structure of another ternary complex, [Arg^{15}]PTI (derived from PTI by a semi-synthetic procedure in which Lys15I, the P_1 residue, is replaced by arginine) with trypsinogen and Val-Val, has recently been reported (Bode *et al.*, 1984).

Several complexes that do not involve PTI have also been determined. The structure of the complex between trypsinogen and pancreatic secretory trypsin inhibitor (PSTI:BTn) has been refined at high resolution (Bolognesi *et al.*, 1982). Two complexes involving inhibitors of the potato-inhibitor I family (Laskowski and Kato, 1980) have recently been determined at high resolution: barley inhibitor CI-2 with subtilisin Novo (McPhalen *et al.*, 1985a) and leech inhibitor eglin with subtilisin Carlsberg (McPhalen *et al.*, 1985b). In addition, the structure of *Streptomyces* subtilisin inhibitor (SSI) with subtilisin BPN' has been partially refined at 2.2Å resolution (Hirono *et al.*, 1984) and the unrefined structure of soybean trypsin inhibitor with porcine trypsin (STI:PT, Sweet *et al.*, 1974) has been reported.

A. Solution of Crystal Structures

The two complexes, OMTKY3:SGPB and OMTKY3:CHT, both crystallize in the space group $P2_1$. For both, diffraction data were collected to 1.8Å resolution on a Nonius CAD4 diffractometer, and absorption effects were corrected by the method of North *et al.* (1968). Crystal data for the two crystals are summarized in Table III.1.

Table III.1

Crystal Data for OMTKY3 Complexes

Complex	OMTKY3:SGPB	OMTKY3:CHT
Cell dimensions		
a	45.34Å	44.92Å
b	54.52Å	54.52Å
c	45.65Å	57.18Å
β	119.2°	103.9°
No. of reflections measured	20227	27564
No. of unique reflections	18082	24883
Max. absorption correction factor	1.63	1.44
Max. decay correction factor	1.74[1]	1.44[2]

[1] Function of θ and time (Hendrickson, 1976).

[2] Function of time as measured by standard reflections.

Molecular Replacement

The structures were solved by molecular replacement (Rossmann, 1973), using the native proteinases as search models. The model of SGPB had been refined at 1.7Å resolution to an *R*-factor of 0.177 at the time of this work, and the model of CHT (Birktoft and Blow, 1972) was obtained from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977). A similar approach was followed in each case. Rotation functions were calculated with normalized structure factors ($|\mathbf{E}|$s) in the resolution shell 10.0-3.5Å. For the native data, $|\mathbf{E}_o|$s were determined with the program ORESTES (Thiessen and Levy, 1973). $|\mathbf{E}_c|$s were computed for search models placed in a cubic cell of symmetry P1, with a cell edge of 60Å for SGPB and 65Å for CHT. The rotation functions used integration limits of 3-21Å, and the results were unambiguous; the peaks were 11.4 standard deviations above the mean for SGPB in OMTKY3:SGPB, and 11.2 standard deviations above the mean for CHT in OMTKY3:CHT. Translation searches were performed by a brute force calculation of structure factor agreement for models translated over possible positions in the unit cell. Since $P2_1$ has a polar y axis, this search need only cover the quarter of the xz plane in which x and z vary from 0 to 1/2. For OMTKY3:SGPB, structure factor agreement was measured by the *R*-factor, which was 0.35 at the correct translation for data in the 5-4Å resolution shell. For OMTKY3:CHT, the correlation-coefficient between $|\mathbf{F}_o|$ and $|\mathbf{F}_c|$ was used instead;

this was 0.50 for 5-4Å data at the correct translation.

Since OMTKY3:SGPB was determined first, the structure of OMTKY3 was not yet known at that time. Using maps with the coefficients $(|\mathbf{F}_O|-|\mathbf{F}_C|)\exp(i\alpha_C)$ and $|\mathbf{F}_O|\exp(i\alpha_C)$, it was possible to construct a model of 50 residues of the inhibitor. (The first 6 residues in the sequence have never been visible in the electron density maps.)

When the OMTKY3:CHT structure was determined, the refined structure of the OMTKY3:SGPB complex was already known, so OMTKY3 was oriented in the unit cell by superimposing the active site of SGPB in its complex on the active site of the oriented CHT model. The electron density for the inhibitor was not very clear at this point, and this model did not fit well into density, but it was used nonetheless for the initial refinement cycles. After 8 cycles of refinement, a new map was computed, and it was clear that the inhibitor was grossly misoriented. Phases were calculated using just CHT from the partially refined structure, and a new map was computed with the coefficients $(2|\mathbf{F}_O|-|\mathbf{F}_C|)\exp(i\alpha_C)$. With a rigid body reorientation of OMTKY3, it was possible to achieve a reasonable fit to the density.

Both structures have been refined using the restrained-parameter least-squares refinement program of Hendrickson and Konnert (1980). Information concerning the final refinement parameters of the two structures is summarized in Table III.2. In the OMTKY3:SGPB structure, it was never

Table III.2

Final Refinement Parameters and Results

Structure	OMTKY3:SGPB	OMTKY3:CHT
No. of refinement cycles	65	79
Resolution limits(Å)	10.0-1.8	8.0-1.8
Data acceptance criterion	$I>\sigma(I)/2$	$I>\sigma(I)$
No. of reflections	16245	19178
R-factor	0.145	0.168
No. of protein atoms	1697	2151
No. of solvent atoms	182	222
No. of variable parameters	7699	9715
Rms deviations from ideal values[1]		
distance restraints(Å)		
bond distance	0.016(0.010)	0.016(0.012)
angle distance	0.035(0.020)	0.041(0.020)
planar 1-4 distance	0.039(0.020)	0.036(0.020)
plane restraint(Å)	0.021(0.012)	0.017(0.012)
chiral-center restraint($Å^3$)	0.181(0.080)	0.176(0.080)
non-bonded contact restraints(Å)		
single torsion contact	0.260(0.400)	0.331(0.300)
multiple torsion contact	0.150(0.400)	0.188(0.300)
possible hydrogen bond	0.199(0.400)	0.236(0.300)
conformational torsion angle		
planar(ω) restraint(°)	3.8(2.8)	3.3(3.0)
B-factor restraints($Å^2$)		
main-chain bond	2.192(1.500)	1.755(1.000)
main-chain angle	3.002(2.000)	2.617(1.500)
side-chain bond	4.102(2.500)	3.853(2.000)
side-chain angle	5.688(3.000)	5.215(2.500)

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

possible to locate the first 6 residues in the inhibitor sequence. However, in the refinement of OMTKY3:CHT, 2 additional residues were located, so that this structure contains 52 out of the 56 amino acids in the sequence of OMTKY3.

Since the structure of the OMTKY3:SGPB complex was refined first, it has been subjected to closer scrutiny.

Therefore, much of the discussion will concentrate on that complex.

Estimation of Error for OMTKY3:SGPB

Much of the structure analysis was performed after cycle 58 of least-squares refinement of OMTKY3:SGPB; little changed in the structure during the succeeding cycles of refinement. The analysis of coordinate error was also carried out at that point. A more complete discussion of the error estimation can be found in Read *et al*. (1983); the summarized results follow.

Three methods for estimating coordinate error were used, with surprising agreement obtained among the methods. From the method of Luzzati (1952), an estimate of 0.14Å for the overall rms coordinate error was obtained, but this method does not provide any information about the errors for individual parts of the structure. Individual estimated errors were obtained from the formula of Cruickshank (1949, 1954, 1967), which gives the error as a function of atom type and *B*-factor. The rms value of the estimated error for all of the atoms in the structure was, again, 0.14Å. A third method is to refine the structure for several cycles without structural restraints, then to compare the coordinates to those determined with restraints (Chambers and Stroud, 1979). Not only did this method give the same rms error (0.14Å), the coordinate differences as a function of atom type and *B*-factor agreed very well with the errors obtained from Cruickshank's (1949, 1954, 1967) equation.

B. Crystal Structures of the OMTKY3 Complexes

Structure of OMTKY3

OMTKY3 contains 56 amino acids, of which 50 are visible in the complex with SGPB, and 52 in the complex with CHT. The structure of OMTKY3 from the complex with SGPB is shown in Figure III.1. The inhibitor has the overall appearance of a wedge-shaped disc, with a diameter of approximately 30-32Å. It is about 19Å thick at the thickest part, and the thin edge of the wedge is the segment of polypeptide chain that contains the reactive bond (Leu18I-Glu19I). The secondary structural features of OMTKY3 consist of a three-stranded antiparallel β-sheet C-terminal to the reactive bond region (strand 1: Pro22I to Ser26I; strand 2: Asp27I to Tyr31I; strand 3: Ser51I to Gly54I) and a central α-helix (Asn33I to Ser44I).

Two disulfide bridges, Cys8I-Cys38I and Cys16I-Cys35I, and two asparagine residues position the N-terminal polypeptide chain (residues 8-20, which includes the reactive bond) relative to the central helix [Figure III.1(b)]. The side-chain of Asn39I forms hydrogen bonds to the main-chain of Lys13I. Asn33I anchors the two peptide bonds that are immediately adjacent to the reactive bond (Fujinaga *et al.*, 1982): the $N^{\delta 2}$ atom of Asn33I donates hydrogen bonds to the carbonyl-oxygen atoms of Thr17I and Glu19I and the $O^{\delta 1}$ atom

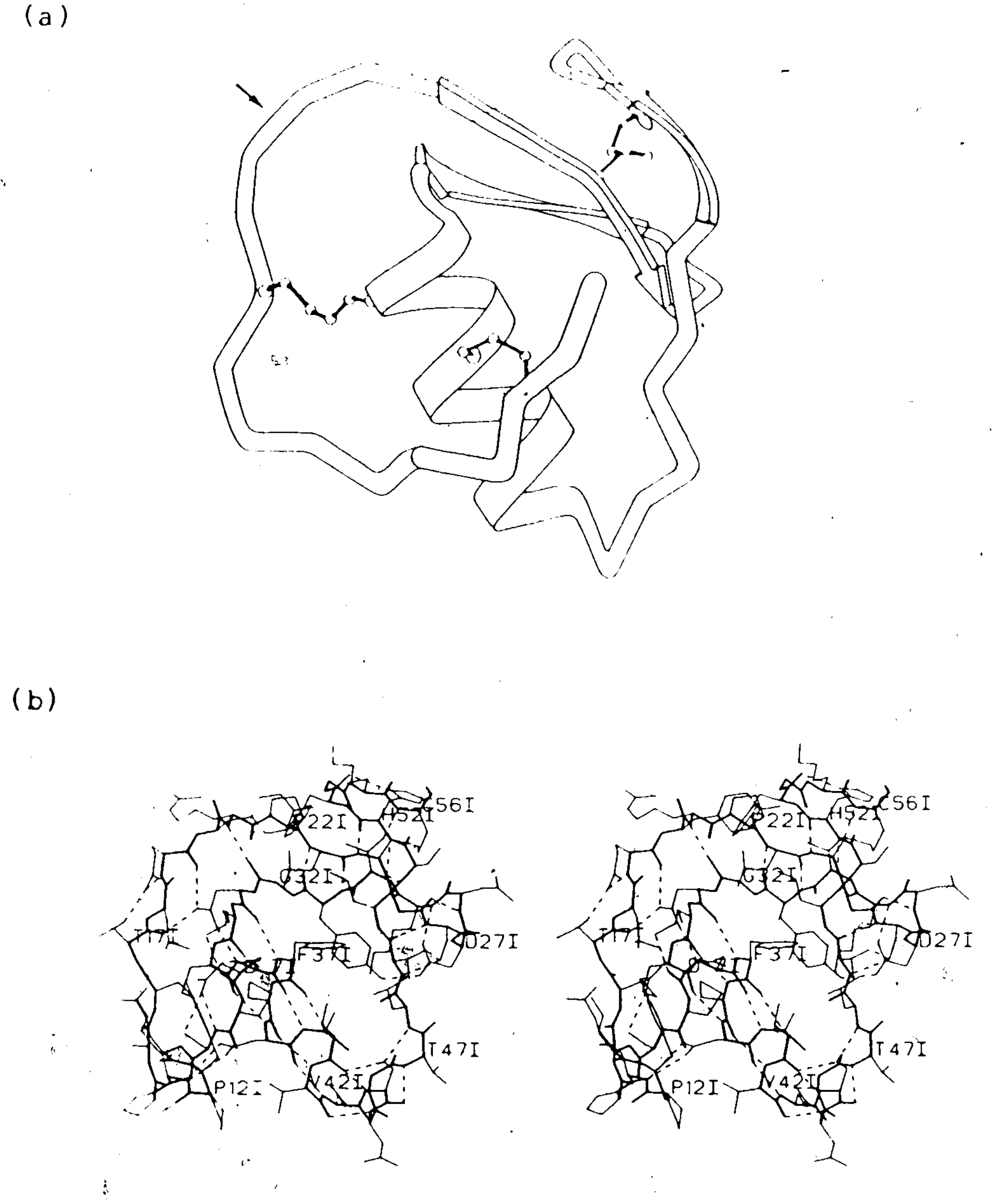


Figure III.1. Structure of OMTKY3 from Complex with SGPB. (a) Stylized representation of the chain fold, showing secondary structure and disulfide bridges. The small arrow indicates the position of the scissile peptide, Leu18I-Glu19I. Broad arrows represent strands of polypeptide chain in a β-sheet conformation, and the α-helix is represented as a coiled ribbon. (b) Stereographic view of the OMTKY3 molecule in a similar orientation to (a). Thick lines indicate the main-chain, thin lines represent the side-chains and dashed lines show hydrogen bonds. Every fifth amino acid residue is labelled.

is hydrogen bonded to the N-terminal position of the helix at Asn36I. Asn33I thereby acts as a spacer between the reactive site loop (primary contact region) and the secondary contact region (Papamokos *et al.*, 1982).

A least-squares minimization of the differences in α-carbon atom positions of the three ovomucoid third domains shows that they have very similar structures (Bode *et al.*, 1985). OMSVP3 and OMTKY3 are the most similar [rms deviation for 231 main-chain atoms is 0.36Å (Bode *et al.*, 1985)]. The main differences in conformation among these inhibitors involve the first 6 or 7 amino acids; these residues form a novel β-channel in OMJPQ3, whereas in OMTKY3 bound to SGPB, the first 6 residues are disordered and are not seen in the electron density map. Comparison of the porcine PSTI structure with those of the avian ovomucoids (Bolognesi *et al.*, 1982; Bode *et al.*, 1985) shows much larger structural differences, and different intermolecular contacts (at the N-termini and in the region of residues 37 to 48).

Common Structural Features of the Protein Inhibitors

The major common element in the structures of the protein inhibitors of serine proteinases is the primary contact region, or the reactive site loop. In all of the inhibitors of known structure, the conformation of this loop is highly complementary to the surface of the enzymes and likely resembles that of an oligopeptide substrate when bound to the serine proteinase active site. These reactive site loops project out from the inhibitors, so that they are accessible to the active sites of proteolytic enzymes. The pointed shape also serves to minimize potentially unfavourable contacts involving inhibitor residues other than those in the reactive site loop. The comparison of PTI and OMTKY3 shown in Figure III.2 demonstrates two very different ways in which similar reactive sites can be constructed. The conformations of residues P_2 through P_3' are extremely similar, but the parts of the inhibitors providing the framework

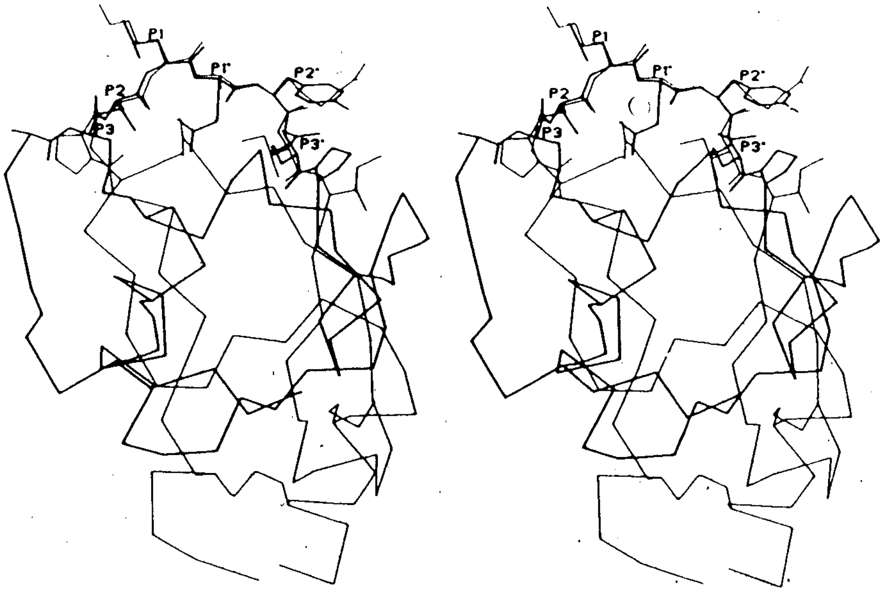


Figure III.2. Superposition of OMTKY3 and PTI Reactive Sites. OMTKY3 is shown in thick lines and PTI in thin lines. The relative orientation was determined by a comparison of the complexes OMTKY3:CHT and PTI:BT using a program of W. Bennett. All atoms are shown for structurally equivalent residues of the reactive site loops, and for the peptide bond Gly36I-Gly37I in PTI; the carbonyl-oxygen atom of Gly36I forms a hydrogen bond with the peptide nitrogen of the P_1' residue, as does the side-chain of Glu19I in OMTKY3. For other residues only C^α atoms and disulfide bridges are shown.

are quite different.

Bolognesi *et al.* (1982) have noted some other common structural features. In all of the inhibitors of known structure, the segment on the P_n' side of the reactive peptide bond is involved in β-sheet structure. This can be seen, for example, in the view of OMTKY3 in Figure III.1. However, Figure III.2 shows that the β-sheets of OMTKY3 and PTI are not structurally equivalent. In addition, many of the serine proteinase inhibitors have a disulfide bridge close to the reactive site on the P_n side (Laskowski and Kato, 1980).

Structures of the Complexes

To a first approximation, the formation of the inhibitor:enzyme complexes is an association of rigid bodies. The enzyme active site is preformed to accept a substrate; the inhibitor has a reactive site conformation that very likely resembles that of a bound substrate. Only relatively minor readjustments are required for complex formation. Some departures from this simplified picture will be discussed below. Figure III.3 shows alpha-carbon representations of the two complexes OMTKY3:SGPB and OMTKY3:CHT. Each complex is presented with the enzyme active site in approximately the same orientation.

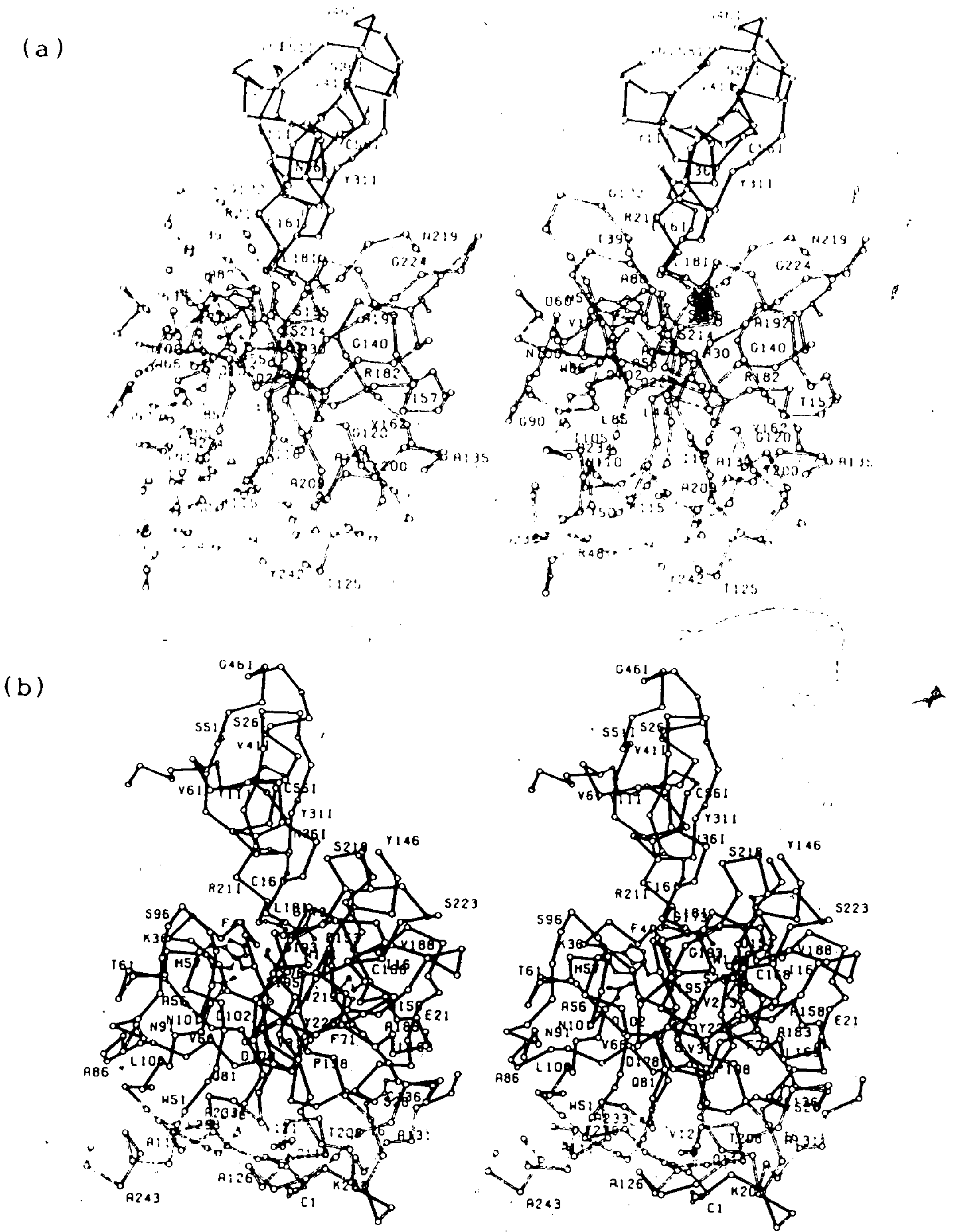


Figure III.3. C^{α}-atom Representations of Complexes. In both parts, OMTKY3 is shown with solid bonds and the enzyme [SGPB in part (a) and CHT in (b)] with open bonds. Every fifth amino acid residue is labelled. Side-chains are shown for disulfide bridges, for the catalytic residues His57, Asp102 and Ser195 in the enzyme and for the P_1 residue of the inhibitor. The atoms comprising the scissile bond are also shown.

C. Implications of the Structures

Geometry at the Reactive Site Carbonyl Group

Before discussing the mechanism of inhibition, it is appropriate to consider a subtle aspect of the interaction between enzyme and inhibitor that has sometimes been considered important to inhibitor action.

The initial 2.8Å structure of PTI:BT (Rühlmann *et al.*, 1973) and the 2.6Å structure of STI:PT (Sweet *et al.*, 1974) were both interpreted as showing a tetrahedral adduct in which there was a covalent bond between Ser195 O^{γ} in the enzyme active site and the carbonyl-carbon atom of residue P_1 in the inhibitor reactive site. The adduct was presumed to be stable because the enzyme stabilizes the transition state (Pauling, 1946). This attractive proposal had to be modified when refinement of PTI:BT at 1.9Å resolution showed that the distance from Ser195 O^{γ} to the carbonyl-carbon atom of the reactive peptide bond was too long for a normal covalent bond, but too short for a van der Waals contact (Huber *et al.*, 1974; Huber *et al.*, 1975). It was concluded that the geometry at the reactive site carbonyl was intermediate between that of a planar peptide and that of a tetrahedral intermediate.

The chemical environment of the reactive site carbonyl group in the protein inhibitors of serine proteinases has also been examined using ^{13}C-NMR (Baillargeon *et al.*, 1980; Richarz *et al.*, 1980). In all of the complexes STI:PT, PTI:BT, PTI:BTn and PTI:BTan, the NMR results were interpreted to rule out a covalent, fully tetrahedral adduct at the carbonyl-carbon atom. However, it was not possible to distinguish between a planar trigonal carbon atom and a tetrahedrally distorted one.

Bürgi *et al*. (1973) studied analogous interactions in small molecule crystal structures, which are much more precisely determined than protein structures. It was found that deviations from planarity at the carbonyl-carbon atom increased as a nitrogen nucleophile approached more closely. The carbonyl-carbon atom is displaced from the plane defined by the three atoms to which it bonds; the size of this out-of-plane displacement towards the nucleophile is a measure of the distortion from planarity. Bürgi *et al*. (1974) extended this work to oxygen nucleophiles, for which it was found that the distortion was generally about one-third as large as the distortion that would be found with nitrogen as a nucleophile. It was concluded that the amount of distortion induced by the close approach of a nucleophile depends on the strength of the nucleophile (Bürgi, 1975).

There are two problems in deducing the amount of distortion expected at the reactive site carbonyl of an inhibitor from the work of Bürgi and co-workers. We might expect that O$^{\gamma}$ of Ser195 would be a stronger nucleophile than a typical oxygen nucleophile; how much stronger is difficult to say. In addition, the carbonyl-oxygen atom of the reactive bond sits in what has been termed the oxyanion hole

(Robertus *et al.*, 1972), where it forms two strong hydrogen bonds to main-chain NH groups. These hydrogen bonds enhance the polarization of the carbonyl bond, and would also be expected to encourage a distortion from planarity. The distortion observed for PTI:BTan, in which Ser195 of trypsin has been converted to dehydroalanine, has been attributed to this effect (Marquart *et al.*, 1983). As a result, one might expect to see a distortion of the reactive peptide group somewhat greater than that seen for typical oxygen nucleophiles, but the exact amount could not be predicted.

An additional problem arises from the small size of the distortion. If the distortion were of the magnitude found with a nitrogen nucleophile, it would be predicted that, for the nucleophile-electrophile distances observed in inhibitor:proteinase complexes, the reactive peptide carbonyl-carbon atoms would show out-of-plane displacements of less than 0.1Å. For well-refined structures at 1.8 to 1.9Å resolution, even the best-determined atoms have positional errors of the order of 0.1Å. As discussed above, the structure of OMTKY3:SGPB has estimated coordinate errors in this range. Since none of the other complexes of serine proteinases and their protein inhibitors has been determined at higher resolution, it is unlikely that any of them will have significantly more accurate coordinates.

One potential concern is that, even if a peptide bond distortion were significant compared to the coordinate errors, the application of geometrical restraints in refinement would tend to damp deviations from ideality. One way to test whether deviations are being masked is to refine the structure without planar restraints. When this experiment was performed on OMTKY3:SGPB, the small distortion observed in the restrained structure became somewhat larger, but it was still small compared to both the coordinate error and the deviations found for other residues (Read *et al.*, 1983).

The combined effect of real deviations from ideality, coordinate error and the application of restraints can be studied by examining the distribution of geometries for all the peptide bonds in a structure. The out-of-plane displacements observed in a number of high resolution, well-refined structures of inhibitor:proteinase complexes are summarized in Table III.3. [The coordinates for complexes other than those of OMTKY3 were obtained from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977).] All of these structures show a distortion in the expected direction. However, in none of these structures taken in isolation can the observation be considered significant compared to the coordinate error; none of the out-of-plane displacements is large compared to a reasonable coordinate error of 0.1Å, and none is strikingly large compared to the rms displacement in the structure.[2] Also, in each complex there are peptide bonds, presumably in normal environments, that show much

[2] The variations in the rms out-of-plane displacements for these structures probably result from variations in the tightness of the geometrical restraints.

Table III.3

Geometry of Scissile Peptides in Several Complexes

Structure	Distance $O^{\gamma}\cdots C$(Å)[1]	Out-of-plane displacement(Å)[2]			
		observed	predicted[3]	rms[4]	maximum (residue)[4]
OMTKY3:SGPB	2.70	0.066	0.084	0.027	0.080 (Cys161)
OMTKY3:CHT	2.95	0.014	0.060	0.021	0.070 (Ala401)
PTI:BT	2.68	0.089	0.086	0.066	0.177 (Asn233)
PTI:BTan[5]	—	0.070	0.0	0.070	0.286 (Ser147)
PTI:BTn	2.94	0.127	0.060	0.074	0.211 (Val27)
PTI:BTn:IV	2.80	0.082	0.073	0.075	0.264 (Ser147)
PSTI:BTn	2.67	0.104	0.087	0.058	0.179 (Pro161)

[1] Distance from O^{γ} of Ser195 to the carbonyl-carbon atom in the scissile peptide bond.
[2] Displacement of the carbonyl-carbon atom of the scissile peptide from the plane defined by the atoms to which it is bonded, i.e., C^{α}, O and N.
[3] Out-of-plane displacement calculated from equation (1) of Bürgi *et al.* (1973); the assumption is that O^{γ} of Ser195 is a nucleophile of similar strength to a nitrogen atom.
[4] The rms and maximum out-of-plane displacements are determined from all of the peptide bonds in the complex.
[5] In anhydrotrypsin, Ser195 has been converted to dehydroalanine, so that O^{γ} is not present.

greater distortions from planarity. Nonetheless, the combination of observations from several independent structures suggests that some distortion is being observed in the crystallographic experiments (Marquart *et al.*, 1983).

The fact that the reactive peptide carbonyl is in a special environment in the complexes can be deduced quite satisfactorily from protein crystallography. The observed distances of about 2.7Å between Ser195 O^{γ} and the carbonyl-carbon atom of residue P_1 are about 0.5Å shorter than the expected van der Waals contact distances of about 3.2Å. This is large compared to the coordinate error. From the work of Bürgi and co-workers, one would expect a corresponding distortion of the reactive peptide carbonyl. The results of protein crystallography suggest that a distortion exists, but the available structures are not sufficiently accurate to measure precisely the size of the distortion.

The Standard Mechanism of Inhibition

Most protein inhibitors of serine proteinases act by a standard mechanism (Laskowski and Kato, 1980). The basic elements of this mechanism have been recognized for some time. As early as 1954, it was speculated that ovomucoids are competitive inhibitors that form stable complexes with proteinases, but are acted upon slowly (Sri Ram *et al.*, 1954). The finding that inhibitors bind tightly to inactive anhydro-enzymes (Foster and Ryan, 1965) implied that inhibition did not require the formation of any covalent intermediate (Feinstein and Feeney, 1966). Studies of the kinetics of the reactions between enzyme and inhibitor showed that k_{cat}/K_M is characteristic of a very specific enzyme-substrate interaction, but that both k_{cat} and K_M are unusually small (Laskowski and Kato, 1980). This leads to very tight binding (low K_M) and slow hydrolysis (low k_{cat}).

The inhibitor and inhibitor:enzyme structures determined by crystallography are consistent with the general ideas of this mechanism. The mode of binding to the active site is quite similar to that expected for substrates. For example, Mitsui *et al*. (1979) pointed out the similarity between the conformation of SSI in the active site of subtilisin and that of the hypothetical subtilisin substrate proposed by Robertus *et al*. (1972). As another example, Figure III.4 shows a comparison of the reactive site loop of OMTKY3 in the active site of SGPB with a tetrapeptide product in the active site of SGPA (James *et al*., 1980b), a closely related proteinase. The tetrapeptide product can be considered as a virtual substrate for which oxygen exchange with the solvent is catalyzed.

The crystal structures allow one to rationalize the equilibrium constants and rate parameters that govern the interaction of enzyme and inhibitor. Since most of the

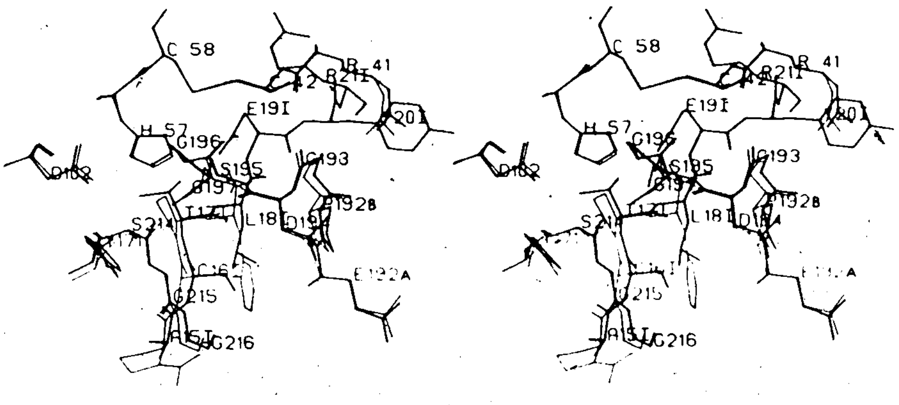


Figure III.4. Comparison of Interactions with Inhibitor and Product. The active site regions of OMTKY3:SGPB (thick lines) and the tetrapeptide product Ac-Pro-Ala-Pro-Phe-OH bound to SGPA (thin lines) were superimposed using a program of W. Bennett. Residues of OMTKY3:SGPB are labelled.

inhibitors are still functional when equilibrium has been reached, it is useful to consider the situation at equilibrium. In simplified form, the reaction between enzyme and inhibitor can be represented as

$$E + I \rightleftharpoons E \cdot I \rightleftharpoons E + I*$$

where I is the virgin inhibitor and I* is the cleaved, or modified inhibitor. The amount of complex E·I that is present at equilibrium is governed by

$$K_{assoc} = [E \cdot I]/([E][I + I*]) \qquad (1)$$

(Finkenstadt *et al.*, 1974). An inhibitor will be potent at equilibrium if it has a large value for K_{assoc}. The equilibrium constant for the hydrolysis of inhibitor is given by

$$K_{hyd} = [I*]/[I] \qquad (2)$$

Finally, the equilibrium constant for the association of virgin inhibitor with enzyme can be denoted by

$$K_a = [E \cdot I]/([E][I]) \qquad (3)$$

Some manipulation of equations (1)-(3) leads to the following:

$$K_{assoc} = K_a/(1 + K_{hyd}) \qquad (4)$$

This formulation separates inhibition into two logically distinct parts. A good inhibitor has a large value for K_a, which is determined by the interactions between E and I, and a small value for K_{hyd}, which is a function of the inhibitor structure alone. Some values for K_a are summarized in Table III.4.

Table III.4

Strength of Association of Some Inhibitor:Enzyme Complexes

Complex	$K_a (M^{-1})$	pH	Reference
OMTKY3:SGPB	5.6×10^{10}	8.3	Laskowski *et al.*, 1983
OMTKY3:CHT	3.2×10^{11}	8.3	Empie and Laskowski, 1982
OMTKY3:elastase	5.7×10^{10}	8.3	Empie and Laskowski, 1982
PTI:BT	1.6×10^{13}	8.0	Lazdunski *et al.*, 1974
PTI:BTn	4.3×10^{5}	8.0	Bode, 1979
PSTI:BT	1.0×10^{10}	8.0	Antonini *et al.*, 1983
PSTI:BTn	3.6×10^{4}	8.0	Antonini *et al.*, 1983

Maximizing K_a

It has long been recognized that a major reason for the tight binding of inhibitors is the fact that the free inhibitors have conformations complementary to the active sites of the enzymes'they inhibit (Huber *et al.*, 1974; Sweet *et al.*, 1974; Blow, 1974). (As discussed above, it is no longer believed that a covalent bond in a tetrahedral intermediate contributes to the strength of binding.) The binding energy of a substrate to an enzyme is the result of a large number of favourable and unfavourable terms. Up to a point, the balance sheet comparing favourable and unfavourable free energy terms would be similar for the binding of either a substrate or an inhibitor. For example, almost all of the interactions between OMTKY3 and SGPB involve residues P_6 through P_3' (100 out of 108 intermolecular contacts less than 4Å) and are thus possible for a small peptide substrate to achieve. However, a small peptide free in solution will adopt many conformations, only one of which is that required to bind to the enzyme active site, and this binding conformation does not necessarily have the lowest free energy. Therefore, on binding there are unfavourable free energy terms: a decrease in entropy from the loss of internal degrees of freedom, and probably an increase in free energy from selecting a conformation other than that of lowest energy. These unfavourable terms have no counterparts in the case of a rigid inhibitor that is already complementary to the enzyme (Huber *et al.*, 1974; Sweet *et al.*, 1974; Blow, 1974). As a result, the inhibitor has a considerably higher binding energy than a substrate.

More recently, it has become apparent that, although inhibitors are rigid compared to good substrates such as small peptides or floppy external loops of globular proteins, they are not absolutely rigid. In fact, as will be discussed below, a certain amount of flexibility can be important to inhibitor binding.

Minimizing K_{hyd}

In most inhibitors, K_{hyd} is close to unity at neutral pH (Finkenstadt *et al.*, 1974; Laskowski and Kato, 1980). It can be seen from equation (4) that K_{assoc} will not increase much if K_{hyd} is lower than unity. The region of the reactive site in inhibitors must have considerably less freedom to relax after hydrolysis than regions containing hydrolyzable bonds in globular proteins, which undergo almost complete hydrolysis (Finkenstadt *et al.*, 1974). It has been noted that in all of the inhibitors of known structure, the segment on the P_n' side of the reactive peptide is involved in β-sheet structure, and that there is often a disulfide bridge close to the reactive site on the P_n side (Bolognesi *et al.*, 1982). In addition, inhibitors of the PSTI (Kazal) family have a disulfide-linked cysteine at position P_6'. It is not necessary, however, for the reactive site loop to be covalently closed. Even when STI is cleaved by subtilisin at Met84I, it is still active as a trypsin inhibitor. If it is then converted to the modified inhibitor by cleavage at Arg63I, the segment Ile64I to Met84I does not dissociate, even though it is held only non-covalently (Laskowski *et al.*, 1974).

Additional evidence that disulfide bridges are not essential to the standard mechanism comes from the structures of inhibitors in the PI-1 family. The interaction of barley inhibitor CI-2 or eglin with subtilisin is equivalent to that seen in the complexes of the other serine proteinase inhibitors (McPhalen *et al.*, 1985a,b). Therefore, it is reasonable to assume that the standard mechanism applies equally to this inhibitor family. CI-2 and eglin lack disulfide bridges; the extensive β-sheet structure, which includes segments flanking the reactive site on both sides (McPhalen *et al.*, 1985a,b), should contribute to the necessary conformational constraints.

The covalent and non-covalent interactions that would be expected to reduce the conformational freedom of modified

OMTKY3, for example, can be seen clearly in Figure III.1(b). In addition to the disulfide bridges and β-sheet structure already mentioned, three hydrogen bonds involving Gly32I and Asn33I link the main-chain of residues near the reactive bond to the rest of the inhibitor.

Slow Approach to Equilibrium

For some inhibitors, like PSTI, the modified forms are more susceptible to additional cleavage than the virgin forms (Schneider *et al.*, 1974; Tschesche *et al.*, 1974), so that they become gradually inactivated. Presumably, there is some physiological advantage to the temporary nature of the inhibition rendered by inhibitors such as PSTI. Nevertheless, if equilibrium between virgin and modified inhibitor were approached too rapidly, such inhibitors would be quite ineffective. To a certain extent, the extreme strength of inhibitor binding sets a limit on the rate at which equilibrium can be approached. In addition, a specific barrier against hydrolysis has been suggested because of the observation that, for most inhibitor:enzyme systems, formation of the complex from modified inhibitor is slower than from intact inhibitor.

Several suggestions have been made for interactions that would slow the approach to equilibrium (Fujinaga *et al.*, 1982; Read *et al.*, 1983). However, most of these suggestions are difficult to reconcile with data showing that the size of the barrier to hydrolysis differs markedly for different enzymes interacting with the same inhibitor. In the simplified kinetic scheme

$$E + I \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} E\cdot I \underset{k_{on}*}{\overset{k_{off}*}{\rightleftharpoons}} E + I*$$

the ratio $k_{off}/k_{off}*$ is a measure of the relative heights of the activation energy barriers on the virgin and modified inhibitor sides of the reaction. For PTI interacting with the two enzymes CHT and BT, this ratio is equal to 2×10^5 and 2×10^2, respectively (data summarized in Quast *et al.*, 1978). Even more impressive are recent data involving complexes of OMTKY3 with various enzymes for which the ratio $k_{off}/k_{off}*$ varies over six orders of magnitude (Ardelt and Laskowski, 1985). It is fortunate that two of the more extreme examples involve the two enzymes for which the structures of complexes with OMTKY3 are known: CHT ($k_{off}/k_{off}* = 1.25\times10^6$) and SGPB (4.3). To rationalize this pronounced difference, based on two structures that reveal very similar inhibitor: enzyme interactions, is a challenging problem.

Such marked differences in the barrier to hydrolysis are hard to explain with mechanisms based only on intramolecular interactions within the inhibitor. These interactions are expected to be the same in complexes with different enzymes. For example, in both complexes with OMTKY3, hydrogen bonds are observed from the side-chain amide nitrogen atom of Asn33I to the carbonyl-oxygen atoms of Thr17I and Glu19I. In addition, the main-chain amide nitrogen atom of the scissile peptide forms a hydrogen bond to the side-chain carboxyl group of Glu19I. It has been proposed that these interactions contribute to an increased activation energy barrier to hydrolysis (Fujinaga *et al.*, 1982; Read *et al.*, 1983). The new kinetic parameters discussed above rule out this possibility. Similar considerations invalidate a proposal by Rühlmann *et al.* (1973) that a favourable hydrogen-bond from the NH of Ala16I to the carbonyl-oxygen atom of Gly36I in PTI would not be possible in the acyl-enzyme complex. The different behaviours of CHT and BT towards PTI in terms of the ratios of k_{off} to $k_{off}*$ cannot be explained by this intramolecular interaction because it should be present in both complexes.

An equivalent problem arises if one tries to invoke interactions with the active site residues of the serine proteinases to explain the large differences in $k_{off}/k_{off}*$ for different enzymes. These residues adopt extremely similar conformations in all of the serine proteinases of known structure. Thus, the fact that solvent is excluded from approaching the His57 residue of BT will be equally valid for the complexes of PTI with CHT, so that such proposals (Rühlmann *et al.*, 1973) cannot account for the kinetic differences between BT and CHT.

An explanation for the barrier to hydrolysis that is consistent with the kinetic data must involve properties and structural features that differ among the enzymes studied. One possibility would be that the dynamic properties of different enzymes result in some complexes being more rigid, hence less reactive, than others. Another possibility could be that the differences in the ratio $k_{off}/k_{off}*$ arise from altered strength of interactions on the P_n' side of the scissile bond. In the formation of the acyl-enzyme, the new N- and C-termini must move apart. This, presumably, would change the interactions between at least P_1' and S_1' (Rühlmann *et al.*, 1973), with an energy cost that could vary considerably among enzymes. For a normal substrate, the leaving group is expected to diffuse away from the active site, thus making room for a water molecule that will participate in the de-acylation step. In an inhibitor such as OMTKY3, covalent and non-covalent interactions prevent the loss of the leaving group and limit the possible change in conformation on the P_n' side. It is easy to imagine that the rearrangement of the P_n' residues necessary for de-acylation could be much more difficult with some enzymes, since there is considerable variability in the S_n' regions of serine proteinases (e.g., compare the two complexes of OMTKY3 in parts (a) and (b) of Figure III.3). This question could be addressed by modelling the acyl-enzyme complex, then performing extended numerical simulations of the dynamics of this complex.

Role of Flexibility in Inhibitors

As noted above, a conformationally labile inhibitor would probably be a good substrate. To take the other extreme, we can imagine an inhibitor that is rigidly fixed in the best conformation to bind to a particular enzyme. Because there would be no strain in the bound inhibitor and no change of entropy from loss of internal degrees of freedom, this inhibitor would bind very tightly to its cognate enzyme. However, the rigid inhibitor would not bind to another enzyme that differed in structure to any significant extent, whereas a more flexible inhibitor might be able to adapt to the active sites of several enzymes. These considerations suggest that one way of increasing specificity is to increase the rigidity of an inhibitor. In inhibitors with broad specificity such as the ovomucoids, there will be a trade-off between flexibility and strength of binding.

Flexibility is indicated in refined crystallographic structures by the thermal motion parameters, or *B* values. For free OMJPQ3 (Weber *et al.*, 1981; Papamokos *et al.*, 1982) and free OMSVP3 (Bode *et al.*, 1985), the *B* values at the reactive sites are among the highest in each molecule, indicating a considerable amount of flexibility in this region of the uncomplexed Kazal domains. In contrast, in each of OMTKY3:SGPB (Fujinaga *et al.*, 1982), OMTKY3:CHT, and PSTI:BTn (Bolognesi *et al.*, 1982), the *B* values of the reactive site loop are among the lowest in the complexes. Thus, the association of a Kazal domain with its cognate enzyme is accompanied by a large reduction in the conformational flexibility of the reactive site region of the free inhibitor. However, inhibitor binding to trypsin does not have a striking effect on the relative *B* values of the reactive loop in

PTI. This suggests a lack of flexibility in free PTI that may be related to its extremely large association constant (see Table III.4).

A comparison of the structures of OMTKY3 bound to SGPB and to α-chymotrypsin gives evidence in support of the role of flexibility. Figure III.3 shows the two complexes with the enzyme active sites in the same orientation. From this view, one would think that the binding mode differed significantly for the two enzymes. However, the binding interactions with the reactive site loop (Figure III.5) are virtually identical in the two complexes. There has been a conformational change (Figure III.6) that has the effect of altering the relative orientation of the reactive site loop (P_5-P_3') and the rest of the inhibitor.

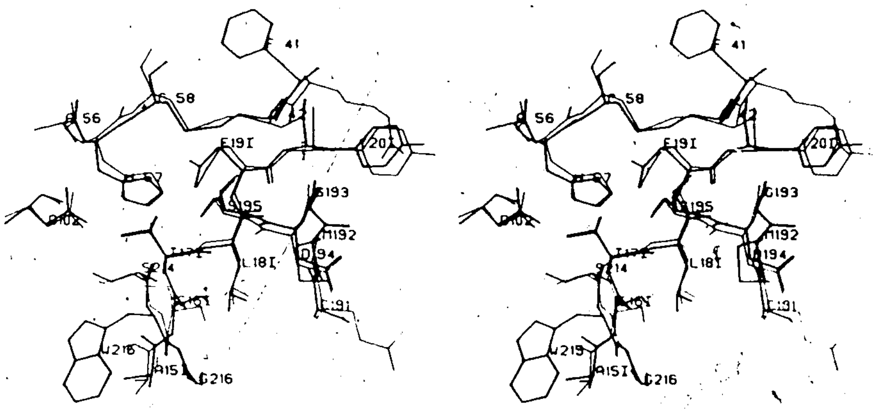


Figure III.5. Comparison of Active Site Regions of the OMTKY3 Complexes. OMTKY3:CHT is shown in thick lines and OMTKY3:SGPB in thin lines. Residues of OMTKY3:CHT are labelled. Main-chain atoms of residues 57, 191-195, 214-216 and 17I-19I were superimposed using a program of W. Bennett; the rms deviation for these 48 atoms was 0.21Å.

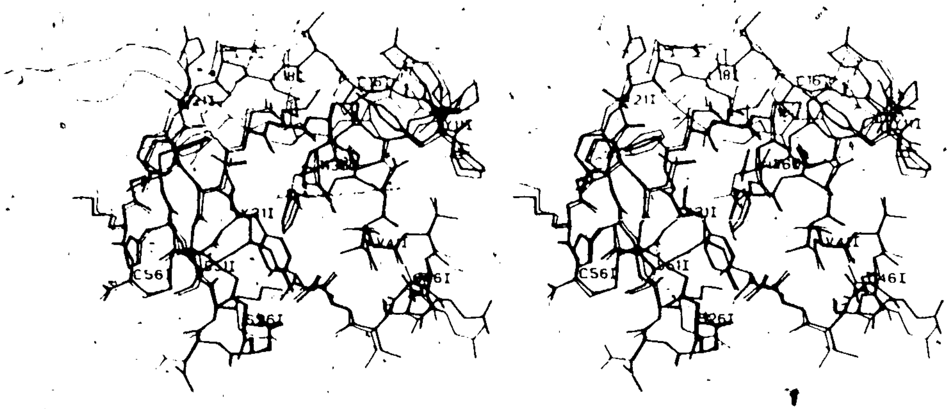


Figure III.6. Comparison of OMTKY3 Structures. The structures of OMTKY3 in its complexes with SGPB (thick lines) and with CHT (thin lines) were superimposed by a least-squares procedure comparing the main-chain atoms of residues involved in secondary structure (rms deviation for 112 atoms is 0.34Å). In both structures the N-terminus of OMTKY3 is disordered. Two additional residues are visible in the electron density of the complex with CHT.

The necessity for this conformational change can be demonstrated by a simple model-building experiment. If OMTKY3(B) is the structure of the inhibitor from the complex with SGPB and OMTKY3(C) the structure of the inhibitor from OMTKY3:CHT, then we can construct the hypothetical complexes OMTKY3(C):SGPB and OMTKY3(B):CHT by superimposing the two enzyme active sites, as for Figure III.5, and exchanging the two forms of OMTKY3. These hypothetical complexes are shown in Figure III.7.

In OMTKY3(C):SGPB [Figure III.7(a)], Lys13I would make unfavourable close contacts with Tyr171 and Gly172. In addition, Tyr20I and Arg21I would clash with the segment Thr39 to Arg41 on SGPB. In OMTKY3(B):CHT [Figure III.7(b)], the situation would be even worse. There would be close

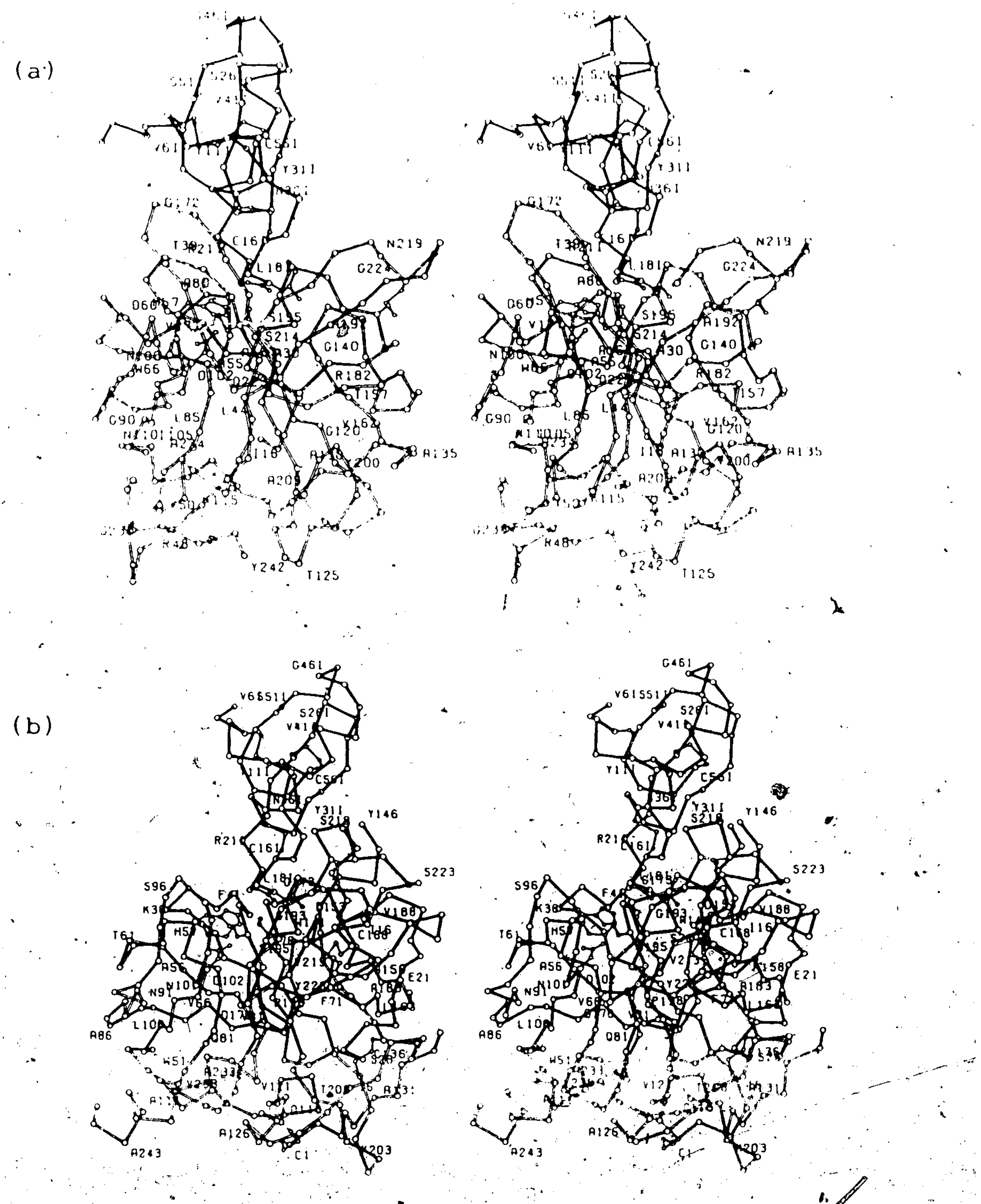


Figure III.7. C$^{\alpha}$-atom Representations of Hypothetical OMTKY3 Complexes. In both figures, OMTKY3 is shown with solid bonds and the enzyme with open bonds. Side-chains are shown for disulfide bridges, the catalytic residues His57, Asp102 and Ser195, and the P$_1$ residue Leu18I. The atoms comprising the scissile bond are also shown. Part (a) shows OMTKY3(C):SGPB, and (b) shows OMTKY3(B):CHT.

contacts of the inhibitor segment Pro14I-Cys16I with Trp172, Trp215 and Gly216, and of Lys29I with Tyr146. Asn36I would have a number of unfavourable contacts with Tyr146, Met192 and Ser218, and the position of Asn39I would conflict with that of Ser218. Therefore, a single rigid conformation for OMTKY3, either OMTKY3(B) or OMTKY3(C), will not permit binding to both SGPB and CHT. The broadness of specificity that this inhibitor displays requires some flexibility.

Model-building suggests the presence of additional flexibility in the OMTKY3 domain. This inhibitor has a large value of K_a for its interaction with porcine pancreatic elastase (Table III.4), which suggests that the binding interactions are similar to those in the OMTKY3:SGPB or OMTKY3:CHT complexes. However, an attempt to dock either OMTKY3(B) or OMTKY3(C) to elastase results in prohibitively close non-bonded contacts in the region of Arg217A of elastase. Thus, conformational adjustments in the inhibitor, and possibly in the enzyme, would be required to form the complex of OMTKY3 with elastase.

D. Comparison of SGPB Structures

The crystal of native SGPB and the crystal of its complex with OMTKY3 differ in crystallization conditions, in packing interactions, and in the presence or absence of the inhibitor molecule, all of which are expected to affect the conformation to some degree. A comparison of the two structures can indicate the nature and extent of these effects.

The level at which two structures can be compared depends on their accuracy. Both structures of SGPB are well-refined, high-resolution structures with estimated rms coordinate errors of 0.1 to 0.2Å, so that comparisons can be quite detailed. The native SGPB structure has been refined at 1.7Å resolution to an *R*-factor of 0.149 (L. Sawyer, A. R. Sielecki and M. N .G. James, unpublished).

When the two SGPB structures are compared by a least-squares superposition of all 1310 atoms (program of W. Bennett), the rms deviation in atomic positions is 0.58Å (for the 741 main-chain atoms, the rms deviation is 0.52Å; for the 569 side-chain atoms, 0.66Å). The relative orientation derived from this superposition has been used in a detailed structural comparison.

Effect of Crystal Packing

In order to study the effect of crystal packing contacts on structure, the symmetry operations for each crystal were applied to the appropriately oriented SGPB molecule from the other crystal. In this way, contacts that are present can be compared with those that would exist if the conformation were that of SGPB in the other crystal.

When such comparisons are made, one can see that differences in crystal packing environment are often accompanied by changes in van der Waals and hydrogen bonding interactions. It is generally evident that changes are necessary to avoid unfavourable contacts, but it is sometimes difficult to ascribe cause and effect.

Some of the largest conformational differences observed involve the C-terminal residues of SGPB. In the crystal of the native enzyme there is an intermolecular contact between the segment Gly172-Gly173 and the C-terminal residues Val241 and Tyr242 (Figure III.8). In the crystal of the complex, both segments are essentially exposed to solvent, although Tyr171 is in contact with the bound inhibitor (see Figure III.4 and Figure III.10 below). If SGPB with the

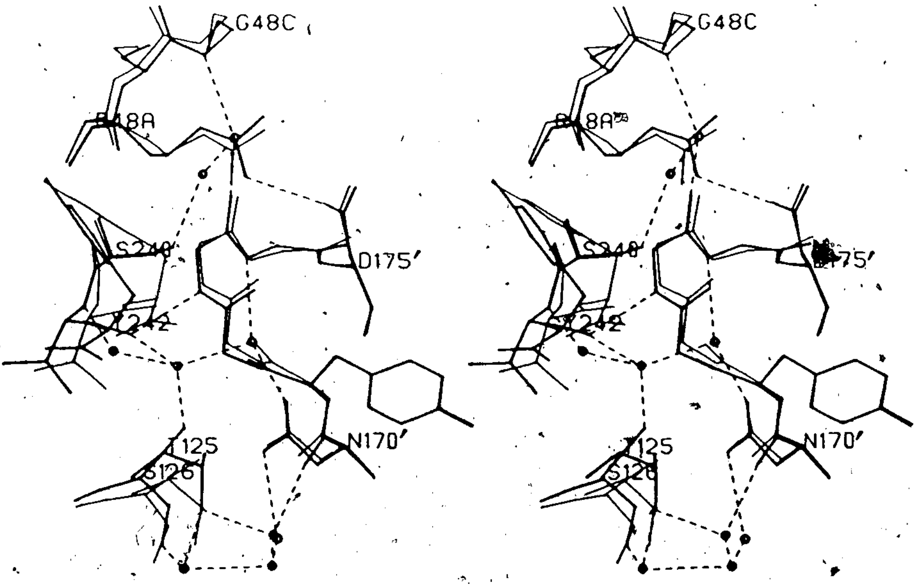


Figure III.8. Intermolecular Contact in Native SGPB Crystal Packing. The segment of the molecule denoted N170' to D175' is related by the following coordinate transformation to the molecule at x,y,z: x,y,-1+z. Thick lines in the representation indicate the structure observed in the native enzyme. Thin lines correspond to the superimposed structure of SGPB in the conformation of its complex with OMTKY3. Had the C-terminal residues Val241 and Tyr242 retained their orientation observed in the crystals of the complex, unacceptably short non-bonded contacts with Gly172' and Gly173' would occur (Table III.5). Only those hydrogen bonds (dashed lines) and solvent molecules (circles) involved in bridging the interface are shown.

conformation observed in the complex were placed in the crystal of the native enzyme, unfavourable van der Waals contacts would exist (Figure III.8). The conformation in the native structure avoids these unfavourable interactions and, in addition, allows Val241 O to form hydrogen bonds both to Gly173 N and to a water molecule involved in the solvent structure of the interface (Table III.5). The small shift of Gly172 and Gly173 towards the contacting protein molecule might be attributed to the inhibitor binding interaction at Tyr171 (see below). It is not clear why the C-terminus, instead of the diglycyl β-bend, adjusts to avoid the potential bad contacts.

The shift of the C-terminal residues propagates by intramolecular contacts to the segment from Lys115 to Ala127 (Figure III.9). Without concerted shifts of Lys115 and, especially, Gly117, prohibitively close contacts would result (Table III.6). These shifts lead to a reorganization of the whole segment, even those residues relatively far from the C-terminus. In fact, the largest main-chain difference between the two SGPB structures (2.8Å for Gly120 C$^{\alpha}$) occurs in the β-bend from Val119 to Gln122. As can be seen in Figure III.9, the pattern of hydrogen bonding involving the residues Lys115 to Ala127 is also somewhat altered. One of these hydrogen bonds, that from the carboxyl group of Tyr242 to Gly117 O in native SGPB, is perhaps a contributing factor in the reorganization. The existence of this hydrogen bond implies that the carboxy-terminus is protonated, which is

Table III.5

Intermolecular Contacts in the Crystal of Native SGPB

Atom 1	Atom 2[1]	Observed distance (Å)	Potential contact[2] (Å)	Shift of atom 1 (Å)	Shift of atom 2 (Å)
Val241 C	Gly173 C^{α}	4.3	3.4	1.1	0.5
Val241 C	Gly173 N	3.8	2.8	1.1	0.4
Val241 O	Gly172 C	3.7	3.0	1.4	0.5
Val241 O[3]	Gly173 N	2.9	1.9	1.4	0.4
Val241 O	Gly173 C^{α}	3.7	2.6	1.4	0.5
Tyr242 C^{α}	Gly172 C^{α}	4.3	3.2	2.0	0.5
Tyr242 C^{α}	Gly172 C	4.5	3.0	2.0	0.5
Tyr242 C^{α}	Gly173 N	4.1	2.8	2.0	0.4
Tyr242 C^{α}	Gly173 C^{α}	4.8	3.5	2.0	0.5
Tyr242 C^{β}	Gly172 C	3.9	3.2	1.9	0.5
Tyr242 C^{β}	Gly173 N	3.6	2.9	1.9	0.4
Tyr242 C^{β}	Gly173 C^{α}	4.0	3.0	1.9	0.5

[1] Coordinates generated by the symmetry operation x,y,-1+z.
[2] Distances calculated with coordinates of both atoms from the superimposed structure of SGPB in the complex.
[3] Hydrogen-bonded interaction in the crystal of native SGPB.

possible for the native SGPB structure, determined at a pH of 4.2, but not for the structure of the complex (pH 6.3). It would appear from Figure III.9 that the large movement of the β-bend, Val119 to Glu122, is in turn propagated through to Ser126, as well as to the neighboring β-bend from Ser201 to Arg208.

The comparison of the two structures of SGPB shows that, in the regions of intermolecular contact, the differences are not large and tend to be smaller than those discussed in the above example. For SGPB there does not seem to be much propagation of the conformational changes from the surface residues to the more internal residues of the structural core.

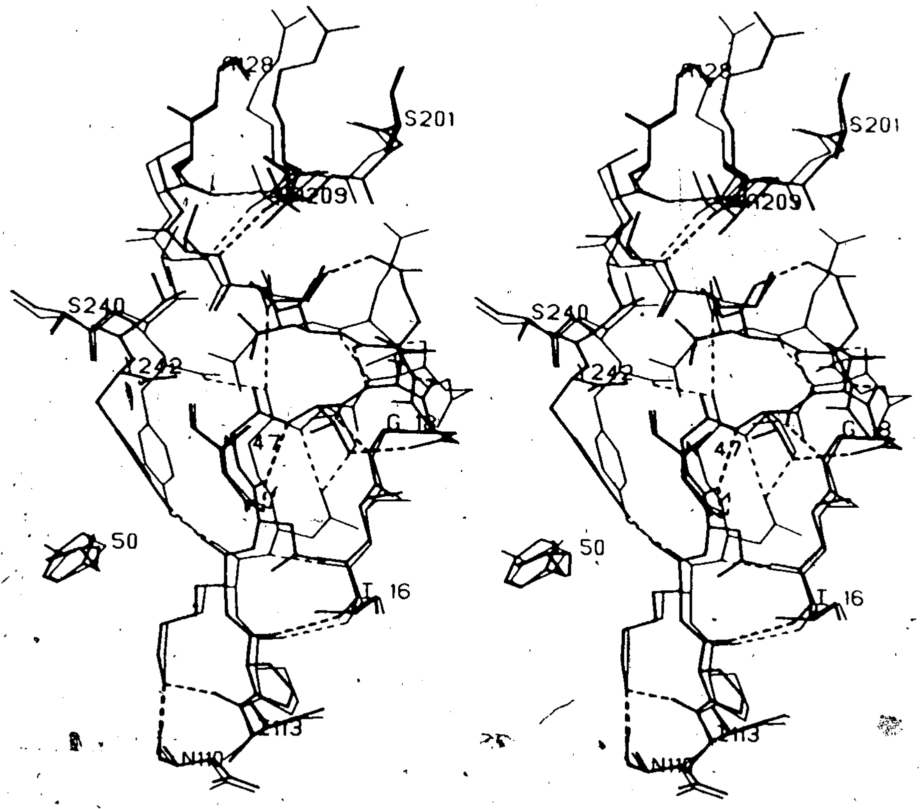


Figure III.9. Propagation of Shift of SGPB C-terminus. The shift of the C-terminal residues Val241 and Tyr242 is propagated to the peptide chain Lys115 to Ser126. Thick lines represent the conformation of SGPB in the crystal of its complex with OMTKY3; thin lines indicate the native structure. Dashed lines represent hydrogen bonds. The major differences are seen for the β-bend Val119-Glu122, where conformational differences up to 2.8Å occur. This β-bend makes few intramolecular contacts and is thus relatively free to move.

Effect of Inhibitor Binding

The residues of SGPB that interact with OMTKY3 are mostly exposed to solvent in the crystal of the native enzyme. Therefore, their conformational changes should closely parallel the movements that occur on binding in solution. Figure III.10 shows the contact region, with the

Table III.6

Contacts Leading to Propagation of Shift in SGPB

Atom 1	Atom 2	Observed distance(Å)	Potential contact(Å)[1]	Shift of atom 2(Å)
Tyr242 C	Gly117 C^{α}	3.7	2.9	1.0
Tyr242 C	Gly117 C	3.9	3.1	0.8
Tyr242 C	Gly117 O	3.3	2.9	0.6
Tyr242 O	Gly117 C^{α}	3.2	2.6	1.0
Tyr242 O	Gly117 C	3.3	2.7	0.8
Tyr242 O	Gly117 O	2.6	2.1	0.5
Tyr242 C^{ζ}	Lys115 O	3.5	2.9	0.7
Tyr242 $C^{\epsilon 2}$	Lys115 O	3.3	2.8	0.7
Tyr242 $C^{\epsilon 2}$	Asp116 C	4.5	3.5	0.8
Tyr242 $C^{\epsilon 2}$	Gly117 N	3.8	3.0	0.9
Tyr242 $C^{\epsilon 2}$	Gly117 C^{α}	3.7	3.1	1.0
Tyr242 $C^{\delta 2}$	Gly117 C^{α}	3.6	2.8	1.0
Tyr242 O^{η}	Lys115 C	3.7	3.1	0.6

[1] Distances calculated with coordinates of atom 2 from the superimposed structure of SGPB in the complex.

native SGPB structure superimposed on that of the complex. Contact distances from OMTKY3 to the SGPB molecule in both conformations are given in Table III.7.

The conformational changes of the catalytic residues, His57 and Ser195, are important in the hydrolytic mechanism. As discussed above, Ser195 O^{γ} is only 2.7Å away from Leu181 C, in what appears to be an attractive interaction. Nonetheless, upon binding, this residue moves slightly away from the inhibitor. As well, the side-chain of His57 rotates away from close contacts with Thr171. The net result of these two movements is the formation of a strong hydrogen bond between His57 $N^{\epsilon 2}$ and Ser195 O^{γ}. It has been proposed that the formation of this hydrogen bond is important in activating O^{γ} as a nucleophile and in allowing the

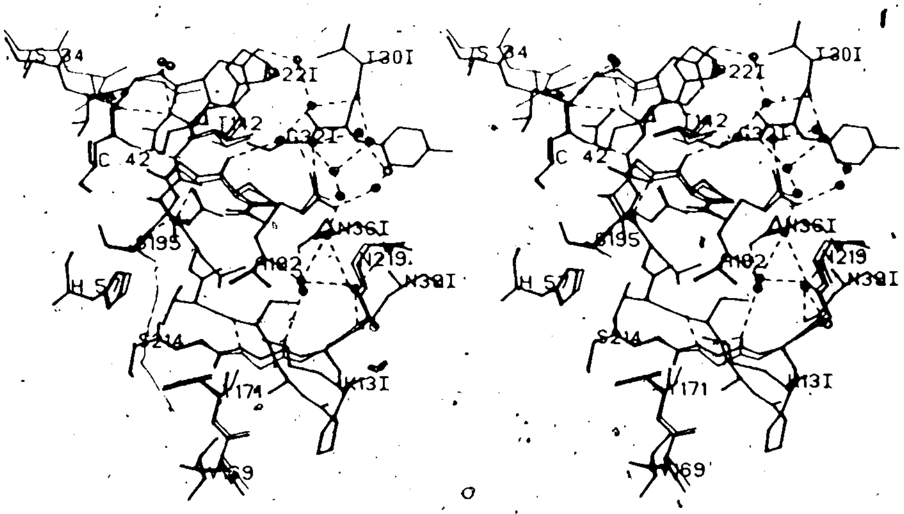


Figure III.10. Intermolecular Contacts Between OMTKY3 and SGPB. Thick lines correspond to the observed structure in the crystal of the complex; thin lines represent the suitably transformed coordinates of native SGPB for comparison. Conformational changes in SGPB induced upon inhibitor binding are relatively small (see text and Table III.7). Only those hydrogen bonds (dashed lines) and solvent molecules (circles) involved in bridging the contact are included in the figure.

transfer of the proton from Ser195 via His57 to the leaving group amide during hydrolysis (Huber and Bode, 1978).

In the binding of OMTKY3 to SGPB, the conformational changes are mostly small, though significant. This observation favours the lock and key model of enzyme-substrate interaction (Fischer, 1894). On the other hand, the small conformational changes in the catalytic residues could have quite profound effects on the mechanism of hydrolysis.

Table III.7

Intermolecular Contacts Between OMTKY3 and SGPB

SGPB	OMTKY3	Observed distance (Å)	Potential contact (Å)	Shift of SGPB atom (Å)	Comments
Thr39 C$^{\gamma 2}$	Arg21I N$^{\eta 1}$	4.1	2.9	2.7	main-chain and side-chain movements alleviate close contact; note, however, that the preceding residue, Ser34, forms an H-bond to a neighbouring molecule of SGPB in the crystal of the complex.
His57 C$^{\epsilon 1}$	Thr17I C$^{\beta}$	3.8	3.5	0.4	His57 side-chain reorientation avoids close contact, and makes H-bond to Ser195 possible.
His57 N$^{\epsilon 2}$	Thr17I C$^{\beta}$	3.6	3.3	0.5	
His57 C$^{\delta 2}$	Thr17I C$^{\beta}$	3.6	3.5	0.3	
Tyr171 C	Lys13I C$^{\delta}$	4.0	3.6	0.5	main-chain shift relieves bad contacts, while allowing H-bond formation.
Tyr171 O	Lys13I C$^{\delta}$	3.5	2.9	0.7	
Tyr171 O	Lys13I C$^{\epsilon}$	3.8	3.3	0.7	
Tyr171 O^{2}	Lys13I N$^{\zeta}$	2.9	2.6	0.7	
Pro192B C$^{\alpha}$	Leu18I C	3.9	3.7	0.4	several unfavourable contacts relieved by main-chain shift; note, however, that this segment lies near a 2-fold axis in the native crystal, and there is a carboxyl-carboxylate interaction between Glu192A and its 2-fold related mate (Sawyer and James, 1982); also, Thr142, which is H-bonded to Pro192B, shifts to avoid contacts with the 2-fold related molecule (Figure III.10).
Pro192B C	Leu18I O	3.5	3.2	0.4	
Gly193 N	Leu18I C	3.6	3.2	0.5	
Gly193 N^{2}	Leu18I O	2.6	2.2	0.5	
Gly193 C$^{\alpha}$	Leu18I O	3.4	2.9	0.6	
Gly193 C	Leu18I O	3.5	3.3	0.3	

(Table III.7 continued)

Gly215 C^{α}	Leu18I C^{γ}	3.9	3.7	0.4	main-chain shift alleviates bad contacts.
Gly215 C^{α}	Leu18I $C^{\delta 1}$	3.8	3.5	0.4	
Gly216 C^{α}	Ala15I C^{β}	4.0	3.4	0.6	
Gly216 C	Ala15I C^{β}	3.8	3.5	0.4	
Gly216 O	Ala15I C^{α}	3.5	3.2	0.3	

[1] Distances calculated with coordinates of the superimposed SGPB native structure.
[2] Hydrogen-bonded interaction in the complex.

Bibliography

Antonini, E., Ascenzi, P., Bolognesi, M., Gatti, G., Guarneri, M., & Menegatti, E. (1983) *J. Mol. Biol. 165*, 543-558.

Ardelt, W., & Laskowski, M. Jr. (1985) *Biochemistry 24*,5313-5320.

Baillargeon, M. W., Laskowski, M. Jr., Neves, D. E., Porubcan, M. A., Santini, R. E., & Markley, J. L. (1980) *Biochemistry 19*, 5703-5710.

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

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

Birktoft, J. J., & Blow, D. M. (1972) *J. Mol. Biol. 68*, 187-240.

Blevins, R. A., & Tulinsky, A. (1985) *J. Biol. Chem. 260*, 4264-4275.

Blow, D. M. (1974) in *Bayer Symp. V, Proteinase Inhibitors* (Fritz, H., Tschesche, H., Greene, L. J., & Truscheit, E., Eds.) pp 677-678, Springer-Verlag, Berlin.

Bode, W. (1979) *J. Mol. Biol. 127*, 357-374.

Bode, W., Schwager, P., & Huber, R. (1978) *J. Mol. Biol. 118*, 99-112.

Bode, W., Walter, J., Huber, R., Wenzel, H. R., & Tschesche, H. (1984) *Eur. J. Biochem. 144*, 185-190.

Bode, W., Epp, O., Huber, R., Laskowski, M. Jr., & Ardelt, W. (1985) *Eur. J. Biochem. 147*, 387-395.

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

Bürgi, H. B. (1975) *Angew. Chemie 14*, 460-473.

Bürgi, H. B., Dunitz, J. D., & Shefter, E. (1973) *J. Amer. Chem. Soc. 95*, 5065-5067.

Bürgi, H. B., Dunitz, J. D., & Shefter, E. (1974) *Acta Cryst. B30*, 1517-1527.

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

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* (Kasper, J. S., & Lonsdale, K., Eds.) Vol II, pp 318-340, Kynoch Press, Birmingham, England.

Delbaere, L. T. J., Hutcheon, W. L. B., James, M. N. G., & Thiessen, W. E. (1975) *Nature (London) 257*, 758-763.

Empie, M. W., & Laskowski, M. Jr. (1982) *Biochemistry 21*, 2274-2284.

Feinstein, G., & Feeney, R. E. (1966) *J. Biol. Chem. 241*, 5183-5189.

Finkenstadt, W. R., Hamid, M. A., Mattis, J. A., Schrode, J., Sealock, R. W., Wang, D., & Laskowski, M. Jr. (1974) in *Bayer Symp. V, Proteinase Inhibitors* (Fritz, H., Tschesche, H., Greene, L. J., & Truscheit, E., Eds.) pp 389-411, Springer-Verlag, Berlin.

Fischer, E. (1894) *Chem. Ber. 27*, 2985-2993.

Foster, R. J., & Ryan, C. A. (1965) *Fed. Proc. 24*, 473.

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., & Kauffman, D. L. (1966) *Biochem. J. 101*, 229-231.

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.

Hirono, S., Akagawa, H., Mitsui, Y., & Iitaka, Y. (1984) *J. Mol. Biol. 178*, 389-413.

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.

Bibliography

Antonini, E., Ascenzi, P., Bolognesi, M., Gatti, G., Guarneri, M., & Menegatti, E. (1983) *J. Mol. Biol. 165*, 543-558.

Ardelt, W., & Laskowski, M. Jr. (1985) *Biochemistry 24*,5313-5320.

Baillargeon, M. W., Laskowski, M. Jr., Neves, D. E., Porubcan, M. A., Santini, R. E., & Markley, J. L. (1980) *Biochemistry 19*, 5703-5710.

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

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

Birktoft, J. J., & Blow, D. M. (1972) *J. Mol. Biol. 68*, 187-240.

Blevins, R. A., & Tulinsky, A. (1985) *J. Biol. Chem. 260*, 4264-4275.

Blow, D. M. (1974) in *Bayer Symp. V, Proteinase Inhibitors* (Fritz, H., Tschesche, H., Greene, L. J., & Truscheit, E., Eds.) pp 677-678, Springer-Verlag, Berlin.

Bode, W. (1979) *J. Mol. Biol. 127*, 357-374.

Bode, W., Schwager, P., & Huber, R. (1978) *J. Mol. Biol. 118*, 99-112.

Bode, W., Walter, J., Huber, R., Wenzel, H. R., & Tschesche, H. (1984) *Eur. J. Biochem. 144*, 185-190.

Bode, W., Epp, O., Huber, R., Laskowski, M. Jr., & Ardelt, W. (1985) *Eur. J. Biochem. 147*, 387-395.

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

Bürgi, H. B. (1975) *Angew. Chemie 14*, 460-473.

Bürgi, H. B., Dunitz, J. D., & Shefter, E. (1973) *J. Amer. Chem. Soc. 95*, 5065-5067.

Bürgi, H. B., Dunitz, J. D., & Shefter, E. (1974) *Acta Cryst. B30*, 1517-1527.

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

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* (Kasper, J. S., & Lonsdale, K., Eds.) Vol II, pp 318-340, Kynoch Press, Birmingham, England.

Delbaere, L. T. J., Hutcheon, W. L. B., James, M. N. G., & Thiessen, W. E. (1975) *Nature (London) 257*, 758-763.

Empie, M. W., & Laskowski, M. Jr. (1982) *Biochemistry 21*, 2274-2284.

Feinstein, G., & Feeney, R. E. (1966) *J. Biol. Chem. 241*, 5183-5189.

Finkenstadt, W. R., Hamid, M. A., Mattis, J. A., Schrode, J., Sealock, R. W., Wang, D., & Laskowski, M. Jr. (1974) in *Bayer Symp. V, Proteinase Inhibitors* (Fritz, H., Tschesche, H., Greene, L. J., & Truscheit, E., Eds.) pp 389-411, Springer-Verlag, Berlin,

Fischer, E. (1894) *Chem. Ber. 27*, 2985-2993.

Foster, R. J., & Ryan, C. A. (1965) *Fed. Proc. 24*, 473.

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., & Kauffman, D. L. (1966) *Biochem. J. 101*, 229-231.

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.

Hirono, S., Akagawa, H., Mitsui, Y., & Iitaka, Y. (1984) *J. Mol. Biol. 178*, 389-413.

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.

Huber, R., Bode, W., Kukla, D., Kohl, U., & Ryan, C. A. (1975) *Biophys. Struct. Mechanism 1*, 189-201.

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

James, M. N. G., Brayer, G. D., Delbaere, L. T. J., Sielecki, A. R., & Gertler, A. (1980a) *J. Mol. Biol. 139*, 423-438.

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

Kato, I., Kohr, W. J., & Laskowski, M. Jr. (1978) *Proc. FEBS Meet. 47*, 197-206.

Laskowski, M. Jr., & Kato, I. (1980) *Ann. Rev. Biochem. 49*, 593-626.

Laskowski, M. Jr., Kato, I., Leary, T. R., Schrode, J., & Sealock, R. W. (1974) in *Bayer Symp. V, Proteinase Inhibitors* (Fritz, H., Tschesche, H., Greene, L. J., & Truscheit, E., Eds.) pp 597-611, Springer-Verlag, Berlin.

Laskowski, M. Jr., Tashiro, M., Empie, M. W., Park, S. J., Kato, I., Ardelt, W., & Wieczorek, M. (1983) in *Proteinase Inhibitors: Medical & Biological Aspects* (Katunuma, N., Umezawa, H., & Holzer, H., Eds.) pp 55-68, Japan Scientific Societies Press, Tokyo/Springer-Verlag, Berlin.

Lazdunski, M., Vincent, J.-P., Schweitz, H., Péron-Renner, M., & Pudles, J. (1974) in *Bayer Symp. V, Proteinase Inhibitors* (Fritz, H., Tschesche, H., Greene, L. J., & Truscheit, E., Eds.) pp 420-431, Springer-Verlag, Berlin.

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

Marquart, M., Walter, J., Deisenhofer, J., Bode, W., & Huber, R. (1983) *Acta Cryst. B39*, 480-490.

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

McPhalen, C. A., Svendsen, I., Jonassen, I., & James, M. N. G. (1985a) *Proc. Natl. Acad. Sci. U.S.A.*, in press.

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

Mitsui, Y., Satow, Y., Watanabe, Y., & Iitaka, Y. (1979) *J. Mol. Biol. 131*, 697-724.

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

Papamokos, E., Weber, E., Bode, W., Huber, R., Empie, M. W., Kato, I., & Laskowski, M. Jr. (1982) *J. Mol. Biol. 158*, 515-537.

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

Quast, U., Engel, J., Steffen, E., Tschesche, H., & Kupfer, S. (1978) *Eur. J. Biochem. 86*, 353-360.

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

Richarz, R., Tschesche, H., & Wüthrich, K. (1980) *Biochemistry 19*, 5711-5715.

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

Rossmann, M. G. (1973) *The Molecular Replacement Method* International Science Review 13, Gordon & Breach, New York.

Rühlmann, A., Kukla, D., Schwager, P., Bartels, K., & Huber, R. (1973) *J. Mol. Biol. 77*, 417-436.

Sawyer, L., & James, M. N. G. (1982) *Nature (London) 295*, 79-80.

Schneider, S. L., Stasiuk, L., & Laskowski, M. Sr. (1974) in *Bayer Symp. V, Proteinase Inhibitors* (Fritz, H., Tschesche, H., Greene, L. J., & Truscheit, E., Eds.) pp 223-234, Springer-Verlag, Berlin.

Sri Ram, J., Terminiello, L., Bier, M., & Nord, F. F. (1954) *Arch. Biochem. Biophys. 52*, 451-463.

Sweet, R. M., Wright, H. T., Janin, J., Chothia, C. H., & Blow, D. M. (1974) *Biochemistry 13*, 4212-4228.

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

Tschesche, H., Reidel, G., & Schneider, M. (1974) in *Bayer Symp. V, Proteinase Inhibitors* (Fritz, H., Tschesche, H., Greene, L. J., & Truscheit, E., Eds.) pp 235-242, Springer-Verlag, Berlin.

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

Weber, E., Papamokos, E., Bode, W., Huber, R., Kato, I., & Laskowski, M. Jr. (1981) *J. Mol. Biol. 149*, 109-123.

IV. Electron Density Map Coefficients[1]

Accurate phase probabilities are important for combining independent sources of phase information (Rossmann and Blow, 1961; Hendrickson and Lattman, 1970) or for calculating probability-weighted electron density maps (Blow and Crick, 1959). Woolfson (1956) and Sim (1959, 1960) derived expressions for phase probabilities from partial structures for centric and non-centric structure factors respectively. Srinivasan and co-workers extended this work to include coordinate errors in the partial structure (Srinivasan and Ramachandran, 1965; Srinivasan, 1966).

Two problems arise when one attempts to apply the results on phase probabilities to the calculation of electron density maps using partial structure phases. The first problem is that of estimating either Σ_Q or σ_A from the observed and calculated structure factor amplitudes, in order to obtain accurate phase probabilities. The parameter Σ_Q, which measures the amount of missing scattering matter, is used in the expressions of Woolfson (1956) and Sim (1959,1960); σ_A, which is a combined measure of the completeness and the accuracy of the partial structure, is required for the expressions of Srinivasan (1966). (These terms and others are defined in Table IV.1, and the relationships among the important structure factor vectors are illustrated in Figure IV.1.) Part A deals with the

[1] A version of this chapter has been accepted for publication [Read, R. J. (in press) *Acta Crystallographica*, Section A].

Table IV.1

Definitions of Terms and Notation

Term	Definition
$\overline{x}$	= mean value of x
$<x>$	= expected value, or probability weighted average, of x
$\mathbf{F}_N$	= $\sum_{j=1}^{P} f_j \exp(2\pi i \mathbf{s}\cdot\mathbf{r}_j) + \sum_{j=P+1}^{N} f_j \exp(2\pi i \mathbf{s}\cdot\mathbf{r}_j)$, where s is the reciprocal lattice vector ($\lvert s\rvert=2\sin\theta/\lambda$) and the $\mathbf{r}_j$ are the atomic coordinates (in Å) = $\mathbf{F}_P + \mathbf{F}_Q$, where the P atoms constitute the partial structure and the Q atoms the missing structure = $\lvert\mathbf{F}_N\rvert\exp(i\alpha_N)$
$\mathbf{F}_P^C$	= $\sum_{j=1}^{P} f_j \exp[2\pi i \mathbf{s}\cdot(\mathbf{r}_j+\Delta\mathbf{r}_j)]$ where $\Delta\mathbf{r}_j$ are positional errors = structure factor of partial structure with errors
D	= $<\cos(2\pi \mathbf{s}\cdot\Delta\mathbf{r})>$ (Luzzati, 1952)
ϵ	= correction factor for expected intensity in reciprocal lattice zone
Σ_N	= $\sum_{j=1}^{N} f_j^2 = <\lvert\mathbf{F}_N\rvert^2/\epsilon>$
$\mathbf{E}_N$	= $\mathbf{F}_N/(\epsilon\Sigma_N)^{1/2}$
σ_A	= $D(\Sigma_P/\Sigma_N)^{1/2}$
m	= $<\cos(\alpha_N-\alpha_P^C)>$ = $\frac{I_1(X)}{I_0(X)}$ for non-centric reflections, where I_0 and I_1 are the zero and first order modified Bessel functions, respectively, or

(Table IV.1 continued)

$= \tanh(X/2)$ for centric reflections, where

$$X = \frac{2|F_N||F_P|}{\epsilon\Sigma_Q}$$ for a partial structure with no errors (Woolfson, 1956; Sim, 1959, 1960)

$$= \frac{2\sigma_A|E_N||E_P^c|}{1-\sigma_A{}^2}$$ for a partial structure with errors (Srinivasan, 1966)

'Some terms are not given explicitly, but are analogous to terms defined here.

estimation of these parameters and with the evaluation of their associated phase probability distributions. The second problem, which is treated in part B, is that of minimizing the bias towards the model in an electron density map using model or combined phases.

A. Estimating Phase Probabilities for Partial Structures

Estimating Σ_Q

It has been common to assume that the probability expressions for partial structures with no coordinate errors provide a reasonable approximation for the case with errors. These expressions require an estimate for Σ_Q, modified perhaps to include a contribution from errors (Rossmann and Blow, 1961). Because of thermal motion and the finite size of atoms, Σ_Q is a function of resolution and is generally estimated for several resolution ranges. Blundell and Johnson (1976, p. 418) suggest the mean square deviation

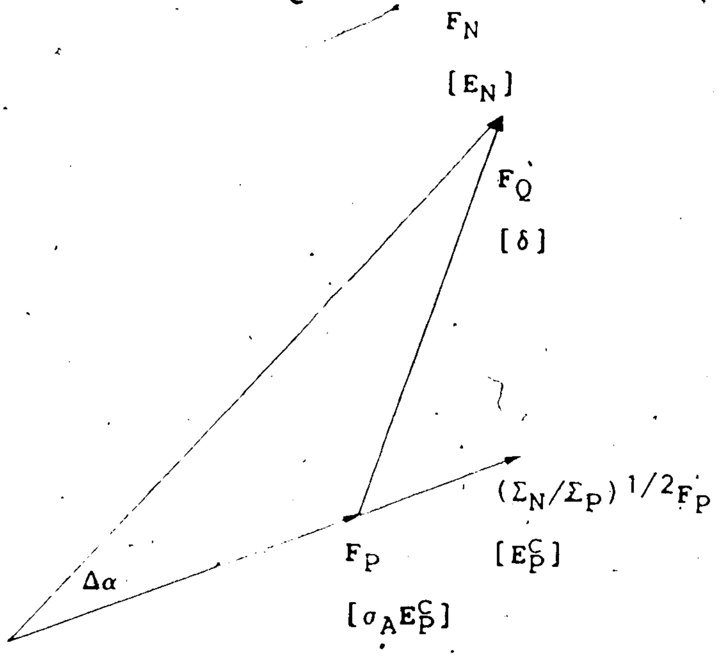


Figure IV.1. Phase Triangles for Probability Distributions. The vector relationships important for partial structure phase probabilities are illustrated as an Argand diagram. For a perfect partial structure (Sim, 1959, 1960), the expected magnitude and direction of the vector F_Q is independent of F_P, and $F_N=F_P+F_Q$. The labels in square brackets indicate that a similar relationship exists among the normalized variables E_N, $\sigma_A E_P^c$ and δ (Srinivasan and Ramachandran, 1965), with δ being independent of $\sigma_A E_P^c$. The figure of merit is equal to $<\cos(\Delta\alpha)>$.

between the structure factor amplitudes for the complete and partial structures, $\overline{(|F_N|-|F_P|)^2}$, as an empirical estimate. However, when F_Q, the missing atom structure factor, is small compared to and independent in direction of F_P (see Figure IV.1), $\overline{(|F_N|-|F_P|)^2}$ is a measure of the variance of each component of F_Q, the components in-phase and out-of-phase with F_P (Henderson and Moffat, 1971). Therefore, a better estimate is given by

$$\Sigma_Q = <|F_Q|^2/\epsilon> \simeq \overline{n(|F_N|-|F_P|)^2/\epsilon} \qquad (1)$$

where n = 2 for non-centric reflections and 1 for centric reflections, which have no out-of-phase component for F_Q. The factor ϵ corrects for the difference in expected intensity for different reciprocal lattice zones. Bricogne (1976) suggests

$$\Sigma_Q \simeq \overline{||F_N|^2-|F_P|^2|}$$

Including a correction for expected intensity, this becomes

$$\Sigma_Q \simeq \overline{||F_N|^2-|F_P|^2|/\epsilon} \qquad (2)$$

This expression is used most commonly. Finally, for non-centric data

$$<|F_Q|^2> = |F_N|^2+|F_P|^2-2|F_N||F_P| \; I_1(X)/I_0(X)$$

where

$$X = \frac{2|F_N||F_P|}{\Sigma_Q}$$

(Srinivasan, 1968). Nixon and North (1976) note this and solve the equation

$$\sum_h <|F_Q|^2> = \sum_h \Sigma_Q = \sum_h [|F_N|^2+|F_P|^2-2|F_N||F_P|I_1(X)/I_0(X)]$$

for Σ_Q by numerical methods. Extending this to include centric data and again including the factor ϵ, this becomes

$$\sum_h <|F_Q|^2/\epsilon> = \sum_h \Sigma_Q = \sum_h [(|F_N|^2+|F_P|^2-2m|F_N||F_P|)/\epsilon] \qquad (3)$$

where m is the appropriate expression for the figure of merit of centric or non-centric data (see Table IV.1). It is instructive to test these methods on calculated data where the correct values of Σ_Q and of the phase error are known.

Any reliable method for estimating Σ_Q should work in the ideal case of a perfect partial structure. Such a case was modelled by taking as $\mathbf{F}_N$ the calculated structure factors for *Streptomyces griseus* trypsin (SGT) at cycle 78 of least squares refinement when the *R*-factor was 0.159 (see Chapter II). About 30% of the atoms were removed randomly to give a partial structure from which $\mathbf{F}_P$ was calculated. This test data set will be referred to as TD1. The correct value of Σ_Q was calculated as a function of resolution from the scattering factors of the missing atoms. Estimates of Σ_Q were calculated using equations (1), (2) and (3) (with the appropriate values of ϵ for the different zones of the space group $C222_1$). In Figure IV.2 the correct values of Σ_Q and the three sets of estimates are shown; in Figure IV.3 $\overline{\cos(\alpha_N-\alpha_P)}$ is compared to $\overline{m}$ calculated from each set of estimates. Equation (3) gives the best results, while equations (1) and (2) lead to a slight underestimate and a large overestimate respectively of Σ_Q.

By the time one has accurate coordinates in protein crystallography, however, there is comparatively little need for phase probabilities. In the early stages of developing a structure, there are large coordinate errors. A more realistic set of test data (referred to as TD2) was constructed, using as $\mathbf{F}_P^c$ the structure factors calculated for SGT at cycle 7 of refinement. At cycle 7, the *R*-factor was 0.455 to 1.7Å resolution. Parts of the structure were

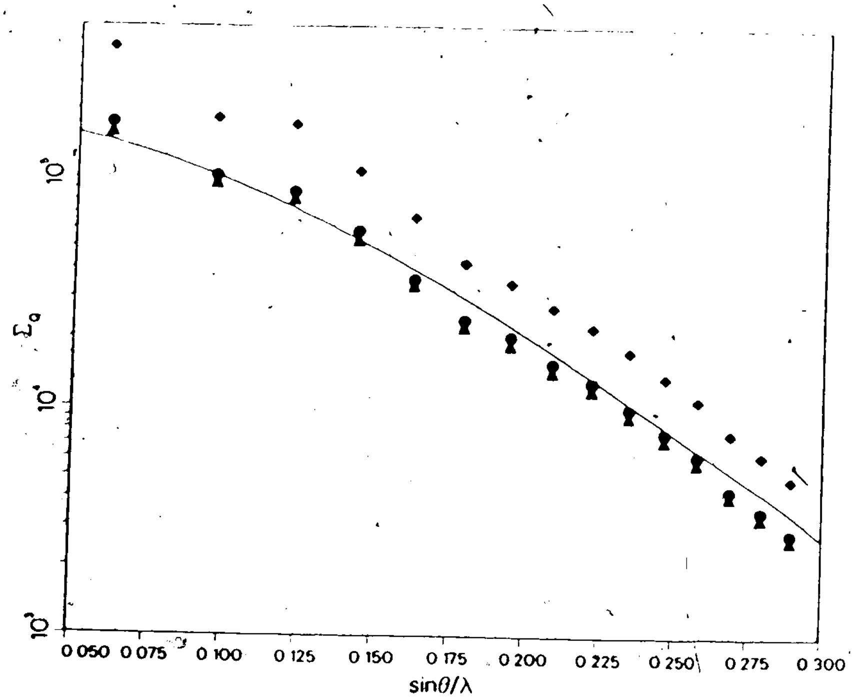


Figure IV.2. Estimates of Σ_Q for Perfect Partial Structure. Methods for estimating Σ_Q were evaluated for the case of a perfect partial structure (TD1). Estimates were calculated for 15 ranges of equal width in $(\sin\theta/\lambda)^2$. The different estimates are indicated by: triangles for estimates of Σ_Q calculated according to equation (1) (Henderson and Moffat, 1971), diamonds for estimates from equation (2) (Bricogne, 1976), circles for estimates from equation (3) (Nixon and North, 1976). The correct value, calculated as a function of $\sin\theta/\lambda$, is shown by the curve.

missing, including all of the solvent molecules, and the model was similar to bovine trypsin in places where SGT is not. Individual thermal motion parameters had not yet been introduced, and the parts of the model that were essentially correct were inaccurate. Figure IV.4 demonstrates that, for TD2, neither the method of Bricogne [equation (2)] nor even that of Nixon and North [equation (3)] gives reliable estimates of phase probabilities. The problem is not one of being unable to estimate Σ_Q correctly [even when the phases

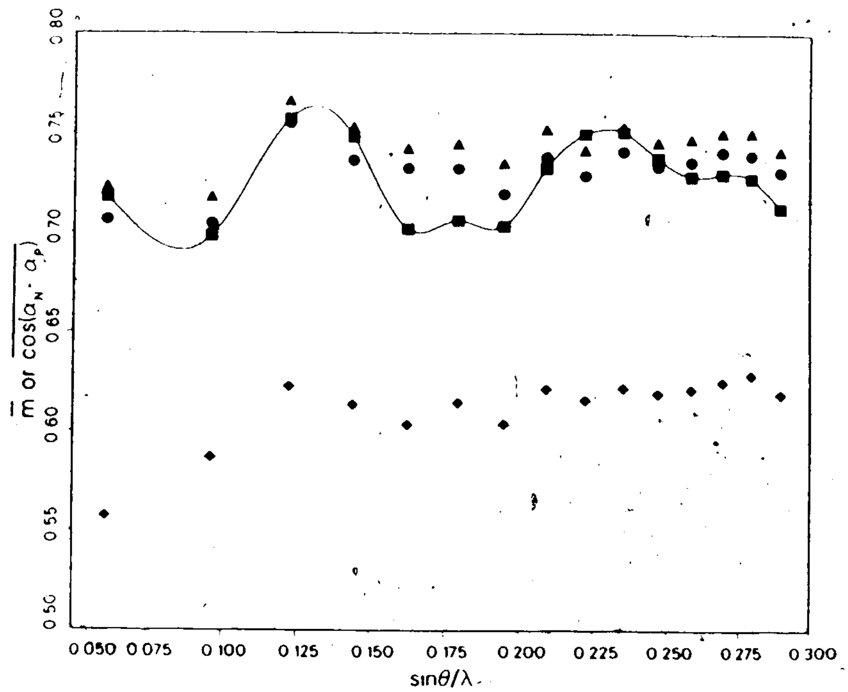


Figure IV.3. Estimates of m for Perfect Partial Structure. The smooth curve connects the mean values of $\cos(\alpha_N-\alpha_P)$ for each resolution range, shown by squares. The other points are the mean values of m calculated from the estimates of Σ_Q shown with the same symbols in Figure IV.2.

are used to estimate Σ_Q via $\overline{|F_N-F_P^c|^2/\epsilon}$, $\bar{m}$ does not agree with $\overline{\cos(\alpha_N-\alpha_P^c)}$]; rather the problem is that the phase probability expressions are no longer valid.

Estimating σ_A

Since one may not safely ignore coordinate errors, it is necessary to use the phase probability distributions of Srinivasan and co-workers. Srinivasan and Ramachandran (1965) showed that, when the probability distributions are cast in terms of normalized structure factors, the effects of missing structure and of coordinate errors are formally

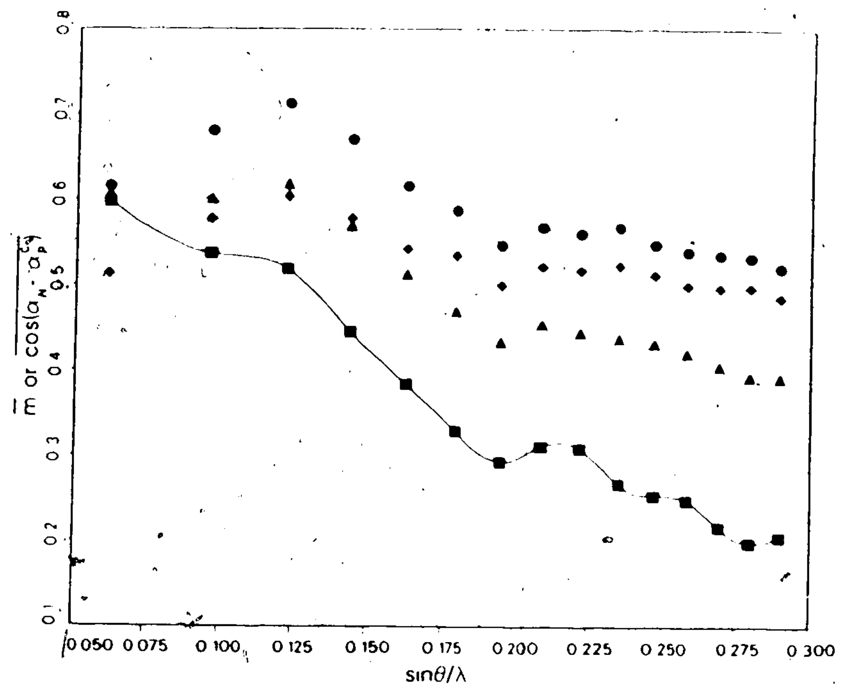


Figure IV.4. Estimates of m from Σ_Q for Partial Structure with Errors. Phase probabilities calculated from estimates of Σ_Q were evaluated for the case of a partial structure with coordinate errors (TD2). The smooth curve connects the mean values (shown by squares) of $\cos(\alpha_N-\alpha_P^C)$ for each resolution range. The estimates are indicated by: diamonds for mean values of m calculated from Σ_Q estimated by equation (2); circles for mean values of m calculated from Σ_Q estimated by equation (3); and triangles for mean values of m calculated from Σ_Q estimated from the mean value of $|F_N-F_P^C|^2/\epsilon$, i.e., using a knowledge of the phases.

equivalent. The parameter σ_A in these expressions varies from zero when the partial structure provides no phase information (no atoms in the partial structure, or an unrelated partial structure) to one, when the partial structure is perfect and complete. The factor D in σ_A (see Table IV.1) varies strongly with resolution when there are significant coordinate errors. As a result, σ_A should generally be estimated for several resolution ranges.

Hauptman (1982) has derived joint probability distributions for structure factors from isomorphous pairs of structures. In these equations, his parameter α plays the same role as σ_A in the expressions of Srinivasan. Hauptman suggests that σ_A can be estimated as the square root of the correlation coefficient between $|E_1|^2$ and $|E_2|^2$.

$$\sigma_A \simeq \left[\frac{\Sigma(|E_1|^2-\overline{|E_1|^2})(|E_2'|^2-\overline{|E_2|^2})}{[\Sigma(|E_1|^2-\overline{|E_1|^2})^2\Sigma(|E_2|^2-\overline{|E_2|^2})^2]^{1/2}} \right]^{1/2} \qquad (4)$$

Lunin and Urzhumtsev (1984) have proposed that parameters defining phase probabilities for partial structures with errors can be estimated from non-centric structure factors by maximizing a likelihood function. This approach can be extended to include centric data and the expected intensity factor ϵ. Since there appear to be misprints in the paper of Lunin and Urzhumtsev (1984), the derivation of their result is repeated here.

Srinivasan and Ramachandran (1965) derived the probability densities of $|E_N|$ conditional on $|E_P^C|$. For the non-centric case,

$$P(|E_N|;|E_P^C|)=\frac{2|E_N|}{1-\sigma_A^2}\exp\left[-\left[\frac{|E_N|^2+\sigma_A^2|E_P^C|^2}{1-\sigma_A^2}\right]\right]I_0\left[\frac{2\sigma_A|E_N||E_P^C|}{1-\sigma_A^2}\right]$$

Since the relationships are symmetrical for the normalized structure factors (Srinivasan and Ramachandran, 1965), the roles of $|\mathbf{E}_N|$ and $|\mathbf{E}_P^C|$ can be interchanged. Changing variables to put the structure factors on absolute scale,

$$P(|\mathbf{F}_P^C|;|\mathbf{F}_N|)=\frac{2|\mathbf{F}_P^C|}{\epsilon\beta}\exp\left[-\left[\frac{|\mathbf{F}_P^C|^2+\alpha^2|\mathbf{F}_N|^2}{\epsilon\beta}\right]\right]I_0\left[\frac{2\alpha|\mathbf{F}_P^C||\mathbf{F}_N|}{\epsilon\beta}\right] \qquad (5)$$

where $\alpha = \sigma_A(\Sigma_P/\Sigma_N)^{1/2}$ and $\beta = \Sigma_P(1-\sigma_A{}^2)$. Equation (5) is the same as equation (7) of Lunin and Urzhumtsev (1984), except for the inclusion of the factor ϵ and two apparent misprints in that paper (μ for the first β, and α for α^2 in the argument of the exponential). For the centric case, one can similarly derive from the results in Srinivasan and Ramachandran (1965) that

$$P(|\mathbf{F}_P^C|;|\mathbf{F}_N|)=\left[\frac{2}{\pi\epsilon\beta}\right]^{1/2}\exp\left[-\left[\frac{|\mathbf{F}_P^C|^2+\alpha^2|\mathbf{F}_N|^2}{2\epsilon\beta}\right]\right]\times$$

$$\cosh\left[\frac{\alpha|\mathbf{F}_P^C||\mathbf{F}_N|}{\epsilon\beta}\right] \qquad (6)$$

Lunin and Urzhumtsev maximize the likelihood function

$$\psi = \Pi P(|\mathbf{F}_P^C|;|\mathbf{F}_N|) \qquad (7)$$

where the expression for $P(|\mathbf{F}_P^C|;|\mathbf{F}_N|)$ is their equation (7). By using equation (5) for non-centric and (6) for centric reflections, all of the data can be used in equation (7). Estimates of α and β that maximize the likelihood function ψ

occur when the partial derivatives of $\ln\psi$ with respect to α and β are both zero. When the appropriate expressions for the figure of merit are used [where $X = 2\alpha|F_P^C||F_N|/(\epsilon\beta)$], the partial derivatives for the non-centric and centric terms differ only by a weighting factor ($w = 2$ for non-centric and 1 for centric).

$$\frac{\partial \ln\psi}{\partial\alpha} = \Sigma\left[\frac{w(m|F_P^C||F_N|-\alpha|F_N|^2)}{\epsilon\beta}\right] = 0 \tag{8}$$

$$\frac{\partial \ln\psi}{\partial\beta} = \Sigma\left[\frac{w(|F_P^C|^2 + \alpha^2|F_N|^2-2\alpha m|F_P^C||F_N|-\epsilon\beta)}{2\epsilon\beta^2}\right] = 0 \tag{9}$$

Equations (8) and (9) can be solved for β to get

$$\beta = \Sigma[w(|F_P^C|^2-\alpha^2|F_N|^2)/\epsilon]/\Sigma w \tag{10}$$

Considering the definitions of α and β, this is not a surprising result. From equation (8), the parameter α is determined by finding the zero of

$$R = \Sigma[w(\alpha|F_N|^2-m|F_P^C||F_N|)/\epsilon] \tag{11}$$

Note that α and β will adjust to compensate for an arbitrary change of scale. If $|F_P^C|$ is scaled by a factor k_P and $|F_N|$ by k_N, then values of α scaled by the factor (k_P/k_N) and of β scaled by k_P^2 will satisfy (8) and (9) while leaving the figures of merit unchanged. Therefore, if we use structure factors normalized so that $\Sigma w|E|^2/\Sigma w = 1$, α is equivalent to σ_A, equation (10) simplifies to

$$\beta = 1-\sigma_A^2$$

and (11) simplifies to

$$R = \Sigma w(\sigma_A - m|\mathbf{E}_P^C||\mathbf{E}_N|) \qquad (12)$$

In summary, the method proposed by Lunin and Urzhumtsev (1984) corresponds to estimating σ_A by finding the zero of the residual function R in equation (12). In equation (12), w=1 for centric and 2 for non-centric reflections, m is the appropriate function of σ_A, and the structure factors are normalized so that $\Sigma w|\mathbf{E}|^2/\Sigma w = 1$. This equation is consistent with the result of Srinivasan and Chandrasekaran (1966) that

$$\sigma_A = \langle|\mathbf{E}_1||\mathbf{E}_2|\cos(\alpha_1-\alpha_2)\rangle/(\langle|\mathbf{E}_1|^2\rangle\langle|\mathbf{E}_2|^2\rangle)^{1/2} \qquad (13)$$

Newton's method is used to solve for the zero of the residual function R; the initial estimate of σ_A is calculated from equation (4).

$$\frac{dR}{d\sigma_A} = \Sigma w\left[1-|\mathbf{E}_P^C||\mathbf{E}_N|\frac{dm}{d\sigma_A}\right] \qquad (14)$$

The expression for m differs for centric and non-centric data. For centric data,

$$\frac{dm}{d\sigma_A} = \frac{(1-m^2)}{2}\frac{dX}{d\sigma_A} \qquad (15)$$

For non-centric data,

$$\frac{dm}{d\sigma_A} = [1-(m/X)-m^2]\frac{dX}{d\sigma_A} \qquad (16)$$

Finally,

$$\frac{dX}{d\sigma_A} = \frac{2|\mathbf{E}_N||\mathbf{E}_P^C|(1+\sigma_A{}^2)}{(1-\sigma_A{}^2)^2} \qquad (17)$$

Equations (14) through (17) define $dR/d\sigma_A$, and the next estimate of σ_A is given by

$$\sigma_{A,i+1} = \sigma_{A,i} - R/(dR/d\sigma_A)$$

In practice, the refinement of σ_A values generally converges in 3 or 4 cycles. Because the parameter σ_A varies with resolution, it is evaluated in shells of equal width in $(\sin\theta/\lambda)^2$. When there are between 500 and 1000 reflections in each shell, the estimates of σ_A seem to vary fairly smoothly and reliably. When there are too few reflections in a shell and the correct value of σ_A is small, there is sometimes a negative correlation between $|E_N|^2$ and $|E_P^C|^2$, in which case any starting value of σ_A refines to zero. (Note from equation (12) that $\sigma_A = 0$ is always a zero of the residual R.) In general, the algorithm is less stable and the results less reliable when the correct value of σ_A is small. In this case, it is necessary to have a larger number of reflections in each resolution shell.

If the method used to normalize the structure factors does not result in the weighted mean of $|E|^2$ being precisely unity, the simplifications made above are not valid. The easiest solution is to renormalize within each resolution shell. Otherwise, one must return to equations (10) and (11).

The results of the σ_A estimates for the two test data sets TD1 and TD2 are shown in Figures IV.5 (σ_A estimates) and IV.6 (figure of merit estimates). Figures of merit calculated using σ_A are much more reliable than those calculated using Σ_Q (compare Figure IV.6 with Figures IV.3 and

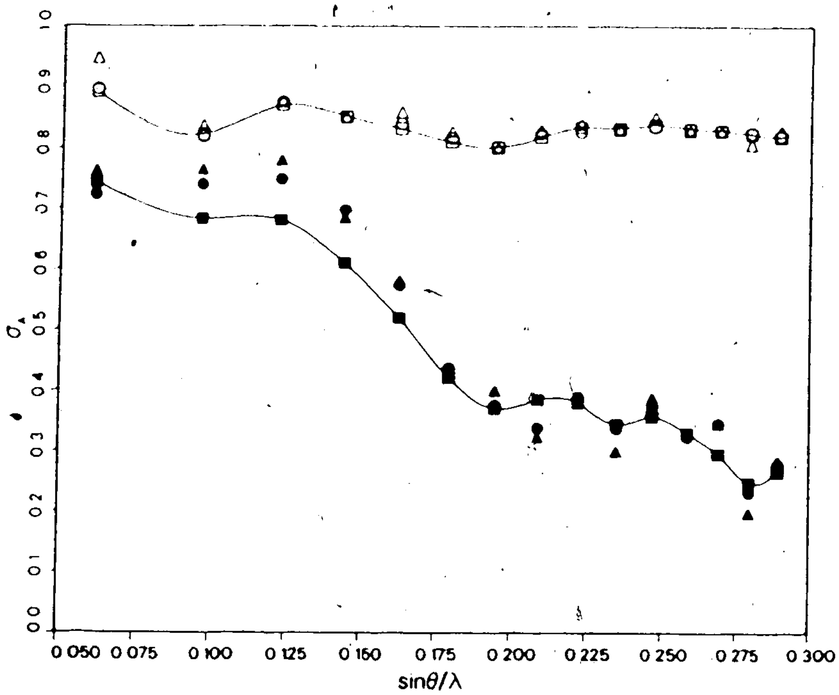


Figure IV.5. Evaluation of Methods for Estimating σ_A. Estimates of σ_A are shown with open symbols for the test with TD1, and closed symbols for TD2. The smooth curves connect the values of σ_A calculated from equation (13) for each resolution range (squares). The triangles show the estimates of σ_A from equation (4); the circles show ~~the~~ refined estimates of σ_A.

IV.4). From Figures IV.5 and IV.6 one sees that the refined estimates of σ_A are slightly better than the estimates from equation (4), and that they lead to somewhat more reliable figures of merit for both sets of test data. (In implementing these methods, one might decide that the increased accuracy does not justify the increased programming effort.)

Lunin and Urzhumtsev (1984) observe that the accuracy of phases obtained from models refined in reciprocal space is overestimated. This effect can be seen in the TD2 data. The first 7 cycles of refinement for SGT used 6.0 to 2.8Å

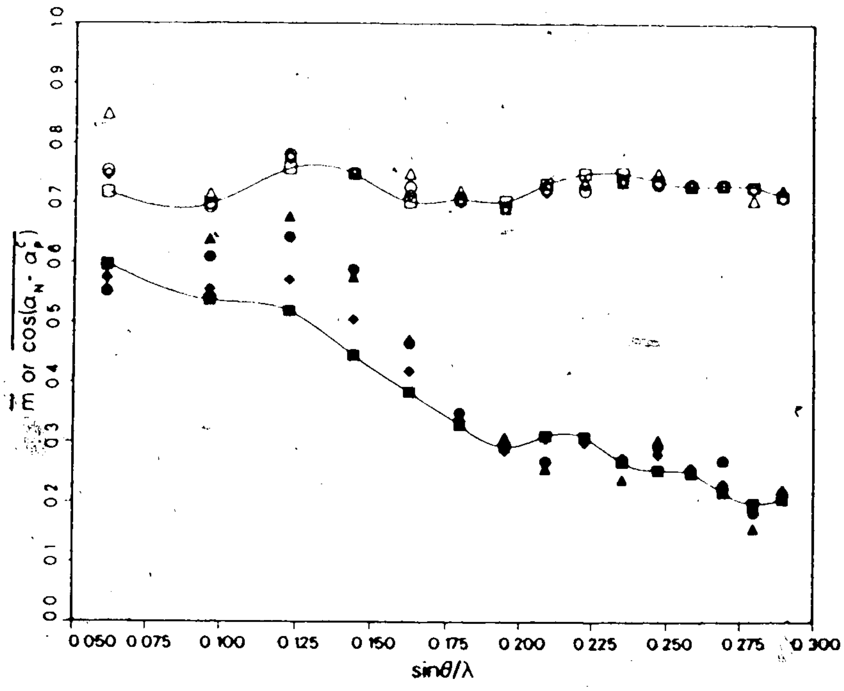


Figure IV.6. Estimates of m from σ_A. The estimated figures of merit are evaluated by comparing the mean values of $\cos(\alpha_N - \alpha_P^C)$ and m. As for Figure IV.5, open symbols are used for TD1 and closed symbols for TD2. The smooth curves connect the mean values of $\cos(\alpha_N - \alpha_P^C)$, shown with squares. Other points indicate the mean values of m calculated from the estimates of σ_A shown with the same symbols in Figure IV.5. In addition, the diamonds show the mean values of m calculated from the values of σ_A determined with equation (13).

data, and the figures of merit are systematically overestimated only within these resolution limits. In some way structure refinement must alter the distribution of errors; even the values of σ_A calculated using the phase differences via equation (13) give slightly high figures of merit.

A potential concern in the use of this method to determine σ_A is the effect of omitting reflections having low $|F_N|$. In macromolecular crystallography, many low intensity measurements are quite unreliable. The practice of discarding these observations has been criticized (e.g., Hirshfeld and Rabinovitch, 1973), but is still quite common. This is probably due in part to concerns of cost and computer memory; in addition, the resulting bias is less pronounced for positional than for thermal motion parameters.

When low intensity reflections are discarded, the distribution of $|\mathbf{F}_N|$ is altered, so that the joint distribution of $|\mathbf{F}_N|$ and $|\mathbf{F}_P^C|$ is altered. Therefore, under these circumstances, phase probabilities determined using equation (4) are unreliable. On the other hand, the distribution of $|\mathbf{F}_P^C|$ conditional on $|\mathbf{F}_N|$ is unaffected, so that equations (5) through (11) are still valid. Numerical tests (results not shown) confirm that figures of merit calculated from refined estimates of σ_A are still reasonable when reflections having small $|\mathbf{F}_N|$ are omitted. However, the reduction in the number of observations can aggravate the instability in the algorithm when σ_A is small, so that figures of merit determined using all of the data are more reliable. In addition, the inclusion of data that were not used in structure refinement might be expected to reduce the overestimation of σ_A.

Estimation of Coordinate Error from σ_A

The Luzzati (1952) plot, which is commonly used to estimate coordinate errors in macromolecular structures, is based on the variation of D (see Table IV.1) with resolution. Srinivasan and Ramachandran (1965) note that, if the variation in σ_A with resolution is ascribed to the factor D, the resolution dependence of σ_A can also be used, in principle, to estimate the mean coordinate error of the atoms comprising the partial structure. However, the approach developed in later papers (e.g., Srikrishnan and Srinivasan, 1968) is to calculate an overall normalized *R*-factor that, in a comparison with theoretical values, leads to a value for the mean coordinate error; this approach requires one to make an accurate *a priori* estimate of the ratio (Σ_P/Σ_N).

If the coordinate errors are assumed to be normally distributed, then

$$D = \exp\,[-\pi^3(<|\Delta r|>)^2(\sin\theta/\lambda)^2]$$

where $<|\Delta r|>$ is the expected value of the coordinate error (in Å) (Luzzati, 1952). The radial standard deviation of atomic position is a more statistically useful quantity than the mean coordinate error. As Chambers and Stroud (1979) note, the radial distribution of coordinate error is equivalent to a Maxwell distribution of velocities, so it is appropriate to make the substitution

$$(<|\Delta r|>)^2 = (8/3\pi)<|\Delta r|^2>$$

and take the natural logarithm on each side to get

$$\ln\sigma_A = (1/2)\ln(\Sigma_P/\Sigma_N) - (8\pi^2/3)<|\Delta r|^2>(\sin\theta/\lambda)^2 \qquad (18)$$

If the ratio (Σ_P/Σ_N) is constant, then a plot of $\ln\sigma_A$ vs. $(\sin\theta/\lambda)^2$ should give a straight line with a slope of $[-(8\pi^2/3)<|\Delta r|^2>]$ and an intercept of $(1/2)\ln(\Sigma_P/\Sigma_N)$. An example of a σ_A plot is shown in Figure IV.7.

By assuming that (Σ_P/Σ_N) is constant, one assumes that the missing atoms are of the same type and have the same overall temperature factor as the atoms included in the partial structure. Clearly, this is invalid for the disordered solvent in a protein crystal; in applying equation (18) it will be necessary to ignore data to which the disordered solvent atoms contribute significantly, i.e., reflections at

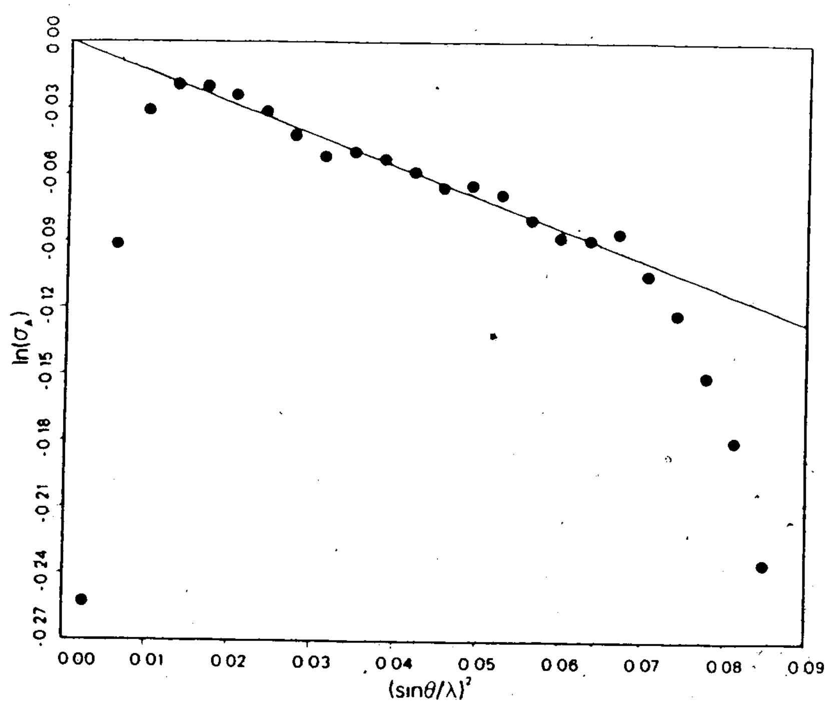


Figure IV.7. σ_A Plot to Estimate Coordinate Error. To show the effects of disordered solvent and measurement errors, observed structure factors are used as $|F_N|$. (For reflections with negative measured intensities, $|F_N|$ is set to zero. In the highest resolution range, 20.9% of the reflections fall into this category.) The values of $|F_P^C|$ are the calculated structure factors from SGT at cycle 78 of least-squares refinement. The line is a least squares fit to all the points excluding the first 3 and the last 4. From the intercept (0.000), (Σ_P/Σ_N) = 1.000, and from the slope (-1.410Å^2), $<|\Delta r|^2>^{1/2}$ = rms $|\Delta r|$ = 0.231Å.

lower than 5 or 6Å resolution (see Figure IV.7). Atoms missing from partial structures often come from the less well-ordered regions. This will hinder the use of equation (18) at intermediate stages of structure refinement.

The σ_A plot will also be affected by the overestimation of σ_A that results from structure refinement (Lunin and Urzhumtsev, 1984). Since the weights applied to structure factors during least-squares refinement are often an empirical function of resolution (Hendrickson and Konnert, 1980), the degree to which σ_A is overestimated could vary as a function of resolution, depending on the precise choice of weights. This means that the estimated coordinate error could be sensitive to the structure factor weights. A possible example of this effect was given in Chapter II.

For highly-refined structures, the observation errors in $|\mathbf{F}_N|$ will contribute significantly to the disagreement with $|\mathbf{F}_P^C|$. Since the proportional error in $|\mathbf{F}_N|$ generally increases with resolution, measurement errors will tend to lead to an overestimate of the coordinate error. The non-linearity at high resolution of the σ_A plot shown in Figure IV.7 can probably be attributed in part to measurement errors. In addition, only a minority of the highest resolution data are above the cutoff used in refinement, so that the overestimation of σ_A due to structure refinement might be reduced at high resolution.

Similar considerations affect the use of the method of

Luzzati (1952), except that the condition on the ratio (Σ_P/Σ_N) is more restrictive: it is assumed implicitly that Σ_P and Σ_N are equal. In a Luzzati plot, one compares a family of theoretical curves to R-factors calculated in resolution shells; the σ_A plot defined by equation (18) requires determining only the slope of a line. Since σ_A is calculated from normalized structure factors, the σ_A plot, unlike the Luzzati plot, is unaffected by scaling errors. Though a σ_A plot must be interpreted with due care, it is therefore preferable in several respects to the Luzzati plot.

B. Removing Model Bias from Maps

Structural information in the Fourier synthesis is contained to a great extent in the phase angles (Ramachandran and Srinivasan, 1970; Oppenheim, 1981). Therefore, electron density maps phased with model phases are biased towards the model. Several suggestions have been made for Fourier coefficients that reduce model bias.

Luzzati (1953) showed that, for an almost complete structure, a map with non-centric coefficients $|F_N|\exp(i\alpha_P)$ will show the missing atoms at half weight, but at less than half weight when more of the structure is missing. Because of this effect, maps with the commonly used coefficients $(2|F_N|-|F_P^C|)\exp(i\alpha_P^C)$ bring missing atoms up towards full weight. For a generalized version of these coefficients, $[n|F_N|-(n-1)|F_P|]\exp(i\alpha_P)$, Vijayan (1980) determined the value of n appropriate for different amounts of missing structure. Following a somewhat different approach, Main (1979) showed that, for non-centric data,

$$m|F_N|\exp(i\alpha_P) \simeq 1/2\ F_N + 1/2\ F_P$$

so non-centric coefficients that reduce model bias are given by $(2m|F_N|-|F_P|)\exp(i\alpha_P)$. Main's approach will be extended here to the case of a partial structure with errors.

Non-centric Case

Following Main (1979), we start with the cosine law.

$$|\delta|^2 = |E_N|^2 + \sigma_A{}^2|E_P^C|^2 - 2\cos(\alpha_N-\alpha_P^C)\sigma_A|E_N||E_P^C|. \qquad (19)$$

We use $\sigma_A E_P^C$ instead of E_P^C because the expected magnitude and direction of δ ($= E_N-\sigma_A E_P^C$) are uncorrelated with those of E_P^C (see Figure IV.1). Thus E_N, $\sigma_A E_P^C$ and δ are interrelated in the same way as F_N, F_P and F_Q, respectively. Replacing both sides of (19) by expected values,

$$<|\delta|^2> = |E_N|^2 + \sigma_A{}^2|E_P^C|^2 - 2m\sigma_A|E_N||E_P^C|$$

Noting that

$$|E_N|^2 = E_N E_N{}^* = E_N(\sigma_A E_P^C{}^* + \delta^*)$$

we can rearrange and multiply both sides by $\exp(i\alpha_P^C)$ to get

$$m|E_N|\exp(i\alpha_P^C) = \frac{E_N}{2} + \frac{\sigma_A E_P^C}{2} + \frac{E_N\delta^*}{2\sigma_A E_P^C{}^*} - \frac{<|\delta|^2>}{2\sigma_A E_P^C{}^*}$$

$$m|E_N|\exp(i\alpha_P^C) = \frac{E_N}{2} + \frac{\sigma_A E_P^C}{2} + \frac{\delta^*}{2}\exp(2i\alpha_P^C) + \frac{|\delta|^2-<|\delta|^2>}{2\sigma_A E_P^C{}^*} \qquad (20)$$

In the Fourier transform of $m|E_N|\exp(i\alpha_P^C)$, the third and fourth terms of (20) will lead to background noise [cf. Main (1979)]. Therefore, ignoring the noise contributions,

$$E_N \simeq (2m|E_N|-\sigma_A|E_p^C|)\exp(i\alpha_p^C)$$

$$F_N \simeq (\epsilon\Sigma_N)^{1/2}(2m|E_N|-\sigma_A|E_p^C|)\exp(i\alpha_p^C) \qquad (21)$$

Substituting into (21) the expression for σ_A from Table IV.1,

$$F_N \simeq (2m|F_N|-D|F_p^C|)\exp(i\alpha_p^C) \qquad (22)$$

Thus, when there are no errors in the coordinates of the partial structure, D = 1 and equation (22) simplifies to give the Fourier coefficients derived by Main (1979). An advantage to working with normalized structure factors is that the scale factor and overall *B* value used in calculating structure factors do not affect the values of E_p^C, so that errors in these quantities have no effect on Fourier coefficients calculated from equation (21). The information on the scale and *B* parameters relative to those of $|F_N|$ is contained in the values of σ_A (see Table IV.1), which vary as a function of resolution.

The omission of low intensity reflections will have an adverse effect on map coefficients calculated with equation (21). This differs from the result above, that truncation of data has no systematic effect on the estimation of figures of merit, because two different conditional probabilities are involved in the two cases. The refined estimates of σ_A depend on the conditional probability $P(E_p^C;E_N)$, whereas the derivation of equation (21) depends on $P(E_N;E_p^C)$. As long as the structure factors are normalized to be on the same scale ($<|E_p^C|^2> = <|E_N|^2>$), these expressions will be symmetrical (Srinivasan and Ramachandran, 1965), so that the value of σ_A in $P(\mathbf{E}_P^C;\mathbf{E}_N)$ is the same as in $P(\mathbf{E}_N;\mathbf{E}_P^C)$. This will be true when the data are complete. When low intensity observations are discarded, normalization over the truncated data changes the relative scale of $|\mathbf{E}_N|$ and $|\mathbf{E}_P^C|$.

Using a prime to indicate variables derived from truncated data, if we let $|\mathbf{E}_N'| = k_N|\mathbf{E}_N|$ and $|\mathbf{E}_P^C'| = k_P|\mathbf{E}_P^C|$, then $\sigma_A' = (k_P/k_N)\sigma_A$ (as discussed above) and $\Sigma_N' = \Sigma_N/k_N^2$. If one uses these values in equation (21)

$$(\epsilon\Sigma_N')^{1/2}(2m|\mathbf{E}_N'|-\sigma_A'|\mathbf{E}_P^C'|) = 2m|\mathbf{F}_N|-(k_P/k_N)^2D|\mathbf{F}_P^C| \quad (23)$$

Therefore, if not accounted for, data truncation will introduce a systematic error into the map coefficients. Normally it would be preferable to avoid this problem by using all of the data. However, if the relative size of k_P and k_N is known, it is possible to apply the appropriate correction.

It is necessary to determine only one of k_P and k_N; the other can then be estimated. From equation (10),

$$\Sigma w(|\mathbf{E}_P^C'|^2-\sigma_A'^2|\mathbf{E}_N'|^2)/\Sigma w = 1-\sigma_A'^2 \quad (24)$$

where the sums are taken over the truncated data. Since the difference vector $(\mathbf{E}_P^C'-\sigma_A'\mathbf{E}_N')$ is independent of $\mathbf{E}_N'$, equation (24) would be valid even if the missing data were added. Therefore,

$$k_P^2-k_N^2\sigma_A'^2 = 1-\sigma_A'^2 \quad (25)$$

Thus, if k_N is determined by comparing Σ_N with Σ_N', k_P can be calculated from equation (25) and the correction implied by equation (23) can be made to the non-centric map coefficients.

Centric Case

In the centric case, we start from (20) and note that

$$\frac{\delta *}{2}\exp(2i\alpha_{P}^{C}) = \frac{\delta}{2}$$

because

$$\alpha_{P}^{C} = \alpha_{\delta} + n\pi$$

Therefore,

$$m|\mathbf{E}_N|\exp(i\alpha_{P}^{C}) = \mathbf{E}_N + \frac{|\delta|^2-<|\delta|^2>}{2\sigma_A \mathbf{E}_{P}^{C}*} \qquad (26)$$

As in the case treated by Main (1979), therefore, the appropriate Fourier coefficients for centric data are simply $m|\mathbf{F}_N|\exp(i\alpha_{P}^{C})$.

Evaluating the Map Coefficients

It is difficult to do objective visual comparisons of electron density maps calculated with different coefficients. One common quantitative measure for comparing maps is the root-mean-square value of the difference electron density (Blow and Crick, 1959). A related measure, which has the virtue of being unaffected by scaling errors, is the coefficient of correlation between electron density maps. For an electron density map omitting the contribution of the F_{000} term, the mean density is zero, so that the correlation coefficient is defined by

$$r = \frac{\int_V \rho_1(\mathbf{x})\rho_2(\mathbf{x})d\mathbf{x}}{[\int_V \rho_1(\mathbf{x})^2 d\mathbf{x} \int_V \rho_2(\mathbf{x})^2 d\mathbf{x}]^{1/2}}$$

Applying the convolution theorem

$$r = \frac{\Sigma|F_1||F_2|\cos(\alpha_1-\alpha_2)}{[\Sigma|F_1|^2\Sigma|F_2|^2]^{1/2}} \qquad (27)$$

where the sums are taken over a hemisphere of reciprocal space. One might note, comparing equations (27) and (13), that the correlation coefficient between two E maps is equivalent to an overall value for σ_A.

Two correlation coefficients will be used as objective criteria to judge electron density maps. The first is the correlation with the correct map (i.e., the Fourier transform of $\mathbf{F_N}$), which should be as close as possible to unity. Following a similar argument to that of Blow and Crick (1959), one can show that the maximum correlation between the correct map and one with Fourier coefficients $(w_1|F_N| + w_2|F_P^C|)\exp(i\alpha_P^C)$ is obtained when $w_1 = m$ and $w_2 = 0$; any coefficients designed to compensate for model bias will lower the correlation with the correct map. Nonetheless, model bias in an electron density map makes it difficult to detect and correct errors in the model. For this reason, coefficients that reduce model bias lead to maps that are subjectively (even if not objectively) improved. Optimal map coefficients will reduce model bias at only a small cost in the correlation with the correct map.

The second correlation coefficient is that with the model map. For a model-biased map this correlation will be higher than that of the correct map with the model map. Map coefficients that reduce model bias should lower this correlation, but not excessively. A correlation lower than that between the correct and model maps would indicate that correct features of the model were being eliminated.

Table IV.2 shows the results of some test calculations with TD2. The two correlation coefficients were evaluated, using equation (27), for several types of map coefficients that have been suggested previously. These results indicate that the coefficients described here are superior in reducing model bias with little cost in the resemblance to the correct map.

All of the map coefficients evaluated in Table IV.2 are phased by the model. The greatest model bias is found for the unweighted coefficient, $|F_N|$. As expected, the highest correlation is given by figure-of-merit weighting $(m|F_N|)$, but this map shows considerable model bias. The non-centric coefficients $(2m|F_N|-D|F_P^C|)$ give a large reduction in model bias with little cost in the correlation to the correct map. Some of this reduction in model bias comes from reflections with small $|F_N|$, for which the map coefficient will often be negative; omitting reflections with small $|F_N|$ leads to a slight increase in model bias. Though the factor D might seem counter-intuitive, its omission in the coefficients $(2m|F_N|-|F_P^C|)$ leads to an excessive reduction in both correlation coefficients because the negative $|F_P^C|$ component is too large. The coefficients $[m(2|F_N|-|F_P^C|)]$ are fairly successful in the case of TD2, but with an accurate partial structure, the figure of merit would not provide a

Table IV.2

Correlation Coefficients Between Electron Density Maps

Non-centric Fourier coefficients of tested map[1]	Correct map (coefficients F_N)	Model map (coefficients F_P^C)
F_N	1.0	0.585
$\lvert F_N\rvert \exp(i\alpha_P^C)$	0.640	0.851
$m\lvert F_N\rvert \exp(i\alpha_P^C)$	0.698	0.833
$(2m\lvert F_N\rvert - D\lvert F_P^C\rvert)\exp(i\alpha_P^C)$	0.663	0.650
$(2m\lvert F_N\rvert - D\lvert F_P^C\rvert)\exp(i\alpha_P^C)$[2]	0.666	0.631
$(2m\lvert F_N\rvert - \lvert F_P^C\rvert)\exp(i\alpha_P^C)$	0.570	0.397
$m(2\lvert F_N\rvert - \lvert F_P^C\rvert)\exp(i\alpha_P^C)$	0.660	0.682
$(2m_B\lvert F_N\rvert - \lvert F_P^C\rvert)\exp(i\alpha_P^C)$[3]	0.588	0.479
$(2\lvert F_N\rvert - \lvert F_P^C\rvert)\exp(i\alpha_P^C)$	0.573	0.630
$(3\lvert F_N\rvert - 2\lvert F_P^C\rvert)\exp(i\alpha_P^C)$	0.490	0.446

[1] Centric Fourier coefficients were $m|F_N|\exp(i\alpha_P^C)$ or $|F_N|\exp(i\alpha_P^C)$ for all figure-of-merit weighted and unweighted maps respectively.
[2] These data were truncated by setting to zero all coefficients having $|F_N|<150e$.
[3] m_B refers to figures of merit calculated by the method of Bricogne (1976).

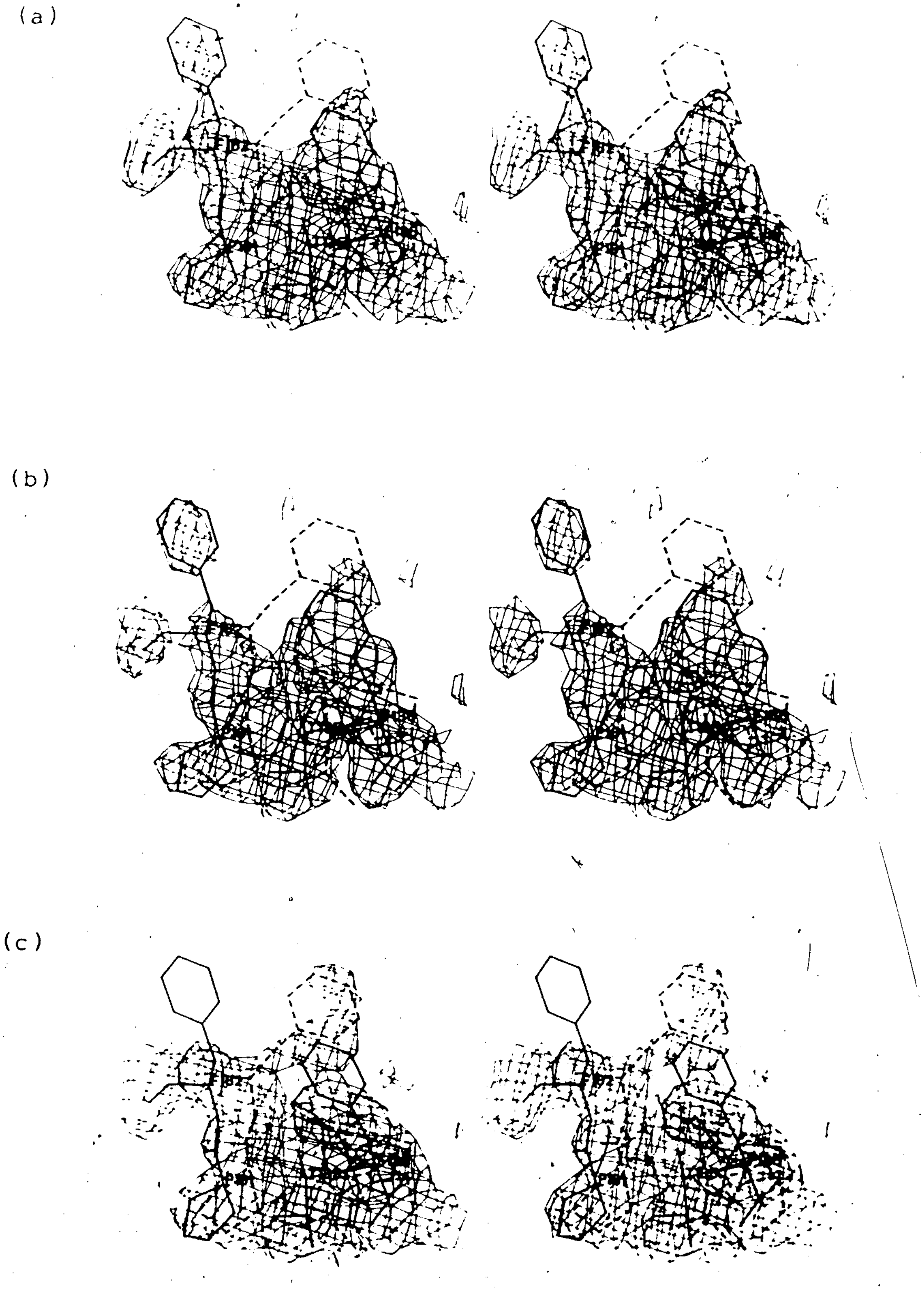
fortuitous compensation for the factor D. When figures of merit calculated by the method of Bricogne (1976) are used in the coefficients $(2m_B|F_N|-|F_P^C|)$, both correlations are quite low; they are not as low as for the related coefficients $(2m|F_N|-|F_P^C|)$ because the overestimation of m_B relative to m compensates in part for the omission of the factor D. Finally, the unweighted coefficients $(2|F_N|-|F_P^C|)$ and $(3|F_N|-2|F_P^C|)$ both lead to a low correlation with the correct map.

Figure IV.8 allows a more subjective comparison. Electron density is shown for a part of SGT where two phenylalanine side chains were positioned incorrectly. The map using the coefficients derived here is compared to a map computed with the non-centric coefficients $(2m_B|F_N|-|F_P^c|)$, a figure-of-merit weighted $(m|F_N|)$ map, and the correct map. The figure-of-merit weighted map displays serious model bias. In contrast, both the $(2m|F_N|-D|F_P^c|)$ and the $(2m_B|F_N|-|F_P^c|)$ maps indicate that the model is incorrect, and both show, at least in part, the correct positions of the side chains. However, the map computed with the coefficients $(2m|F_N|-D|F_P^c|)$ is somewhat superior in clarity and in connectivity of the density. The loss of connectivity in the $(2m_B|F_N|-|F_P^c|)$ map, which is consistent with the correlation coefficients in Table IV.2, would seriously impede the interpretation of some parts of the map.

Combined Phase Map Coefficients

To the extent that model phases influence combined phases, maps computed with combined phases will be biased towards the model. This was noted by Rice (1981) in work on the structure of phosphoglycerate kinase, and model bias caused serious problems in the refinement of SGT (Chapter II).

In order to compute a combined phase map, one must first combine the phase information from different sources. The combined phase probability density is the product of the


(a)
(b)
(c)

(d)

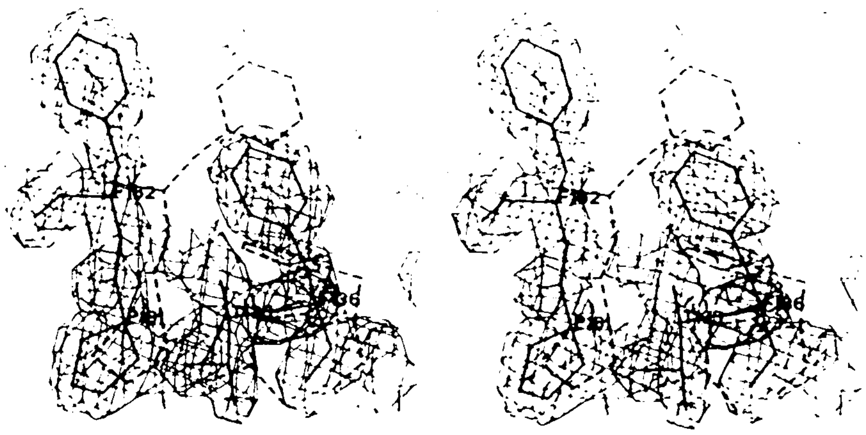


Figure IV.8. Comparison of Model-phased Electron Density Maps. For each electron density map, the correct structure (SGT at refinement cycle 78) is shown in solid lines; the partial structure with errors (SGT at cycle 7) is shown in dashed lines. All three maps are computed using the full 1.7Å TD2 data set. Each map is contoured at 1.25 times the rms electron density of the map. For clarity, only contours within 1.7Å of an atom in the figure are shown. (a) Map calculated with non-centric coefficients $(2m|F_N|-D|F_P^C|)\exp(i\alpha_P^C)$ and centric coefficients $m|F_N|\exp(i\alpha_P^C)$. Contoured at 0.45 e/Å^3. (b) Map calculated with non-centric coefficients $(2m_B|F_N|-|F_P^C|)\exp(i\alpha_P^C)$ and centric coefficients $m_B|F_N|\exp(i\alpha_P^C)$, where m_B is the figure of merit calculated by the method of Bricogne (1976). Contoured at 0.42 e/Å^3. (c) Map calculated with coefficients $m|F_N|\exp(i\alpha_P^C)$ and contoured at 0.35e/Å^3. (d) Map calculated with coefficients F_N, i.e., the correct map, and contoured at 0.47 e/Å^3.

phase probability densities of the independent sources of phase information (Rossmann and Blow, 1961). Instead of multiplying the probability curves point-by-point, it is more convenient to add up the Hendrickson-Lattman coefficients that encode the phase information (Hendrickson and Lattman, 1970). Partial structure phase probabilities are unimodal, so the Hendrickson-Lattman coefficients C_{PAR} and D_{PAR} are both zero. For non-centric data,

$$A_{PAR} = X \cos(\alpha_P^C)$$

$$B_{PAR} = X \sin(\alpha_P^C)$$

where X is the argument for the phase probability expressions, as in Table IV.1 (Hendrickson and Lattman, 1970). For centric data, different expressions must be used for A_{PAR} and B_{PAR} or else the figures of merit, evaluated as in equations (22) of Hendrickson and Lattman (1970), will be too high. Substituting the expression for the centric figure of merit [m=tanh(X/2)] into equations (6) of Hendrickson (1971) gives

$$A_{PAR} = (X/2) \cos(\alpha_P^C)$$

$$B_{PAR} = (X/2) \sin(\alpha_P^C)$$

The reduction of model bias in combined phase maps is not as straightforward as for model-phased maps. With a few assumptions, however, non-centric coefficients to reduce model bias in combined phase maps can be derived. (As noted above, there is no model bias in figure-of-merit weighted centric coefficients.) First, it will be assumed that the coefficients $m_{comb}|F_N|\exp(i\alpha_{comb})$ are biased towards the model by an amount that varies linearly with the extent of influence that the model has on the phases. To make use of this assumption, some measure of the relative influence of several sources of phase information will be needed. Consider the possibility of several sources of model phases (for example, molecular replacement phases from several related proteins - ignoring the possibility that such phase information might not be considered to be completely independent) and one source of non-model phases (e.g., MIR). It will be most convenient to define weights that vary from 0 for a phase source that supplies none of the phase information to 1 for a phase source that supplies all of the phase information.

These weights should not be based on figures of merit, because a figure of merit can be 0 for a sharply peaked bimodal probability density that supplies a great deal of phase information. Turning to information theory, the variation of information, $H(\rho|\rho_o)$, measures the amount of information gained in going from a prior probability density ρ_o to a posterior probability density ρ. (H is related to the entropy, or missing information, of a probability density.) To measure the total information content of ρ, ρ_o must be an uninformative prior: for phase probabilities, $\rho_o=(1/2\pi)$.

$$H(\rho|\rho_o) = \int_0^{2\pi} \rho(\alpha)\ln[\rho(\alpha)/\rho_o(\alpha)]d\alpha$$

(Guiasu, 1977), or

$$H(\rho|\rho_o) = \int_0^{2\pi} \rho(\alpha)\ln[2\pi\rho(\alpha)]d\alpha$$

For simplicity of notation, H_{MIR} will be used for the variation of information of the MIR phase probability density and, for instance, H_1 for the probability density of the first partial structure. Finally, suitable weights are defined by

$$w_{MIR} = H_{MIR}/(H_{MIR} + \Sigma H_i)$$

where the sum is over the partial structures, and

$$w_i = H_i/(H_{MIR} + \Sigma H_i), \text{ so that}$$

$$w_{MIR} + \Sigma w_i = 1$$

When $w_{MIR} = 1$ (i.e., all the phase information is from MIR),

$$m_{comb}|\mathbf{F}_N|\exp(i\alpha_{comb}) = m_{MIR}|\mathbf{F}_N|\exp(i\alpha_{MIR}) \simeq \mathbf{F}_N$$

When, for instance, $w_1 = 1$,

$$m_{comb}|\mathbf{F}_N|\exp(i\alpha_{comb}) = m_1|\mathbf{F}_N|\exp(i\alpha_1) \simeq \mathbf{F}_N/2 + D_1\mathbf{F}_1/2$$

($\mathbf{F}_1$ here refers to the calculated structure factor for partial structure 1.) To deal with cases where the weights have values other than 0 and 1, the two assumptions are invoked: 1) the expression for $m_{comb}|\mathbf{F}_N|\exp(i\alpha_{comb})$ is a linear combination of the two expressions just given; 2) the variation of information gives the correct relative weight. Then, combining the expressions as a linear function of the weights,

$$m_{comb}|\mathbf{F}_N|\exp(i\alpha_{comb}) \simeq [w_{MIR}+\Sigma(w_i/2)]\mathbf{F}_N + \Sigma(w_i/2)D_i\mathbf{F}_i$$

$$\simeq [1-\Sigma(w_i/2)]\mathbf{F}_N + \Sigma(w_i/2)D_i\mathbf{F}_i$$

Solving for $\mathbf{F}_N$ (i.e., for the model-unbiased Fourier coefficient) gives

$$F_N \simeq \frac{m_{comb}|F_N|\exp(i\alpha_{comb}) - \Sigma(w_i/2)D_i F_i}{1 - \Sigma(w_i/2)} \qquad (28)$$

Though the derivation of this expression is not rigorous, it makes intuitive sense. In the extreme cases, for which one of the weights is unity, equation (28) simplifies to the appropriate more rigorously-derived expression, and it varies smoothly between these extremes. Note that the F_i are phased by the model phases; one would not expect to eliminate model bias by subtracting structure factor vectors pointing in the wrong direction. Finally, the influence of bimodal MIR phase probabilities is not underestimated, as it would be if the figure of merit were used instead of H.

Stuart and Artymiuk (1984) have also developed coefficients to remove model bias from maps using combined phases. Their approach differs from the approach used here in several important respects: the relative importance of different sources of phase information is judged by the figure of merit; the coefficients that result when the partial structure is the only source of phase information are different; the calculated phase is not applied to the $(-|F_p^C|)$ component. However, the basic idea, that of having the expression for each map coefficient vary according to the extent to which the model determines the phase, is the same. A systematic comparison of Stuart and Artymiuk's (1984) map coefficients and those given by equation (28) has not yet been attempted.

The combined phase map coefficients derived here were not available in their present form until the refinement of SGT was nearly complete. However, they have been used in work on the structure of pepsinogen (James and Sielecki, 1985). A comparison of the MIR map and a combined phase map using phases from an early model of pepsinogen allows at least a subjective evaluation of the combined phase map coefficients.

This early model of pepsinogen was, in essence, a molecular replacement model derived from penicillopepsin (James and Sielecki, 1983), but only the atoms for which there was a minimal amount of electron density in the previous combined phase map were included (A. R. Sielecki, M. Fujinaga, R. J. Read and M. N. G. James, unpublished). The MIR phases, available to 2.8Å resolution, had a mean figure of merit of 0.63. For the model phases, considering first only those reflections for which MIR phases were available, $\bar{m}$ was 0.42; including all reflections to 2.0Å resolution, $\bar{m}$ was 0.24. Finally, the mean combined figure of merit was 0.71 for the reflections with combined phases, and 0.30 including all reflections to 2.0Å resolution. The poor quality of the model phases is indicative of the serious errors that existed in the model at that time.

It is possible that combined phase electron density maps could be worse than MIR maps in regions where the model is seriously in error. This concern is addressed by the comparison of MIR and combined phase maps in Figure IV.9. In the region shown in this figure, pepsinogen differs markedly from penicillopepsin, so the phasing model is almost completely wrong. Neither map is particularly easy to interpret in this region, though the combined phase map seems to be marginally worse. It is important to note, however, that the combined phase map would not mislead one into believing that the phasing model is correct. Figure IV.10, in contrast, shows a region where pepsinogen and penicillopepsin are reasonably similar. In this region, the combined phase map gives a much closer representation of the final structure than the MIR map. In particular, the improvement in main-chain connectivity makes it much easier to interpret.

C. Summary

It has been shown that, if one takes into account errors in the partial structure, more accurate estimates of phase probabilities can be made. It should be noted that a partial structure can be an atomic model or a density-modified electron density map. The values of σ_A in the phase probability expressions (see Table IV.1) are estimated most easily with equation (4) (Hauptman, 1982). However, estimates from equation (4) are not reliable when low intensity observations are omitted. The effect of errors in the observed structure factor magnitudes has not been considered. However, these errors are likely to be small compared to the

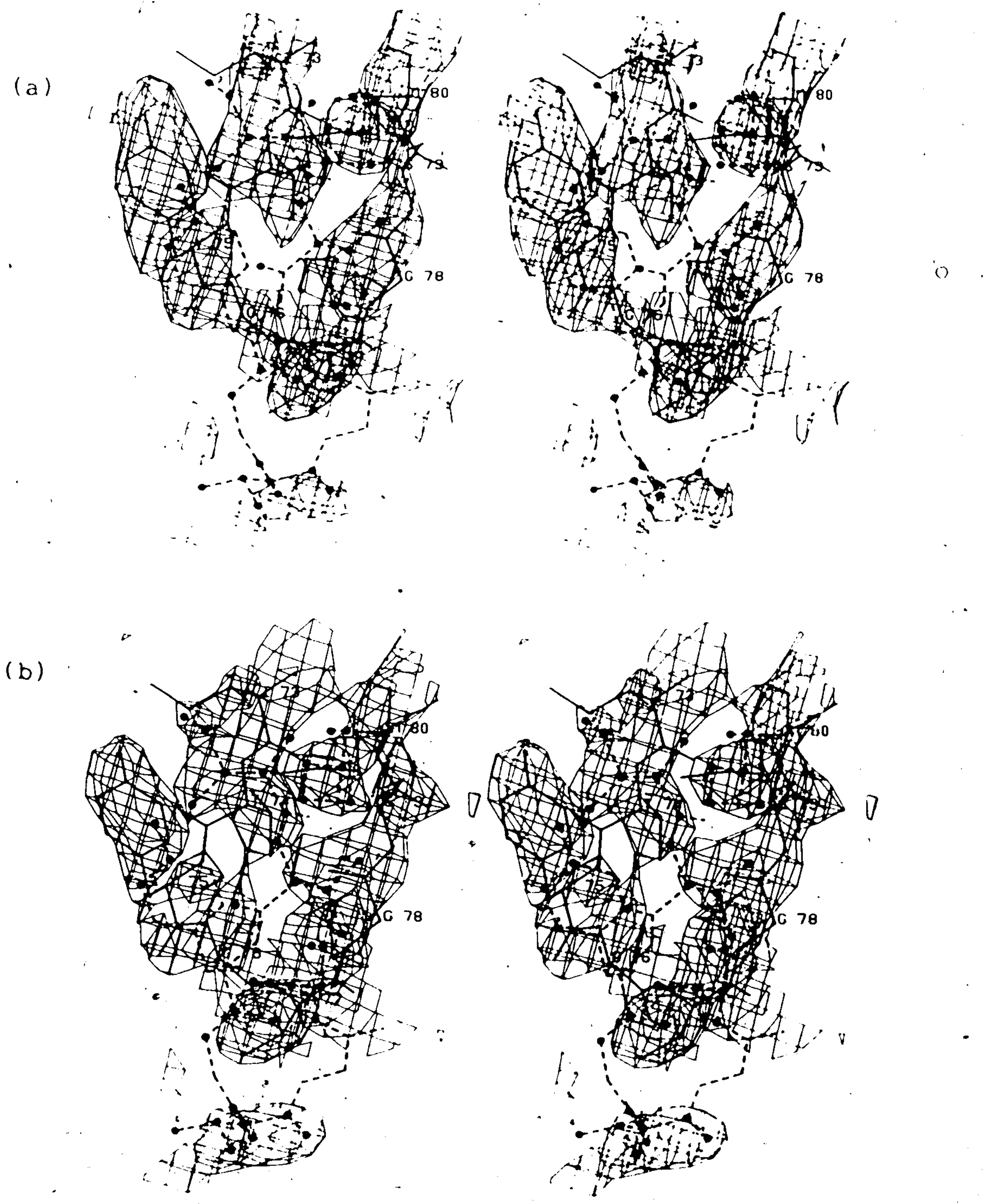


Figure IV.9. Comparing MIR and Combined Phase Maps in Poor Region of Model. Dashed lines show an early model of pepsinogen; circles indicate those atoms that were actually in the phasing model. Solid lines show the refined model of pepsinogen at cycle 63 of refinement. Both maps are contoured at 1.2 times the rms value of the electron density: (a) MIR map, contoured at $0.25e/Å^3$; (b) combined phase map, contoured at $0.35e/Å^3$.

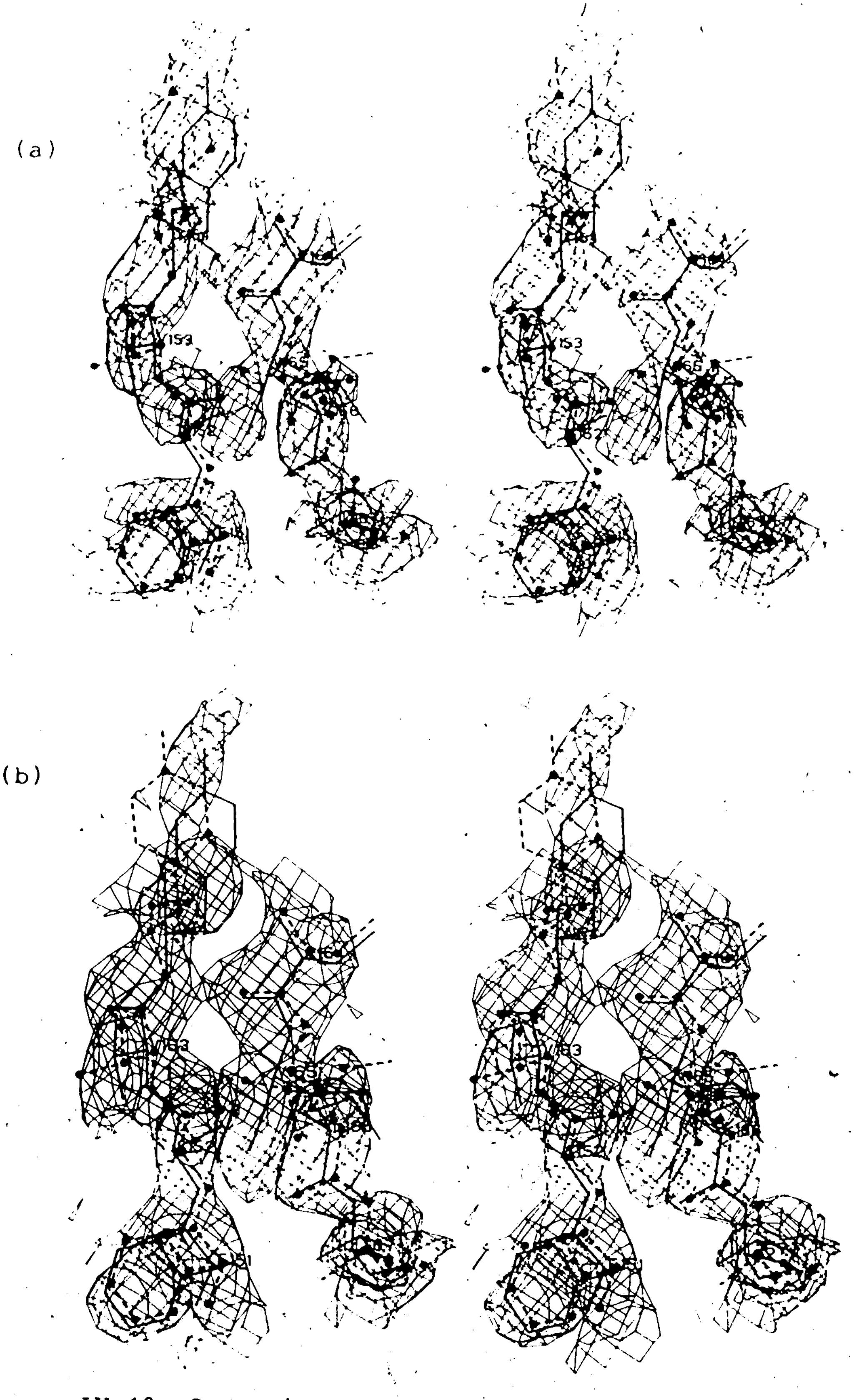


Figure IV.10. Comparing MIR and Combined Phase Maps in Good Region of Model. Line types and contour levels are as for Figure IV.9. (a) MIR map, (b) combined phase map.

difference between the correct and calculated structure factors in a case for which phase probabilities are needed, i.e., when the model is bad.

Main's (1979) work on map coefficients for partial structures has been extended to encompass model errors. In a simulation, the coefficients derived here [$(2m|\mathbf{F}_N|-D|\mathbf{F}_P^C|)\exp(i\alpha_P^C)$ for non-centric, $m|\mathbf{F}_N|\exp(i\alpha_P^C)$ for centric] are objectively superior to coefficients that are currently in general use. Applying a few assumptions, coefficients for combined phase maps have also been derived. Examination of a sample combined phase map shows definite subjective improvements in clarity over an MIR map, while little model bias appears to be introduced.

Accurate phase probabilities for partial structures are important for purposes other than the calculation of electron density maps. The neglect of coordinate errors leads to a significant overestimation of phase accuracy, which would cause combined phases to be skewed towards model phases. In addition, any attempts to improve or extend model phases by direct methods or maximum entropy methods (Bricogne, 1984) should benefit from more accurate phase probabilities.

Bibliography

Blow, D. M., & Crick, F. H. C. (1959) *Acta Cryst. 12*, 794-802.

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

Bricogne, G. (1976) *Acta Cryst. A32*, 832-847.

Bricogne, G. (1984) *Acta Cryst. A40*, 410-445.

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

Guiasu, S. (1977) *Information Theory With Applications*, McGraw-Hill, London.

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

Henderson, R., & Moffat, J. K. (1971) *Acta Cryst. B27*, 1414-1420.

Hendrickson, W. A. (1971) *Acta Cryst. B27*, 1472-1473.

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.

Hendrickson, W. A., & Lattman, E. E. (1970) *Acta Cryst. B26*, 136-143.

Hirshfeld, F. L., & Rabinovitch, D. (1973) *Acta Cryst. A29*, 510-513.

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

James, M. N. G., & Sielecki, A. R. (1985) *Nature (London)*, in press.

Lunin, V. Y., & Urzhumtsev, A. G. (1984) *Acta Cryst. A40*, 269-277.

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

Luzzati, V. (1953) *Acta Cryst. 6*, 142-152.

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

Nixon, P. E., & North, A. C. T. (1976) *Acta Cryst. A32*, 325-333.

Oppenheim, A. V. (1981) *Proceedings of the IEEE 69*, 529-541.

Ramachandran, G. N., & Srinivasan, R. (1970) *Fourier Methods in Crystallography*, Wiley, New York.

Rice, D. W. (1981) *Acta Cryst. A37*, 491-500.

Rossmann, M. G., & Blow, D. M. (1961) *Acta Cryst. 14*, 641-647.

Sim, G. A. (1959) *Acta Cryst. 12*, 813-815.

Sim, G. A. (1960) *Acta Cryst. 13*, 511-512.

Srikrishnan, T., & Srinivasan, R. (1968) *Zeitschrift für Kristallographie 127*, 427-441.

Srinivasan, R. (1966) *Acta Cryst. 20*, 143-144.

Srinivasan, R. (1968) *Zeitschrift für Kristallographie 126*, 175-181.

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

Srinivasan, R., & Ramachandran, G. N. (1965) *Acta Cryst. 19*, 1008-1014.

Stuart, D., & Artymiuk, P. (1984) *Acta Cryst. A40*, 713-716.

Vijayan, M. (1980) *Acta Cryst. A36*, 295-298.

Woolfson, M. M. (1956) *Acta Cryst. 9*, 804-810.

V. General Discussion

The studies in this dissertation are mostly self-contained, and the conclusions to be drawn from them have already been stated. If there is an underlying theme to this work, it is that of how to make the best use of prior structural information in protein crystallography.

For many proteins of interest, the three-dimensional structure of a related protein is known. This structural information can be exploited by the technique of comparative model-building. The evaluation of comparative models of SGT demonstrated a number of sources of serious error in this technique. At the very least, a consideration of the possible errors will discourage over-interpretation of comparative models; in addition, it should lead to improved techniques and more accurate models.

Prior structural information can also be exploited by the technique of molecular replacement, which was used in the solution of all three protein structures described herein. In the case of SGT, the available models had low homology compared to most successful molecular replacement models. To solve this structure, therefore, it was necessary to optimize the strategies employed; the strategies that were developed should apply to other difficult replacement problems. The molecular replacement solution of the structures of the OMTKY3:proteinase complexes was more routine, since accurate models of the enzymes were available. The case of OMTKY3:SGPB exemplifies the use of this technique to determine the previously unknown structure of a protein in a complex with a protein of known structure. For OMTKY3:CHT$_{A}$ molecular replacement facilitated the study of the interaction between two proteins of known structure. One might fear that, without independent phase information, structural differences from the model might be obscured. Nonetheless, these studies showed conformational differences between SGT and BT, and conformational changes in both SGPB and OMTKY3.

Once a model for a crystal structure is available, either from molecular replacement or from electron density map fitting, it provides a source of phase information. The structure factor phases computed from a structural model can be used to solve heavy-atom derivatives, to visualize parts of the structure missing from the model, or to correct fitting errors. For the optimal use of phases calculated from partial structure models, one requires estimates of their accuracy. A reliable method for estimating partial structure phase probabilities has been given here. The use of calculated phases introduces model bias into electron density maps, which can impede the correction of errors in the model. Various Fourier coefficients for electron density map calculation have been proposed to alleviate this problem; the coefficients derived here have been shown to reduce model bias in a more satisfactory fashion than other coefficients in general use.