

**Homology Modelling Of The b Fraction Of**  
**Factor B**

**Mark G. Faller**

Submitted for the qualification: **MSc**

At Glasgow University

Department Of Chemistry

Submitted December 1997

ProQuest Number: 13815588

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 13815588

Published by ProQuest LLC (2018). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code  
Microform Edition © ProQuest LLC.

ProQuest LLC.  
789 East Eisenhower Parkway  
P.O. Box 1346  
Ann Arbor, MI 48106 – 1346

**GLASGOW UNIVERSITY  
LIBRARY**

11305 (copy 1)

## 0.1 Abstract

The b domain of factor B (Bb) is found to be moderately homologous to the mammalian serine proteases. Sequence homology analysis showed that out of all the known serine protease structures, Bb was most homologous to bovine trypsin. Sequence alignment between Bb and bovine trypsin gave a sequence identity of the two sequences of between 17 - 21%. This meant that protein model building of Bb from a known serine protease structure, in this case bovine trypsin, could be attempted. The structure of bovine trypsin was taken from the Brookhaven Database. All the molecular modelling was carried out with the software package "COMMET" that was developed in our laboratory.

In the first instance all of the substitutions necessary to computationally mutate the bovine trypsin structure into the Bb structure were carried out. Substitutions were carried out first as they are the least disruptive of the three modelling techniques used in homology modelling. Substitution only alters the side chain atoms of the residue that is being modified. No alterations to the backbone atoms are necessary at this stage. The software keeps the new side chain position as close to the original as possible. Where this is not feasible the side chain conformation is determined by a conformational search of the side chain's conformation space.

The deletions from bovine trypsin were all small and accomplished by simple removal of the appropriate residues followed by energy minimisation to reposition and rejoin the main chain. Small insertions up to three residues long were built using the "insert" routine. After inserting a residue its side chain's torsion angles were defined by a conformational search. To ease steric strain at the site of small deletions and insertion a segment five residues either side of the insertion or deletion was run through the energy minimiser.

There are a total of eight insertions of three residues in length or longer:

- Gln 30: 3 residues in length
- Ser 186: 3 residues in length
- Gly 129: 7 residues in length
- His 231: 8 residues in length



- Glu 101: 9 residues in length
- Leu 143: 9 residues in length
- Arg 170: 13 residues in length

The conformation of these large insertions was calculated using a sophisticated conformational space sampling procedure which runs on a large parallel computer. The loop conformation generator searches through all of the conformational space and uses filters to eliminate unfavourable conformations.

After all the modifications were carried out the entire protein was run through the energy minimiser. First polar hydrogens, then all hydrogens, and finally the water molecules from the bovine trypsin crystal structure were added to the model. Energy minimisation continued until the first derivatives from the Newton-Ralphson algorithm had become negligibly small.

## 0.2 List Of Contents

<b><u>HOMOLOGY MODELLING OF THE B FRACTION OF FACTOR B</u></b> .....	<b>1</b>
0.1 Abstract .....	2
0.2 List Of Contents .....	4
0.3 List Of Diagrams.....	6
0.4 List Of Tables.....	8
0.5 Acknowledgements .....	10
<b>1 INTRODUCTION TO PROTEIN MODELLING</b> .....	<b>11</b>
1.1 Protein Structure.....	11
1.2 The Primary Sequence .....	18
1.3 The Secondary Structure .....	19
1.3.1 The Hydrogen Bond .....	22
1.3.2 The $\alpha$ Helix .....	23
1.3.3 The $\beta$ Sheet .....	26
1.3.4 The Reverse Turn .....	27
1.3.5 Random Coil.....	28
1.3.6 Ramachandran Maps.....	28
1.3.7 The Handedness Of Proteins .....	29
1.4 Super-Secondary Structure.....	30
1.5 Tertiary Structure.....	31
1.6 Description of Domains.....	31
<b>2 THE COMPLEMENT SYSTEM</b> .....	<b>33</b>
2.1 Introduction. ....	33
2.2 The Alternative Pathway of the Complement System.....	35
2.3 The Proteins.....	40
2.3.1 C3 .....	40
2.4 The Classical Pathway .....	40
2.5 Activation of the MAC (Membrane Attack Complex).....	41
<b>3 THE SERINE PROTEASES</b> .....	<b>47</b>
3.1 Enzymes In General .....	47
3.1.1 Evolution .....	47
3.2 The Serine Proteases.....	53
3.2.1 Scope of the family .....	53
3.2.2 The Reaction Mechanism.....	53
3.2.3 The 3D Crystal Structure Of The Serine Proteases .....	62
<b>4 MOLECULAR MODELLING</b> .....	<b>70</b>
4.1 <b>COMMET</b> .....	70
4.1.1 List Of Functions Available Within COMMET .....	71
4.2 <b>Molecular Mechanics</b> .....	76
4.2.1 Introduction .....	76
4.2.2 Why use Molecular Mechanics.....	80
4.2.3 Quantum Mechanics .....	80
4.2.4 Force Fields .....	81
4.2.5 Force Field Parameters .....	95
4.2.6 Minimisation of Steric Energy.....	96
<b>5 METHODS AND MATERIALS</b> .....	<b>107</b>
5.1 <b>Sequence Alignment</b> .....	107
5.1.1 Homology scan of Brookhaven Database .....	109
5.2 <b>Homology Modelling</b> .....	110
5.3 <b>Substitutions</b> .....	112
5.3.1 Conservative and non - Conservative Substitutions.....	113

5.3.2 SITAR .....	115
5.4 Deletions.....	116
5.4.1 Single amino acid Deletions.....	117
5.4.2 Deletions up to 3 residues .....	118
5.4.3 Larger deletions .....	119
5.5 Insertions .....	119
5.5.1 Where insertions tend to be.....	120
5.5.2 Small Insertions Of Amino Acids .....	120
5.6 Large insertions .....	122
5.6.1 Conformation Loop Generator .....	122
5.7 Addition of Hydrogens .....	126
5.8 Final Energy Minimisation.....	126
5.9 Use of the Transputers .....	129
5.9.1 The Need For Computing Power .....	129
5.9.2 Single Processor Systems .....	131
5.9.3 Multiprocessor Systems .....	134
5.9.4 The Transputer .....	135
6 RESULTS.....	139
6.1 Sequence Alignment of B Fraction of Factor B (Bb) with Bovine Trypsin, low homology	139
6.2 Substitutions.....	147
6.3 Deletions In Bb Sequence.....	147
6.3.1 Small Insertions Not Modelled Using The Conformation Generator.....	148
6.3.2 Large Insertions 3 Residues In Length and Longer.....	148
6.4 Global Minimisation .....	160
6.5 Refinements To The Model.....	167
7 DISCUSSION .....	223
7.1 Discussion of the Model.....	223
7.1.1 Assessment Of The Model .....	223
7.1.2 A Look At The Model .....	232
7.1.3 Discussion on the Active Site.....	239
7.1.4 The S <sub>1</sub> Binding Site .....	244
7.2 Where Protein Modelling Is At.....	246
7.2.1 Why Protein Modelling Is Necessary .....	246
7.2.2 There Are Only Certain Folding Motifs Proteins Adopt.....	247
7.2.3 Modelling Techniques.....	248
7.2.4 Comparison Of The Modelling Techniques .....	251
7.2.5 Methods Used To Model Loop Regions.....	253
7.2.6 Comparison Of The Loop Modelling Techniques.....	255
7.3 How Computer Generated Models Of Proteins Are Used In Research By The Pharmaceutical Industry.....	257
7.4 What Use Is The Model Of Bb To The Pharmaceutical Industry .....	258
7.5 Rational Drug Design .....	259
7.5.1 QSAR.....	259
7.5.2 Protein Modelling.....	261
7.5.3 Building the Compound Around The Protein Model.....	263
7.5.4 Advantage Of Using Protein Model Over QSAR .....	266
7.5.5 Scoring Functions .....	268
7.6 Conclusions .....	271
8 REPORT OF PROTEIN ANALYSIS BY THE WHAT IF PROGRAM. DATE : 1997-22-12 .....	273
8.1 INTRODUCTION.....	273
8.2 Legend.....	273
9 REFERENCES .....	321

## 0.3



## List Of Diagrams

DIAGRAM 1-1: THE GENERAL STRUCTURE OF AN AMINO ACID.....	11
DIAGRAM 1-2: NOMENCLATURE OF THE ATOMS IN AN AMINO ACID.....	16
DIAGRAM 1-3: L CHIRAL CONFORMATION.....	17
DIAGRAM 1-4: THE CYSTINE RESIDUE CONTAINING THE DISULPHIDE BRIDGE.....	18
DIAGRAM 1-5: THE AMINO ACID RESIDUE AND THE PEPTIDE UNIT.....	20
DIAGRAM 1-6: THE PEPTIDE'S DIMENSIONS.....	21
DIAGRAM 1-7: REPRESENTATION OF THE DIHEDRAL ANGLE USING THE NEWMANN PROJECTION.....	22
DIAGRAM 1-8: SIMPLE $\alpha$ HELIX, SHOWING HYDROGEN BONDS BETWEEN RESIDUES I AND (I+4).....	25
DIAGRAM 2-1: OVERVIEW OF THE COMPLEMENT SYSTEM.....	34
DIAGRAM 2-2: INTERNAL THIOLESTER STRUCTURE.....	36
DIAGRAM 2-3: THE DIFFERENT PATHWAY C3 FOLLOWS IN DIFFERENT ENVIRONMENTS.....	37
DIAGRAM 2-4: POSITIVE AMPLIFICATION LOOP FOR THE ACTIVATION OF C3.....	38
DIAGRAM 2-5: SCHEMATIC DRAWING OF C5B-7, C5B-8, C5B-8,6C9 AND C5B-8,POLY C9 BOUND TO A SINGLE BILAYER PHOSPHOLIPID. THE DRAWINGS SHOW TWO SIDE VIEWS OF THE SAME MEMBRANE INSERTED COMPLEX ROTATED BY $90^\circ$ .....	44
DIAGRAM 2-6: 1)THE SUBUNIT ARCHITECTURE OF THE MAC. 2) DIMENSIONS OF THE MAC.....	45
DIAGRAM 3-1: SUPERIMPOSITION OF PORCINE ELASTASE (1ESA) (RED), BOVINE CHYMOTRYPSIN (1GCT) (BLUE) AND BOVINE TRYPSIN (1TPO) (GREEN). THE PLOT IS A TRACE ON THE C ALPHA ATOM FOR EACH PROTEIN.....	49
DIAGRAM 3-2: ACYLATION MECHANISM FOR THE HYDROLYSIS OF A PEPTIDE OR SYNTHETIC ESTER BY A SERINE PROTEASE.....	54
DIAGRAM 3-3: IMIDAZOLE RING OF HIS 57 ACTING AS A BASE TO INCREASE THE NUCLEOPHILICITY OF THE HYDROXYL GROUP ON SER 195.....	55
DIAGRAM 3-4: THE FIRST THREE OF SIX STEPS FOR SERINE PROTEASE HYDROLYSIS OF PEPTIDES OR AMIDES. IN THIS REPRESENTATION THE PROTON SHUTTLE IS CONCERTED.....	57
DIAGRAM 3-5: THE FINAL THREE OF SIX STEPS FOR THE MECHANISM FOR SERINE PROTEASE HYDROLYSIS OF PEPTIDES OR AMIDES. IN THIS REPRESENTATION THE PROTON SHUTTLE IS CONCERTED.....	58
DIAGRAM 3-6: STEREO VIEW OF THE CATALYTIC TRIAD FROM SEVERAL TRYPSIN LIKE SERINE PROTEASES SUPERIMPOSED. 1TRN (RED), 1EPT (GREEN), 2SFA (BLUE), 1SGT (YELLOW), 2HNT (CYAN), 1ESA (MAGENTA).....	65
DIAGRAM 3-7: STEREO VIEW OF THE CATALYTIC TRIAD FOR HUMAN TRYPSIN (1TRN) (RED) AND SUBSTITILISIN (1SBT) (GREEN) SUPERIMPOSED.....	66
DIAGRAM 3-8: THE LABELLING OF THE SUBSITES (S) ON THE ENZYME AND THE SUBSTRATE RESIDUES (P).....	67
DIAGRAM 4-1: LIKELY ELEMENTS ON A POTENTIAL ENERGY SURFACE.....	79
DIAGRAM 4-2: INTERACTION TYPES CONSIDERED IN A FORCE FIELD.....	82
DIAGRAM 4-3: MORSE AND HARMONIC CURVES FOR BOND LENGTH.....	84
DIAGRAM 4-4: MORSE, HARMONIC AND HARMONIC WITH CUBIC CURVES.....	85
DIAGRAM 4-5: SINGLE PAIRWISE INTERACTION.....	89
DIAGRAM 4-6: SINGLE DIPOLE INTERACTION.....	90
DIAGRAM 4-7: TYPICAL VAN DER WAALS INTERACTION WHERE $R_0$ IS THE EQUILIBRIUM DISTANCE.....	91
DIAGRAM 4-8: OUT OF PLANE BENDING.....	93
DIAGRAM 4-9: IMPROPER TORSION ANGLE.....	94
DIAGRAM 4-10: BLOCK DIAGONAL NEWTON - RAPHSON.....	101
DIAGRAM 5-1: A SCHEMATIC REPRESENTATION OF A PIPELINE WITH 5 UNITS. IT SHOWS HOW 7 TASKS MOVE THROUGH THE PIPELINE, HOW SEVERAL CYCLES PASS BEFORE ANY RESULTS ARE USED, AND HOW THE PIPELINE REQUIRES TO BE FLUSHED TO GET THE REMAINING RESULTS.....	132

DIAGRAM 5-2: SCHEMATIC REPRESENTATION OF A VECTOR PROCESS. THE PIPELINES RUN IN PARALLEL SO EACH PIPELINE CAN PRODUCE A RESULT EVERY CLOCK CYCLE. IN THIS EXAMPLE 5 RESULTS ARE PRODUCED PER CLOCK CYCLE.....	134
DIAGRAM 5-3: SCHEMATIC REPRESENTATIONS OF SOME COMMON TOPOLOGIES USED TO CONNECT PROCESSORS IN A PARALLEL COMPUTER. ....	137
DIAGRAM 6-1: SEQUENCE ALIGNMENT BETWEEN C2, FACTOR B (FB) AND BOVINE TRYPSIN (TB). THE * MARK HOMOLOGOUS RESIDUES BETWEEN FB AND TB, AND A   DENOTES IDENTICAL RESIDUES BETWEEN FB AND TB.....	147
DIAGRAM 6-2: THE CHART SHOWS THE DISTANCE BETWEEN THE CENTROID OF THE PROTEIN AND THE CENTROID OF THE MOVING RESIDUES FOR EACH FRAME IN THE MOLECULAR DYNAMICS. ....	170
DIAGRAM 6-3: THIS CHART SHOWS THE TOTAL ENERGY OF THE MOVING RESIDUES IN EACH FRAME OF THE MOLECULAR DYNAMICS.....	170
DIAGRAM 6-4: THE CHART SHOWS THE DISTANCE BETWEEN THE MODEL CENTROID AND THE CENTROID OF THE MOVING RESIDUES FOR EACH STEP IN THE MOLECULAR DYNAMICS SIMULATION.....	171
DIAGRAM 6-5: THIS CHART SHOWS THE TOTAL ENERGY OF THE MOVING RESIDUES FOR EACH STEP IN THE MOLECULAR DYNAMICS SIMULATION. ....	171
DIAGRAM 6-6: NEW AND OLD CONFORMATION OF INS 170. THE OLD CONFORMATION IS IN PURPLE. THIS SHOWS THE NEW CONFORMATION OF THE INSERTION IS LIES MORE ON THE SURFACE.....	172
DIAGRAM 7-7: RAMACHANDRAN PLOT FOR THE REFINED MODEL OF Bb. PRODUCED BY THE PROTEIN STRUCTURE ANALYSIS ROUTINE "FULCHK" IN WHATIF. ....	228
DIAGRAM 7-8: A RIBBON DRAWING TRACING THE C $\alpha$ OF EACH RESIDUE. IT SHOWS THE $\alpha$ HELICES AND SECTIONS OF B STRANDS THAT HAVE REMAINED STABLE DURING THE ENERGY MINIMISATION PROCEDURE.....	234
DIAGRAM 7-9: A RIBBON REPRESENTATION OF THE BACKBONE OF THE Bb MODEL. THE 7 LARGE INSERTIONS ARE SHOWN IN BLACK. ....	234
DIAGRAM 7-10: THE TWO INSERTIONS AT 129 AND 231 ARE SHOWN IN STICK REPRESENTATION. THE SURROUNDING BACKBONE IS SHOWN IN A RIBBON REPRESENTATION. THE DIAGRAM SHOWS HOW CLOSE THE TWO INSERTIONS ARE IN THE THREE DIMENSIONAL STRUCTURE OF THE PROTEIN. ....	235
DIAGRAM 7-11: VIEW OF ALL ACIDIC AND BASIC (ASP, GLU, LYS, ARG, HIS) RESIDUES NOT INVOLVED IN A SALT BRIDGE. THERE IS ALSO A CALPHA TRACE.....	236
DIAGRAM 7-12: THE RESIDUES AROUND GLU 70 ARE SHOWN. GLY 44 IS LABELED. GLU 70 FORMS HYDROGEN BOND WITH NAMIDE OF GLY 44.....	237
DIAGRAM 7-13: THIS SHOWS THE RESIDUES AROUND LYS 160. ASN 189 FORMS A STRONG H-BOND WITH LYS 160. ....	238
DIAGRAM 7-14: THIS SHOWS THE RESIDUES AROUND ARG 153. ARG 153 FORMS H-BONDS WITH CYS 136 AND PRO 8 WHICH ARE ALSO LABELED IN THE DIAGRAM.....	239
DIAGRAM 7-15: STEREO VIEW OF THE ACTIVE SITE OF Bb SUPERIMPOSED ONTO THE ACTIVE SITE OF THE STARTING STRUCTURE, BOVINE TRYPSIN. ....	241
DIAGRAM 7-16: ACTIVE SITE OF THE Bb MODEL SUPERIMPOSED ONTO THE ACTIVE SITE OF BOVINE TRYPSIN. THE BACKBONE OF BOTH STRUCTURES IS SHOWN SEVERAL RESIDUES EITHER SIDE OF EACH RESIDUE OF THE ACTIVE SITE. ....	242
DIAGRAM 7-17: A VIEW ALONG THE CATALYTIC TRIAD OF Bb WITH THE CALPHA TRACE SHOWN IN GREEN. IT SHOWS THE INSERTIONS EXTEND A GOOD DISTANCE PAST THE ACTIVE SITE CAUSING THE ACTIVE SITE TO BE MORE BURIED THAN IN TB.....	243
DIAGRAM 7-18: THE SAME VIEW FOR TB LOOKING ALONG THE CATALYTIC TRIAD, WITH CALPHA TRACE IN GREEN. SHOWS THE ACTIVE SITE MUCH MORE EXPOSED THAN FOR Bb.....	244
DIAGRAM 7-19: CALPHA TRACE OF Bb WITH ASN 189, GLY 216 AND LYS 226 SHOWN. LYS 226 CAN BE SEEN TO BLOCK THE ENTRANCE TO THE S <sub>1</sub> POCKET.....	246

## 0.4 List Of Tables

TABLE 1-1: THE TWENTY NATURAL AMINO ACIDS FOUND IN PROTEINS.....	15
TABLE 3-1: SEQUENCE HOMOLOGIES IN MAMMALIAN SERINE PROTEASES.....	50
TABLE 3-2: SPECIES DIFFERENCES IN SERINE PROTEASES.....	51
TABLE 3-3: EXPERIMENTALLY KNOWN THREE DIMENSIONAL STRUCTURES OF TRYPSIN LIKE SERINE PROTEASES.....	63
TABLE 3-4: EXPERIMENTALLY KNOWN THREE DIMENSIONAL STRUCTURES OF SUBTILISIN LIKE SERINE PROTEASES.....	64
TABLE 6-5: RESULT OF ENERGY MINIMISATION ON THE CONFORMATIONS OF INSERTION GLN 30 PRODUCED BY THE LOOP CONFORMATION GENERATOR.....	175
TABLE 6-6: RESULTS OF ENERGY MINIMISATION ON THE CONFORMATIONS OF 'LONG' INSERTION AT GLN 30 PRODUCED BY THE LOOP CONFORMATION GENERATOR.....	178
TABLE 6-7: RESULTS OF ENERGY MINIMISATION ON THE CONFORMATIONS OF INSERTION 186 PRODUCED BY THE LOOP CONFORMATION GENERATOR.....	183
TABLE 6-8: RESULTS OF ENERGY MINIMISATION ON THE CONFORMATIONS OF INSERTION 127 PRODUCED BY THE LOOP CONFORMATION GENERATOR.....	188
TABLE 6-9: RESULTS OF ENERGY MINIMISATION ON THE CONFORMATIONS OF INSERTION 231 USING CONF 28 FROM INSERTION 127, PRODUCED BY THE LOOP CONFORMATION GENERATOR.....	195
TABLE 6-10: RESULTS OF ENERGY MINIMISATION ON THE FOUR CONFORMATIONS OF INSERTION 231 WITH THE LOWEST ENERGY USING CONFORMATION 28 OF INSERTION 127 NOW TESTED WITH CONFORMATION 26 OF INSERTION 127.....	196
TABLE 6-11: RESULTS OF ENERGY MINIMISATION ON THE TWO CONFORMATIONS OF INSERTION 127 (CONFORMATIONS 26 AND 28) WITH CONFORMATION 49 OF INSERTION 231.....	197
TABLE 6-12: RESULTS OF ENERGY MINIMISATION AFTER EACH RESIDUE IN THE INSERTION IS INSERTED INTO THE MODEL.....	199
TABLE 6-13: RESULTS OF ENERGY MINIMISATION AFTER EACH RESIDUE IN THE INSERTION IS INSERTED INTO THE MODEL.....	201
TABLE 6-14: RESULTS OF ENERGY MINIMISATION AFTER EACH RESIDUE IN THE INSERTION IS INSERTED INTO THE MODEL.....	203
TABLE 6-15: RESULTS OF THE ENERGY MINIMISATION AFTER EACH RESIDUE IN THE INSERTION IS ADDED INTO THE MODEL.....	205
TABLE 6-16: RESULTS OF ENERGY MINIMISATION AFTER EACH RESIDUE IN THE INSERTION IS INSERTED INTO THE MODEL. THE * INDICATES WHERE THIS MODEL WAS SAVED AND THE REST OF THE RESULTS DISCARDED DUE TO THE HIGH ENERGIES.....	206
TABLE 6-17: RESULTS OF THE MINIMISATION AFTER EACH RESIDUE OF THE INSERTION IS INSERTED INTO THE MODEL. THIS IS THE SECOND ATTEMPT WHERE THE HYDROGEN ATOMS ARE ADDED TO THE MODEL.....	207
TABLE 6-18: FURTHER ENERGY MINIMISATION OF THE INSERTION AT RESIDUE 143 BUT WITH THE HYDROGEN ATOMS REMOVED FROM THE MODEL.....	207
TABLE 6-19: ENERGY MINIMISATION OF INSERTION AT RESIDUE POSITION 143 AFTER THE PHENYL RING OF THE PHENYLALANINE AT POSITION 144.....	208
TABLE 6-20: THE PHENYL RING OF THE PHENYLALANINE RESIDUE AT POSITION 144 IS FIXED TO THE TORSION AND PLANAR ANGLES, AND THEN THE FRAGMENT IS PUT THROUGH THE ENERGY MINIMISER.....	210
TABLE 6-21: RESULTS OF ENERGY MINIMISATION AFTER THE HELIX FRAGMENT IS ADDED TO THE MODEL.....	211
TABLE 6-22: RESULTS OF ENERGY MINIMISATION ON THE CONFORMATIONS OF INSERTION 170 PRODUCED BY THE LOOP CONFORMATION GENERATOR.....	220
TABLE 6-23: RESULTS OF MINIMISATION ON THE CONFORMATIONS OF DELETION 63 ASP PRODUCED BY THE LOOP CONFORMATION GENERATOR.....	222



## **0.5 Acknowledgements**

I would like to thank the following people for their ideas and input into the project throughout the three years. Dr. D.N.J. White, my supervisor, for his continual support and ideas when there were any problems. Mr J.N. Ruddock for discussing modelling techniques and who was also of great help making sure the software package "COMMET" done what I asked of it. Finally, Dr. P. Taylor from Ciba Geigy, for his insights into the Complement System and discussions on the three dimensional structure of Bb.

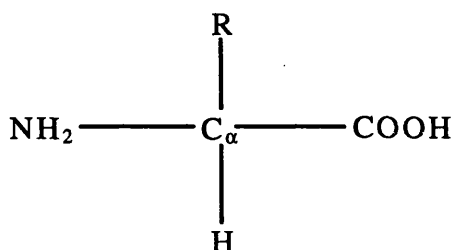
The grant for the project came from Ciba Geigy. I would like to thank Dr. D.N.J. White and Dr. P. Taylor for securing the grant for the full three years of the project.

Finally I would like to thank Dr A.J. Bleasby for the use of and time on the SEQNET computer and printer at Daresbury Laboratory during the final stages of writing this thesis.

# 1 Introduction to Protein Modelling

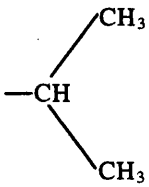
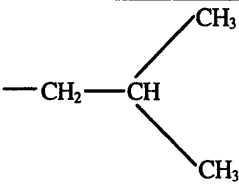
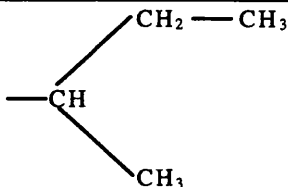
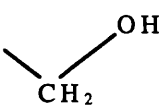
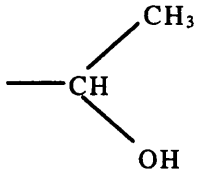
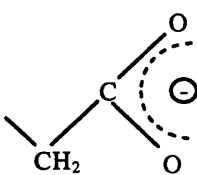
## 1.1 Protein Structure

It has long been known that proteins are made up of 20 fundamental structural units known as the amino acids. These amino acids can be linked together in different orders to produce the wide variety of proteins observed in nature. By the 1920's the structure of the twenty amino acids and how they combined to form proteins had been elucidated, mainly due to the work of Emil Fischer. The amino acids were found to have the general structure found in Diagram 1-1 on page 11:



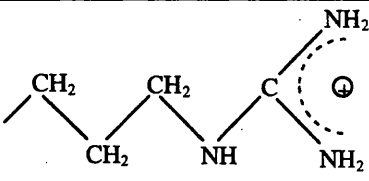
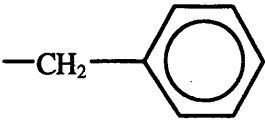
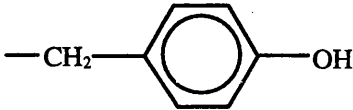
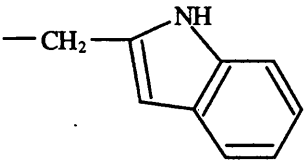
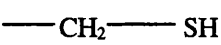
**Diagram 1-1: The general structure of an amino acid**

Name	3 Letter Code	1 Letter Code	Side Chain	Comments
Glycine	Gly	G	—H	uncharged polar, no sidechain other than H atom
Alanine	Ala	A	—CH <sub>3</sub>	nonpolar, methyl sidechain

Name	3 Letter Code	1 Letter Code	Side Chain	Comments
Valine	Val	V		nonpolar, branched at $\beta$ carbon atom
Leucine	Leu	L		nonpolar, branched at $\gamma$ carbon atom
Isoleucine	Ile	I		nonpolar, branched at $\beta$ carbon atom
Serine	Ser	S		uncharged polar, contains hydroxyl group
Threonine	Thr	T		uncharged polar, contains hydroxyl group branched at $\beta$ carbon atom
Aspartate	Asp	D		acidic

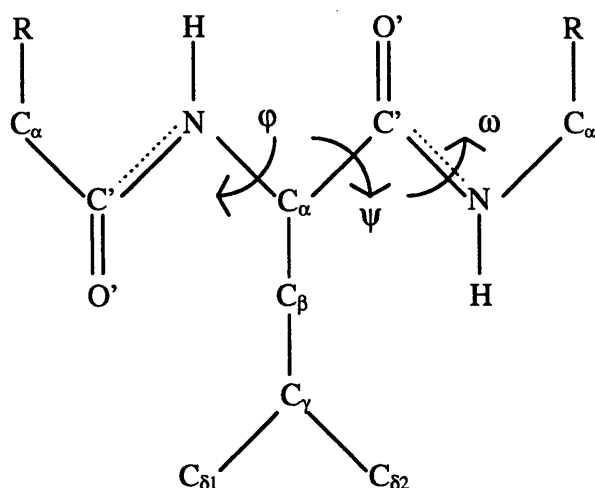
Name	3 Letter Code	1 Letter Code	Side Chain	Comments
Glutamate	Glu	E		acidic
Asparagine	Asn	N		uncharged polar, corresponds to aspartate with amidated sidechain
Glutamine	Gln	Q		uncharged polar, corresponds to glutamate with amidated sidechain
Lysine	Lys	K		basic
Histidine	His	H		basic and cyclic



Name	3 Letter Code	1 Letter Code	Side Chain	Comments
Arginine	Arg	R		basic, contains guanidinyll group
Phenyl - alanine	Phe	F		nonpolar, contains aromatic ring
Tyrosine	Tyr	Y		uncharged polar, contains a para substituted aromatic ring containing hydroxyl group
Tryptophan	Trp	W		nonpolar, contains a double aromatic and heterocyclic ring
Cysteine	Cys	C		uncharged polar, contains sulphhydryl group which may be oxidised to form S—S bridge between residues, then referred to as

Name	3 Letter Code	1 Letter Code	Side Chain	Comments
				cystine
Methionine	Met	M		nonpolar, contains sulphur atom but not as a sulphydryl group
Proline	Pro	P		nonpolar, this is the structure of the whole amino acid. Technically an imino acid as the side chain is bonded to the backbone nitrogen

**Table 1-1: The twenty natural amino acids found in proteins**

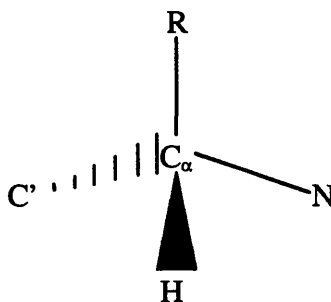


**Diagram 1-2: Nomenclature of the atoms in an amino acid**

The twenty common amino acids all consist of a carboxylic group and an amino group attached to the central carbon atom. This central carbon atom is usually labelled  $C_\alpha$ . The next atom out along the side chain is labelled the  $C_\beta$  carbon atom, the next labelled the  $C_\gamma$ ,  $C_\delta$  the next carbon atom to the end of the side chain, see Diagram 1-2 on page 16. The properties of the twenty amino acid side chains vary in size, shape, hydrophobicity, charge and hydrogen bonding, see Table 1-1 on page 15. The amino acids all share some properties, firstly they avoid the extremes of high chemical reactivity which would make them too non specific. Secondly, they also avoid groups which strongly restrict individual degrees of freedom. The exception to this second point is proline, which is technically an imino acid, due to the side chain forming a five membered ring with the backbone. The five membered ring in proline, which includes the backbone atoms, severely restricts its freedom of rotation.

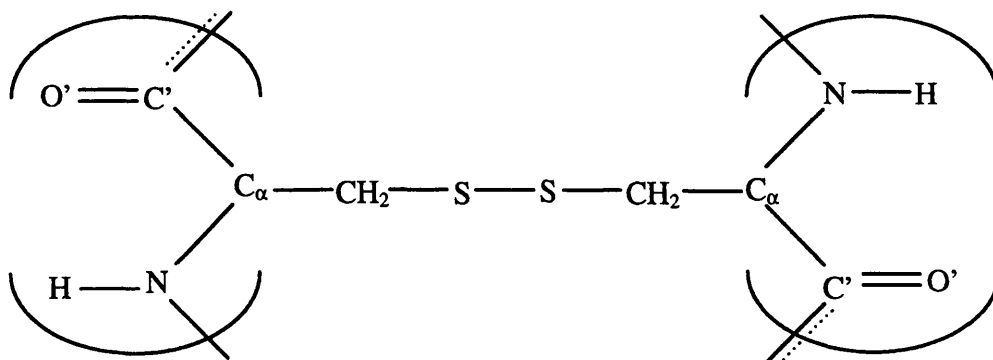
The central carbon atom,  $C_\alpha$ , is chiral for all amino acids except glycine which only has a hydrogen atom as its side chain atom, see Diagram 1-3 on page 17. Throughout nature only the one isomer of each amino acid is used, the L conformation. If the amino acid is viewed along the  $H-C_\alpha$  bond with the  $H$  atom at the front, the other substituents

connected to the  $C_{\alpha}$  atom, in a clockwise direction read CORN: C'-O'- carbonyl, R-side chain, and N- backbone. The amino acids Isoleucine (Ile) and threonine (Thr) have a second chiral centre at the  $C_{\beta}$  carbon atom, and again only the one enantiomer of Ile and Thr occurs naturally.



**Diagram 1-3: L chiral conformation**

In 1902 Emil Fischer and F. Hofmeister independently described how the amino acids are joined in a head to toe fashion to form a polymer. The reaction is a condensation reaction as water is produced in the reaction. The amino group of one amino acid and the carboxylic group of the next amino acid react to form a peptide bond. Therefore proteins are long linear polymers of amino acids. The only other covalent link between amino acid residues is due to the special ability of the cystine residue to form, under oxidising conditions, from two cysteine residues close together in the three dimensional structure of the protein. It is formed by the two sulphur atoms on each of the cysteine amino acid residues combining to produce a disulphide bridge, see Diagram 1-4 on page 18. The disulphide bridge is an important chain cross-linking feature usually within a single polypeptide chain but occasionally between two different chains. It appears as a common feature of proteins secreted from the cell, which has a reducing environment, to the outside with oxidising conditions.



**Diagram 1-4: The cystine residue containing the disulphide bridge**

Whether a chain of amino acids is called a protein, polypeptide or simply peptide is mainly down to the number of amino acid residues the chain contains and its conformation. There are many important biological molecules which contain relatively few amino acid residues joined together. Such short polypeptides are generally called peptides. A prefix denotes the number of residues i.e. a dipeptide has two residues, a decapeptide has ten, and an oligopeptide an unspecified small number. The term polypeptide used to be applied to a chain of a large number of amino acid residues but less than the number of amino acid residues found in an average sized protein but increasingly this view is changing. A polypeptide is now often considered as the most general term to describe a protein molecule. This is how the term polypeptide will be used in the rest of the thesis.

## 1.2 The Primary Sequence

The primary sequence or primary structure of a protein is the specific sequence of the amino acid residues characteristic of each protein. As the chemical structure of each type of amino acid residue is well known, this implies the chemical formula and covalent structure of the protein is known. Up until the 1950's scientists doubted that a protein had a primary sequence at all. Pre 1950's it was believed that any protein of a specific name actually consisted of quite different chemical entities all of which having the biological function of the named protein. This viewpoint was completely demolished with the

advancement in techniques to purify and determine the sequence for a very large number of proteins.

The primary sequence of a protein can be said to hold the information the protein requires to fold into its three dimensional (3D) shape. This can easily be shown by carrying out refolding experiments. In these experiments the purified active protein is denatured, that is unfolded, by the use of chemicals, usually urea or guanidine chloride, and heat. As the protein unfolds it loses its characteristic absorption spectra. Once the protein is denatured the conditions under which it was denatured are removed. The refolding of the protein can be followed by watching its absorption spectra revert back to its original signature. Once the protein sample has been allowed to refold its functionality normally returns.

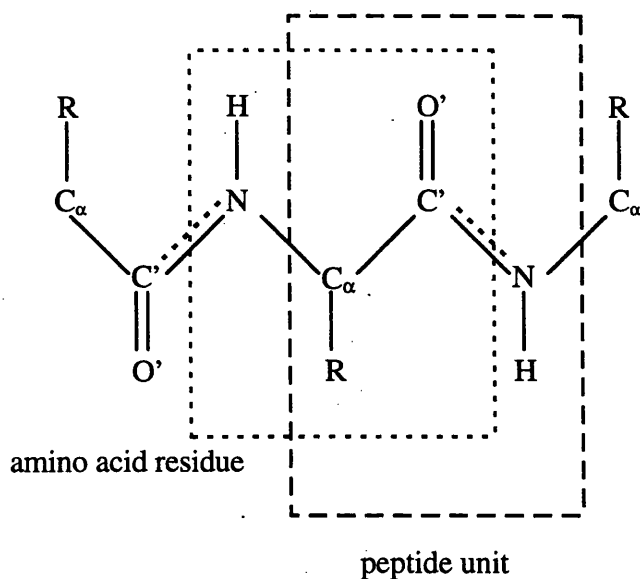
The primary sequence also determines the specificity of an enzyme. Using molecular biology techniques it is possible to selectively alter the proteins primary sequence in any position required. This enables scientists to probe how the enzyme works by altering key amino acids and seeing how this effects the rate of catalysis or the substrate specificity. This technology allows the mechanism of the enzyme reaction being studied to be worked out.

### **1.3 The Secondary Structure**

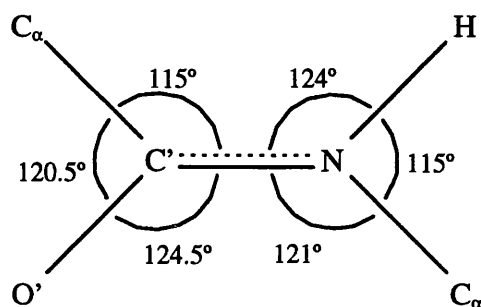
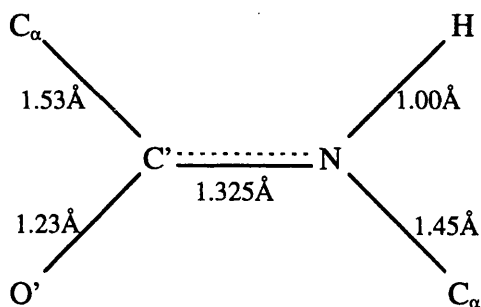
The description of a polypeptide conformation involves the specification of bond lengths, bond angles, and torsion angles about the single bond. A specific convention has been adopted to describe polypeptide conformation. The International Union of Pure and Applied Chemistry and the International Union of Biochemistry (IUPAC - IUB) Commission on Biological Nomenclature has proposed a detailed set of recommendations. The original proposals<sup>12</sup> were modified in 1970<sup>34</sup>.

The proposal defines an amino acid residue as  $-(\text{—NH—CHR—CO—})-$  and peptide units as  $-(\text{—CHR—CO—NH—})-$ , see Diagram 1-5 on page 20. The peptide unit corresponds to the properties of the peptide bond. The peptide bond has characteristics somewhere between a single and double bond due to the delocalisation of electrons between the oxygen and nitrogen atoms of the peptide bond. This means that the N—C' bond length is smaller

than expected for a normal single bond, the barrier of rotation for the  $C_\alpha-N-C'-C_\alpha$  torsion angle is increased from that expected of a single bond by  $\sim 25 \text{ kcal mol}^{-1}$ , and the  $C_\alpha-N-C'-C_\alpha$  torsion angle is planar. The double bond nature of the peptide bond severely limits the conformations possible the backbone in a protein can adopt. The peptide bond is nearly always found in the trans configuration and is unable to rotate at physiological temperatures. The torsion angle across the peptide bond is known as the  $\Omega$  angle. It is normally restricted to the values of  $0^\circ$  (cis conformation),  $180^\circ$  (trans conformation) or values very close to them. Except for proline the peptide bond is nearly always trans. Since the  $\Omega$  angle is fixed this leaves only two torsion angles free to rotate in the protein backbone. The two torsion angles that are free to rotate in the polypeptide backbone are known as the  $\phi$  and  $\psi$  torsion angles. The  $\phi$  angle defines the  $C'-N-C_\alpha-C'$  torsion angle and  $\psi$  defines the  $N-C_\alpha-C'-N$  torsion angle.



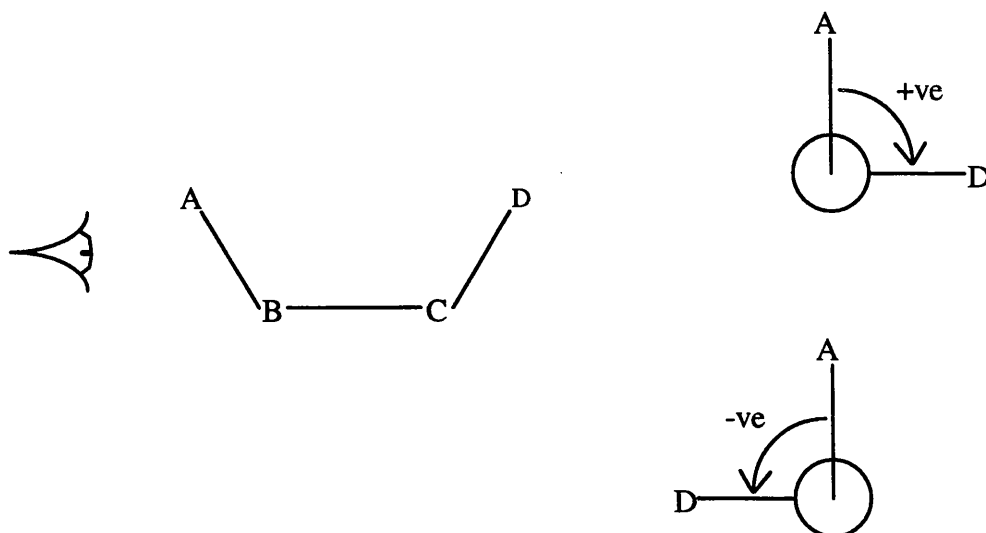
**Diagram 1-5: The amino acid residue and the peptide unit**



**Diagram 1-6: The peptide's dimensions**

The bond lengths and bond angles are generally found to be fixed values. For the peptide unit these values were first determined by Corey and Pauling<sup>5</sup> and revised values were published by Momany et al<sup>6</sup>, see Diagram 1-6 on page 21. The only degrees of internal freedom are those of rotation around the single bonds, see Diagram 1-2 on page 16.





**Diagram 1-7: Representation of the dihedral angle using the Newmann Projection**

### 1.3.1 The Hydrogen Bond

A hydrogen bond occurs when two electronegative atoms compete for the same hydrogen atom. The hydrogen atom is formally bonded covalently to one of the atoms, the donor D, but it also interacts favourably with the other atom, the acceptor A. The main component of the hydrogen bond is an electrostatic interaction between the hydrogen atom covalently bonded to the donor atom, and the electronegative acceptor atom. This arises due to the dipole along the covalent bond between the electronegative donor atom and the hydrogen which results in the hydrogen atom having a partial positive charge.

The hydrogen atom is special in being able to interact strongly with one electronegative atom while being covalently bonded to another. It can do this because of its small size and its substantial charge, which results from its tendency to be positively polarised. In strong hydrogen bonds an additional covalent aspect arises from a transfer of electrons.

The length and strength of hydrogen bonds depends on the electronegativities of the acceptor and donor atoms. The greater this electronegativity is the shorter the distance

between them and the stronger the hydrogen bond. Charged groups also give shorter and stronger hydrogen bonds.

In proteins the dual hydrogen bonding capacity of the backbone peptide group plays a large role in influencing the protein structure e.g.  $\beta$  sheets and  $\alpha$  helices are stabilised by hydrogen bonding. Although hydrogen bonds are weak, noncovalent interactions ( $2-10 \text{ kcal mol}^{-1}$ ), they are fairly directional and specific. Since each peptide can form a hydrogen bond in both directions, the co-operative effect of a network of such interactions can hold the polypeptide together in a strong specific framework. Hydrogen bonding can involve electrostatic interactions; either between actual charges on two amino acid residues, which is more commonly known as a salt bridge, or between dipoles such as the dipeptide dipole, which puts a partial positive charge on the NH and a partial negative charge on the oxygen carbonyl.

The optimum distance for a strong hydrogen bond is about  $3.0 \text{ \AA}$  between the donor and acceptor or  $2.0 \text{ \AA}$  between the hydrogen atom and acceptor<sup>7</sup>. For a charged hydrogen bond this distance can be smaller. The electrostatic part of the interaction only falls off as  $1/d^8$ , so there is still an effect at much greater separations, but beyond a certain point other atoms begin to intervene. The angle between the D—H—A atoms also matters, but again the energy fall off is gradual. The D—H—A angle is fairly critical, with an optimum at  $180^\circ$  and falling to no interaction at  $\sim 120^\circ$ . For the H—A—C angle (where C is the carbon atom to which A is covalently bonded), there are usually optima in the  $120^\circ - 150^\circ$  range, but the interaction is still strong at either  $90^\circ$  or  $180^\circ$ . On the surface of a protein only a hydrogen bond with very good geometry is useful because of the competition with solvent hydrogen bonds, but in the interior even a very long hydrogen bond is better than none at all.

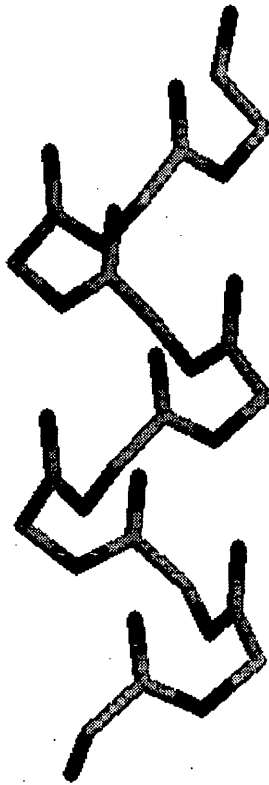
### 1.3.2 The $\alpha$ Helix

The first attempts to carry out x-ray diffraction analysis on proteins were carried out on fibrous proteins as these naturally formed paracrystalline solids. This analysis of fibrous proteins was carried out in the 1930's by Astbury and colleagues. Unfortunately paracrystalline materials do not diffract x-rays very well but the data did suggest that each

molecule is itself a repeating structure, with an organisation repeating itself by translation and rotation along the molecular axis.

Due to this early work by Astbury it became clear that fibrous proteins could be assigned to one of two classes. The  $\beta$  class had a diffraction pattern consistent with a molecular length of  $3.61\text{\AA}$  per residue, which is almost the same length as a fully stretched chain which would have a molecular length of about  $3.7\text{\AA}$  per residue. The classical material studied was silk fibroin but it was also noted that the keratin of bird feathers gave a similar diffraction pattern. The  $\alpha$  class has a x-ray diffraction pattern consistent with a helical pitch of  $5 - 5.5\text{\AA}$ , which suggests that these proteins must be considerably more contracted than an extended chain and having a molecular length of  $1.5\text{\AA}$  per residue. The classic material used in studies was hair keratin.

By studying in detail the x-ray data for much simpler amino acids and small peptides, the dimensions of the atoms and the bonds connecting them were estimated. In 1951 a series of papers published by Pauling, Corey and Branson reported how the theoretical approach had reduced the possible interpretations of the fibre diffraction patterns to a very few probable ones. The model which was consistent with the  $5 - 5.5\text{\AA}$  pitch was the  $\alpha$  helix. The model produced had a right handed thread in screw as a left handed helix of L amino acids would have atomic overlaps between  $C_{\beta}$  atoms and the backbone carbonyl groups. It has a pitch of  $5.41\text{\AA}$  per turn and 3.6 residues per turn. Realising that the number of residues per turn did not have to be integral was obviously an important factor in producing this model. The diameter across the structure was about  $10 - 12\text{\AA}$ , corresponding to the thickness of a fibrous molecule with this conformation. The structure was well packed without either overlaps or significant gaps between atoms. In particular, there was no significant tunnel down the axis of the structure. A major feature, largely responsible for choosing this model in preference to other models, was that all backbone hydrogen bonds formed nicely. In the  $\alpha$  helix a hydrogen bond is made between the  $O_{\text{carbonyl}}$  of every (i)th residue and the NH group of every (i+4)th residue. These hydrogen bonds act as 'staples' holding successive turns of the helix together. Three hydrogen bonds exist in each turn of the  $\alpha$  helix, see Diagram 1-8 on page 25.



**Diagram 1-8: simple  $\alpha$  helix, showing hydrogen bonds between residues  $i$  and  $(i+4)$ .**

In an  $\alpha$  helix all the residues are orientated in the same direction, each turn has three NH groups pointing in the N terminal direction and three CO groups pointing in the C terminal direction. This orientation explains, if qualitatively, the strong features of the CD spectra of many globular proteins if the  $\alpha$  helix was an important feature of globular proteins as well as fibrous proteins. Hence it was assumed that the  $\alpha$  helix model had to be directly relevant to at least certain aspects of globular protein structure.

The first globular protein to have its structure elucidated by x-ray diffraction were myoglobin in 1959. Myoglobin resembles a helical protein which has been bent at certain 'hinge' positions so that the helical regions can be brought together in space. Thus confirmation that the secondary structure feature  $\alpha$  helices play an important role in globular proteins.

### 1.3.3 The $\beta$ Sheet

Another model proposed by Pauling and Corey could account for the  $\beta$  class diffraction pattern observed for the x-ray diffraction of silk fibroin. It was known that the diffraction pattern corresponded to a structure where the chain was nearly fully extended but how could it possibly form hydrogen bonds to stabilise the structure. A stretched structure would not have its NH and CO groups aligned parallel to the axis of the molecule but approximately at right angles to the axis and therefore into the environment of the molecule. This is not electrostatically unfavourable but there should be extensive hydration as the hydrogen bonds form with the present water molecules. The model proposed by Pauling and Corey to get round this problem was that hydrogen bonds did not form within the same chain but across to an adjacent chain. Since this model fixed the direction of all NH and CO groups into an approximate plane, its basic feature consisted roughly of extended polypeptide chains lying side by side in a plane and hydrogen bonding to neighbouring chains. This model was described as a  $\beta$  sheet or  $\beta$  pleated sheet.

The theory at that time could not easily distinguish between  $\beta$  sheets in which all polypeptide chains were parallel, and sheets in which the chains were alternating to point along opposite directions in an anti-parallel arrangement. Good hydrogen bonding is possible in both cases and both models appeared equally stable. It turns out that even modern theoretical approaches do not readily distinguish between the two alternative models. Energy analysis of the two models show that the difference in energy between the two are subtle. This implies both structures are feasible, and indeed both are observed.

After myoglobin and haemoglobin the next protein to have its structure determined by x-ray diffraction was lysozyme by Blake et al in 1965. Unlike myoglobin, where all but the

'hinge' regions of the protein have  $\alpha$  helix conformation, lysozyme contains a relatively small number of residues forming three long  $\alpha$  helices and three distorted helices. The regions between the helices in the lysozyme structure are much longer than those found in myoglobin, with the most important section of chain which weaves back on itself to form three strands of distorted but quite recognisable  $\beta$  sheet. The strands, about six residues in length, were a small beginning to the recognition and importance of  $\beta$  sheets in globular proteins. It was the solving of the carboxypeptidase structure that demonstrated that the  $\beta$  sheet could form a major part of a globular proteins structure. Both parallel, anti-parallel, as well as mixed sheets have subsequently been identified in many proteins. As in lysozyme these are frequently distorted, and in general there is a twist to the sheets due to the handedness of the amino acids, which is of considerable interest. This twisted  $\beta$  sheet structure, in some proteins, looks as though it could bind to the surface of DNA molecules.

#### 1.3.4 The Reverse Turn

For periodic structures such as the  $\alpha$  helix and  $\beta$  sheet each residue has more or less, to a first approximation, the same conformation with the  $\phi$  and  $\psi$  angles repeating themselves throughout the structure. The aperiodic reverse turn, or  $\beta$  bend, structure is different in that each residue in the reverse turn structure has distinct  $\phi$  and  $\psi$  angles. The reverse turn was first described by Venkatachalam in 1968 as a sharp turn about of the direction of the polypeptide chain such that a hydrogen bond formed between the carbonyl of the (i)th residue and the amide of the (i+3)th residue. This structure depends on the conformation of only two residues, the (i+1)th and (i+2)th. Given that the peptide linkages in the reverse turn structure are trans, the middle two residues can have several sterically permissible  $\phi$ ,  $\psi$  angle combinations. Venkatachalam defined six types of reverse turns, types I, II and III, with their backbone mirror images I', II' and III'. These mirror conformations are generated from I, II and III by multiplying the  $\phi$  and  $\psi$  angles by minus one.

Type III reverse turn is identical to two residues in a classical  $3_{10}$  helix, while type I is a distortion of it. In type II the peptide group between the middle two residues, (i+1)th and (i+2)th residues, has 'flipped over' by variation of the  $\phi$ ,  $\psi$  angles either side, not of the peptide group itself which remains trans. In practise, the internally hydrogen bonded turns

are not always seen in proteins, so that the criterion is usually taken that the (i)th and (i+4)th residue are within 7Å of each other.

### 1.3.5 Random Coil

The final secondary structure is the random coil. This is a section of the polypeptide backbone which does not have regular repeating  $\phi$  and  $\psi$  angles. Although the  $\phi$  and  $\psi$  angles are said to be random they still only fall into the allowed regions of the Ramachandran map. Random coils are not found in the interiors of globular proteins but on the protein surface. They often protrude into the solvent and are very flexible. It is often difficult, occasionally impossible, to get the x-ray structure of these large highly flexible random coils because they are so flexible.

### 1.3.6 Ramachandran Maps

Due to the planarity of the peptide bond the conformation of the polypeptide backbone can be defined by the  $\phi$  and  $\psi$  torsion angles. A two dimensional plot of the  $\phi$  and  $\psi$  angles is an important type of representation. This plot is known as a Ramachandran plot. Such plots are used to illustrate properties of repeating conformations, single residues or two successive residues. The asymmetry of the amino acid residues (excluding glycine) due to the  $C_{\beta}$  atom causes the Ramachandran plot to be asymmetrical. The populated regions of  $\phi$   $\psi$  space are generally named after the conformation which results if the  $\phi$  and  $\psi$  values are repeated along the backbone. The major allowed areas are the right handed  $\alpha$  helical region in the lower left quadrant near (  $-60^{\circ}$ ,  $-40^{\circ}$  ); the broad region in the upper left quadrant centred around (  $-120^{\circ}$ ,  $140^{\circ}$  ) is the extended  $\beta$  strand; and the sparsely populated left handed  $\alpha$  helical region in the upper right quadrant near (  $+60^{\circ}$ ,  $+40^{\circ}$  ).

Vacant areas of  $\phi$ ,  $\psi$  space are conformations that place atoms unfavourably close together within the dipeptide unit e.g. near (  $0^{\circ}$ ,  $0^{\circ}$  ) the O' of residue n-1 collides with the C' of residue n. The bridge across the  $\psi = 0^{\circ}$  region joining the  $\alpha$  helix and  $\beta$  sheet regions should be uninhabited based on the hard sphere model but this region is fairly well populated. Using the hard sphere model there is bad contact between successive amide groups. This can easily be relieved either if the N— $C_{\alpha}$ —C' can stretch wider than the

tetrahedral or if the amide hydrogen is slightly 'soft' i.e. another atom can enter the amide hydrogen's van der Waals radius.

Energy calculations of a dipeptide unit give the energy minima in close approximation to the populated regions of  $\phi$ ,  $\psi$  space. This is rather surprising because such calculations leave out both the favourable and unfavourable effects of long range interactions of the backbone as well as specific side chain effects. One of the more remarkable properties of the repetitive secondary structures observed in proteins is that the optimum  $\phi$ ,  $\psi$  values for the permissible range of good long range H-bonding and steric fit are so close to the optimum and range for favourable dipeptide conformations. The presumption is that this neat match is what has, for instance, so strongly selected for the occurrence of the right handed  $\alpha$  helices rather than for any of the slightly different versions such as  $3_{10}$ ,  $\pi$ , or left handed  $\alpha$  helices.

### 1.3.7 The Handedness Of Proteins

Using the same configuration for the chiral centres in the 19 chiral common amino acids causes a handedness at all levels of the protein structure. This handedness is almost always through the interaction of the  $C_\beta$  atom. Helices are right handed because the left handed helices have a slightly unfavourable contact of the  $C_\beta$  of residue  $n$  with the preceding carbonyl of residue  $n-1$ , repeated for every residue in the left handed helix. For  $\beta$  strands the preferred twist is again due to the close contact of  $C_\beta$ s in the strand, but they act in a much more statistical fashion since they are distant in the sequence and can be compensated for by a combination of minor readjustments in other parameters. However, the effect of the strong statistical bias is persuasive and important causing the flaring, saddle shape swirl of parallel  $\beta$  sheets and makes the strands spiral around the axis of a  $\beta$  barrel. In reverse turns the differences in the common types of turns involve bumps of the O' of the central residue  $n+2$  with the  $C_\beta$  of residue  $n+3$ . The congestion is less severe than for a left handed  $\alpha$  helix, but the turn involves only one such awkward position, therefore a single glycine can solve this problem. This is one of the prime uses of the glycine residues in protein structure, to adopt conformations not accessible to other amino acid residues.



This use of glycine is because it is unique among the 20 natural amino acids in that it doesn't contain a  $C_{\beta}$  atom. Instead the  $C_{\alpha}$  atom has two hydrogen atoms attached to it.

The handedness preferences described above for secondary structure produce strongly handed features in the supersecondary and tertiary structure of proteins. Probably the most important of these is the right handedness of crossover connections which dominate the organisation of  $\alpha/\beta$  proteins. All  $\alpha$  proteins reflect the handed nature of helix - helix packing, which in turn reflects the handed spiral of side chains on the surface of right handed  $\alpha$  helices. Antiparallel  $\beta$  barrels are handed with twisted  $\beta$  strands, twisted barrel cross sections, and handedness in the swirl direction of their Greek Key motifs.

## 1.4 Super-Secondary Structure

This level of structure describes how the individual periodical secondary structures i.e.  $\alpha$  helix and  $\beta$  sheets come together in space. In considering the dynamics of the folding of a polypeptide chain the super secondary structure starts to form as favourable interactions between the individual secondary structures. This leads to a packing of the secondary structures. From a non-dynamic viewpoint, however, super secondary structure features can be regarded as a hierarchical level of structure representation between the secondary and tertiary levels.

Regular surface features on helices and pleated sheets become manifest if models are built from the one amino acid, say all alanine side chains. A pattern of grooves is apparent which implies a limited number of possible mutual orientations corresponding to the optimal meshing of the grooves from two secondary structures. These suggest preferred orientations for which there is some evidence when the distribution of orientations in known structures is examined. What is found in known protein structures is that the variation of sidechains within helices and sheets, with the flexibility of the side chains taken into account, causes the groove pattern to be nearly 'destroyed', and hence the simple packing rules to be lost.

On further analysis of the super-secondary structure features are strongly influenced by the chiral nature of the amino acids.

## 1.5 Tertiary Structure

The next level of structure is the tertiary structure which describes how the fragments of secondary structure and super-secondary structure are arranged about each other to give a protein its overall shape. For globular proteins the main driving force behind the folding of the polypeptide chain is the hydrophobic nature of many of the amino acids. The hydrophobic amino acids will be attracted to each other and form a hydrophobic core where all the water molecules have been expelled. Within this hydrophobic core nearly all the polar groups including those on the backbone of the protein are hydrogen bonded to other polar groups belonging to the protein. Only those polar groups found on the surface of a protein form hydrogen bonds with water molecules. The hydrophobic core of a globular protein is highly structured with nearly all the residues part of an  $\alpha$  helix or  $\beta$  sheet. The reason being that all possible hydrogen bonds in the backbone atoms are formed when a residue is folded into one of these secondary structures. The random coil secondary structure is not found in the hydrophobic core of globular proteins but on the surface. The reason being that not all the potential backbone hydrogen bonds are formed. This leaves polar groups exposed to the surrounding environment. In the core of the protein, where the environment is hydrophobic the polar group would not be able to make favourable interactions, whereas on the protein surface the exposed polar group can easily interact with the water solvent.

## 1.6 Description of Domains

The tertiary structure can also be viewed as the packing of individual domains of the protein. These domains are distinct structure within globular proteins. Domains are typically compact, rather hydrophobic clusters of residues formed by a local bunching up of the polypeptide chain of between 50 - 150 residues. They can be identified by this means although the problem is often knowing where to stop to avoid resolving domains into further domains. In more obvious cases, domains resemble protein subunits except they are not separate molecules but rather connected to each other by the continuity of the polypeptide.

There is often not much difference between the description of the domains of a protein and a description of the super secondary structures of the protein. Typically a domain may be primarily a single super secondary structure motif, or a set of smaller super secondary motifs. For example in serine proteases there are two domains with the active site of the enzyme represented in a cleft between the two domains. Each of the two domains in the serine protease is a  $\beta$  pleated sheet barrel and extensions to the polypeptide chain at the ends of the barrel.

## **2 The Complement System**

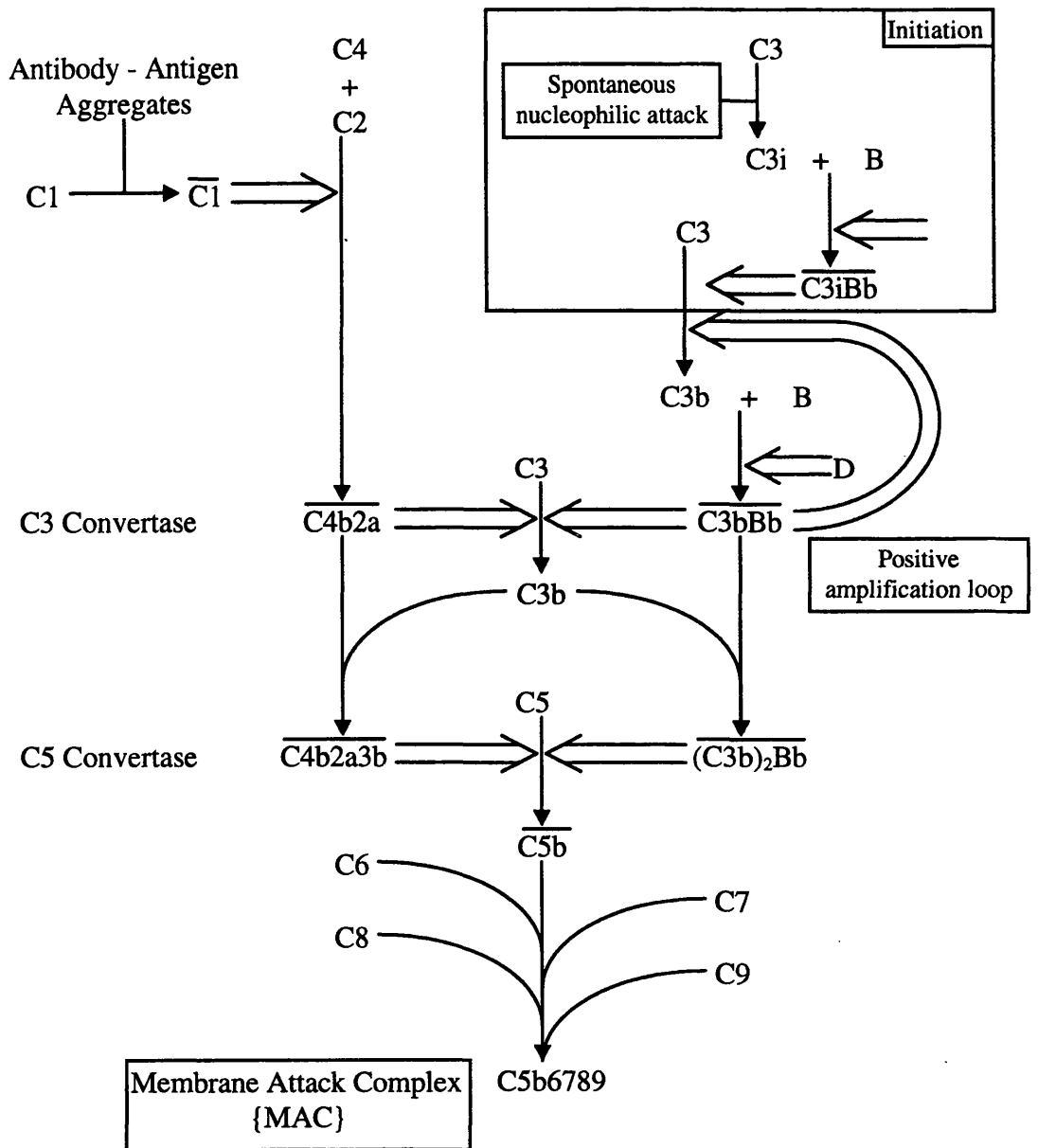
### **2.1 Introduction<sup>10,11,12</sup>.**

The complement system is an effector mechanism in the immune defence against infections by micro-organisms. It is a complete mechanism in which activation products of the complement components cause lysis of cellular antigens, attract phagocyte cells to the place of activation, and facilitate uptake and destruction by the phagocytes.

In the complement system there are two pathways. The Classical and Alternative pathways:

Classical Pathway

Alternative Pathway

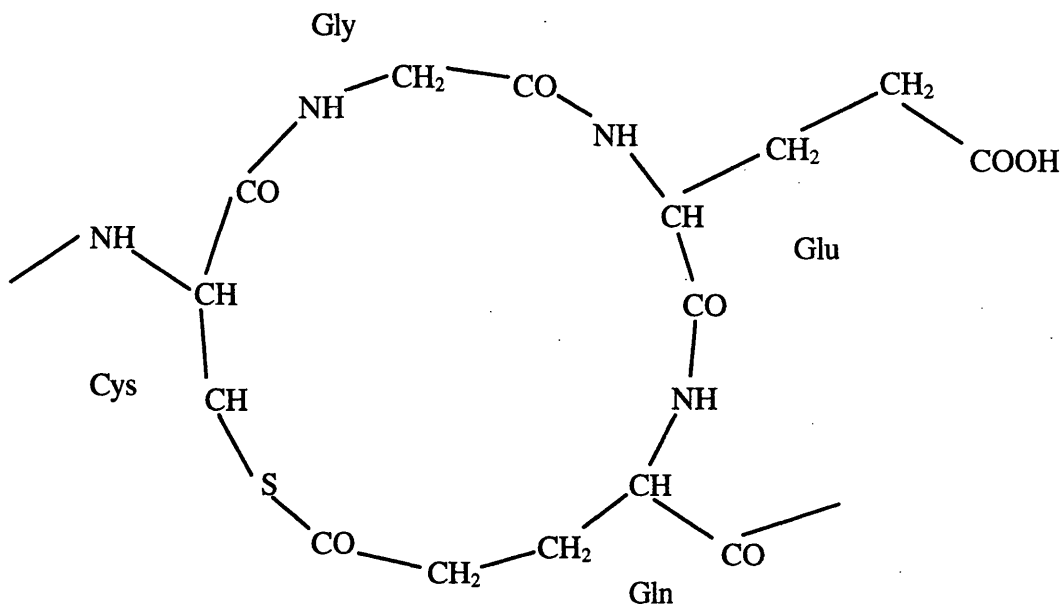


**Diagram 2-1: Overview of the Complement System**

## 2.2 The Alternative Pathway of the Complement System

The activation of the alternative pathway does not require any antibody - antigen aggregate to activate it. Neither does it only attack targets which have been specifically identified by antibodies. Instead C3b is attached indiscriminately to all particles, including the hosts own cells. A system of control factors then rapidly inactivate the C3b molecules bound to the host cells or other non-activating particles, but on surfaces recognised by the system as sites of activation these control factors work very slowly on C3b

The exact mechanism by which the first C3b molecules are produced is still unclear. The best documented mechanism of the initiation of the alternative pathway involves the spontaneous formation of the chemically and conformationally altered form of C3, C3i. It is known that C3 is continuously activated at a slow rate in the fluid phase. This is possible by small nucleophiles e.g. ammonia, or more probably water, that manage to gain access to the internal thiolester see Diagram 2-2 on page 36, or simply the perturbation of the C3 structure by any means leading to the exposure of the thiolester. C3 with a hydrolysed thiolester without the loss of the C3a fragment is called C3i, or C3(H<sub>2</sub>O). C3i has a molecular conformation similar to C3b and is able to form C3 convertase with Factor B (B) in the presence of Factor D (D). These processes operate at a very low rate (0.005% per minute) in aqueous solutions, and the probability that the activated C3b or C3i will bind covalently to a cell surface before being inactivated is very small. However, if C3b or C3i is deposited on an activating surface it can serve as a seed for the positive amplification loop.



**Diagram 2-2: Internal Thiolester Structure**

The activated thiolester of C3b is able to react with water, to be inactivated, and nucleophilic groups e.g. hydroxyl groups ( $\text{—OH}$ ) to form esters or amine groups ( $\text{—NH}_2$ ) to form amides. This allows C3b to be deposited onto any biological surface. Discrimination between non activating and activating surfaces is a result in the reduction in the effectiveness of the regulatory factors to control the amplification process when the initial C3b molecules are bound to the activating surface. In the fluid phase factor I with cofactor H cleave C3b to form inactive iC3b which can no longer form the C3 convertase with B. In contrast when C3b is bound to activating particles both C3b and the C3/C5 convertase are relatively protected from inactivation by the fluid phase regulatory proteins. This appears to be determined by how effective factor H can interact with the surface bound C3b. C3b bound to activating particles exhibits a reduced affinity for factor H, while the binding of B, factor I and properdin to C3b is unaffected. This suggests that the ability to distinguish activating and non activating surfaces is a property of C3b or factor H, or is expressed jointly by these two proteins at the surface of a particle see Diagram 2-3 on page 37.

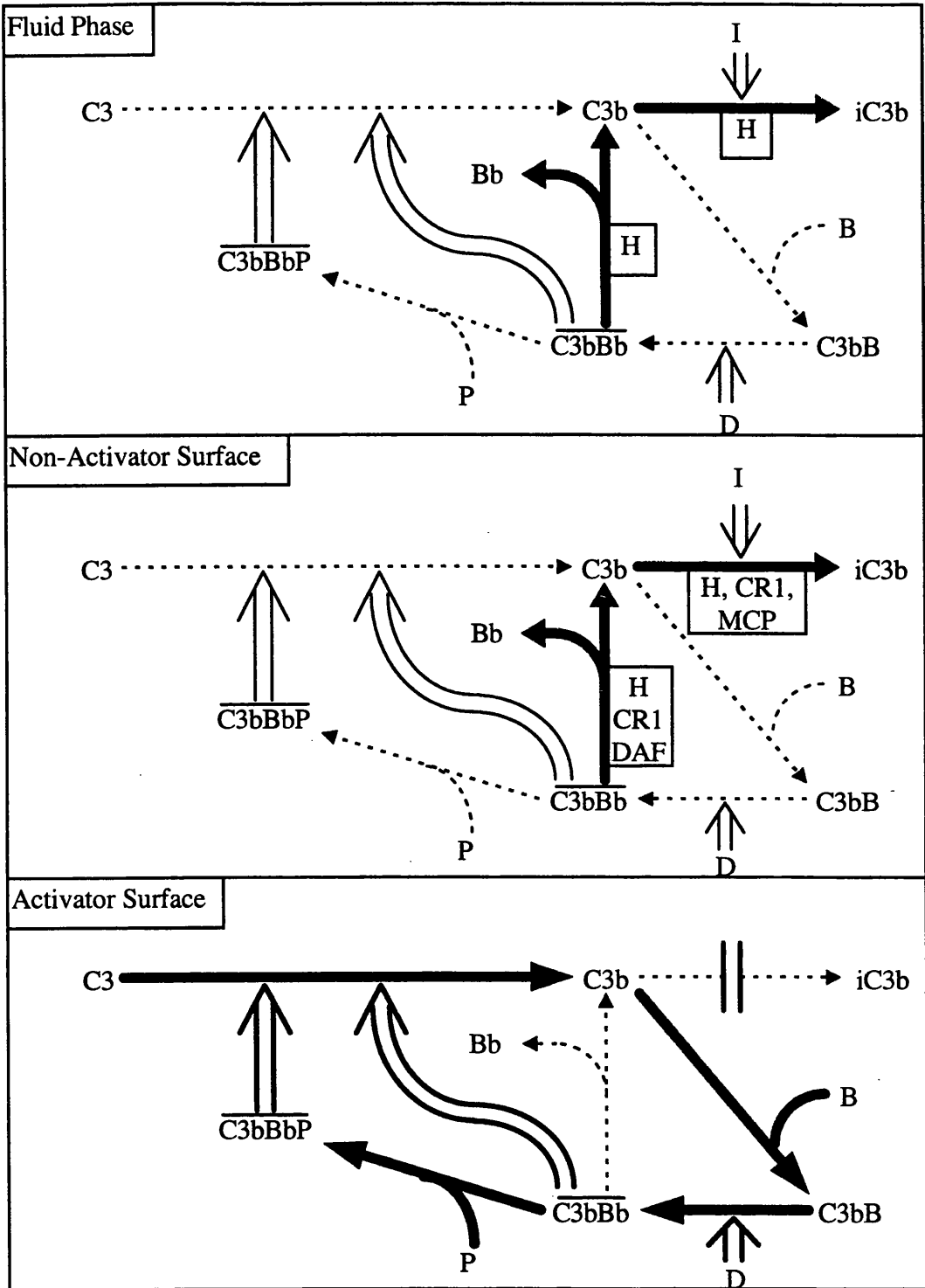
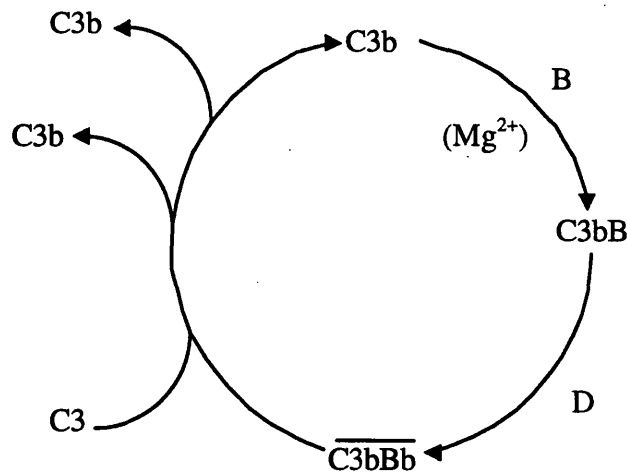


Diagram 2-3: The different pathway C3 follows in different environments.



It is not yet clear what molecular structures are recognised by the alternative pathway. Activators include many pure polysaccharide, lipopolysaccharides, certain immunoglobulins, viruses, bacteria, fungi, animal tumour cells and parasites. The only common factor of these activators is the presence of carbohydrates, but the complexity and variety of the different carbohydrate structures makes it difficult to envisage the shared molecular determinants which are recognised. One feature shared by most activators is the absence of surface sialic acid. Certain extremely weak activators (e.g. sheep erythrocytes) of the alternative pathway can be converted to efficient activators by the removal or modification of surface sialic acid. This has been observed in many systems establishing the connection between low sialic acid content and activation. However low sialic acid concentration is unlikely to be the only crucial factor of activation, since in several systems activators have been generated by the addition of foreign molecules to cell surfaces without the removal of sialic acid.

When C3i or C3b is bound to activators the reduction in the effectiveness of the regulatory factors allows the C3b dependent positive feedback process to occur, see Diagram 2-4 on page 38.



**Diagram 2-4: Positive amplification loop for the activation of C3**

The enzyme primarily responsible for the amplification is C3bBb<sup>\*</sup>, or the properdin stabilised form of the enzyme. The two subunits C3b and Bb are non-covalently bound. The pro-enzyme C3bB is activated by D which cleaves B only when B is bound to C3b. This cleavage produces Ba which is released into the serum and Bb which remains bound to C3b. The enzyme C3bBb is inherently labile, and spontaneous dissociation of the subunits results irreversible loss of enzymatic activity. Properdin can bind to the C3 convertase complex increasing its stability, slowing its dissociation. Both spontaneous dissociation and factor H accelerated dissociation of the Bb subunit are slowed five to ten fold. This makes the half life of PC3bBb bound to the activator surface approximately two orders of magnitude longer than that of C3bBb on non activating surfaces. This leads to efficient C3b deposition on the activator particles and accounts for the high rate of C3b deposition that occurs during the amplification phase.

PC3bBb cleaves the  $\alpha$  chain of C3 at exactly the same site as C4b2a to release C3a and leave C3b. Therefore C3b, a product of the enzymatic action of C3 convertases on C3, is itself a constituent of the alternative pathway's C3 convertase. Thus a positive feedback loop is established which, in the absence of regulatory factors, should continue to cleave C3 until the supply of either C3 or B becomes exhausted.

When additional C3b molecules are present C3bBb can function as a C5 convertase. The role of these C3b molecules is the same as the role C3b plays in the classical pathway, that is, to bind C5. The alternative pathway is an efficient activator of C5 because of the large number of C3b molecules bound to the surface and the numerous C3/C5 convertases formed on activating surfaces.

---

\*The normal convention to show that this is an active enzyme is to put an overline above the enzyme name (as in the diagrams in this chapter). This cannot easily be done in the document and I have therefore adopted a double underline to show the active enzyme.

## 2.3 The Proteins

### 2.3.1 C3

C3 holds a key position in the complement system as it is at the activation of C3 that the Classical and the Alternative pathways merge. In the classical pathway C3 is cleaved by C42a into C3a and C3b. C3a is released into the serum, and C3b is bound to the C3 convertase C42a to form the C5 convertase C42a3b. C3a is an anaphylatoxin that consists of the first 77 amino acids of the  $\alpha$  chain of C3. The removal of C3a causes a conformational change in the C3b fragment of the molecule to expose an internal thiolester. this thiolester is buried and inaccessible in active C3. It consists of a  $\gamma$  - carbonyl group of a glutamic acid residue and a thiol group on an adjacent cysteine. Covalent bond formation results from the transfer of an acyl group from the thiol to a nucleophilic group contained on suitable acceptor molecules, such as polysaccharides.

## 2.4 The Classical Pathway

The classical pathway is considered to be activated *in vivo* primarily by the interaction of the C1q portion of the C1 complex with immune complexes or aggregates containing IgG or IgM. Activation of C1 can also be achieved by its direct interaction with a variety of polyanions, e.g. DNA, RNA, certain small polysaccharides, and viral membranes, but the physiological importance of this type of activation is unclear. C1 contains three subcomponents: C1q, whose function is the binding of C1 to immune complexes and membranes; and C1r and C1s which are proenzymes. The binding of the C1q subunit to antibody is followed by the autocatalytic conversion of C1r to an active esterase, which then converts C1s to a similar active enzyme. It is the C1s that is the active enzyme used to cleave C4 into C4a and C4b.

The three chain C4 molecule like C3 contains an internal thiolester. When C4 is cleaved by C1s the C4b fragment formed undergoes a conformational change in its structure to expose the thiolester. The thiolester is highly reactive, binding either with the target surface or with water. Due to the reactive nature of the thiolester C4b does not dissociate far from the C1 complex before being deposited on the target surface or becoming deactivated by the water

in the serum. The formation of surface bound C4b is an amplification step as each C1s can cleave many C4 molecules.

C2 is able to attach onto the surface bound C4b to form C4b2. C2 is cleaved by C1s in the bound C1 complex. Therefore only those C2 molecules within reach of the bound C1s can be cleaved, and the C4b2 complexes outside the reach of the C1s remain inactivated.

Free C2 in the serum can also be cleaved by C1s bound to surfaces but this cleaved fluid phase C2b is incapable of binding to C4b to become C3 convertase. This cleavage of C2 in the serum proceeds to a much greater extent than the cleavage of C2 bound to C4b on the surface.

The C3 convertase C4b2a has a very short half life due to the irreversible dissociation of C2a from the bound C4b. This dissociation is also hastened by factor I and cofactor C4 binding protein (C4bp). The C3 convertase cleaves C3 at one point to produce C3a and C3b. The C3a molecule is released into the serum. A conformational change occurs in the C3b fragment to expose a thiolester. It was first thought that C3b combined with the C3 convertase complex C4b2a to form the C5 convertase complex. It is now thought that there is probably the one enzyme (C4b2a) which cleaves C3 and C5. C3 can be cleaved when it is free in the serum but C5 must be bound to C3b, which itself must be bound to a surface, before cleavage by C4b2a can take place. The necessity for C5 to be attached to the surface bound C3b before it can be cleaved confines the activity of the complement cascade to the targets under attack.

## **2.5 Activation of the MAC (Membrane Attack Complex)**

The lytic activity of the complement system was the first well defined function attributed to the system and it is now well established that the five plasma glycoproteins C5, C6, C7, C8 and C9 undergo a hydrophilic to amphiphilic transition to produce the typical cytolytic complement lesions. Together the terminal components can produce a complex referred to as the MAC. The MAC forms transmembrane channels which displace lipid molecules and other constituents, thus disrupting the phospholipid bilayer at target cells leading to osmotic cell lysis.

Proteolytic activation of C5 is achieved by the classical (C4b2a3b) or alternative (C3bBb3b) pathway C5 convertase. On binding to C3b, C5 undergoes a slight conformational change making it susceptible to proteolytic cleavage by the C5 convertase. Cleavage of C5 liberates C5a, where as the C5b fragment remains bound to the C3b molecule.

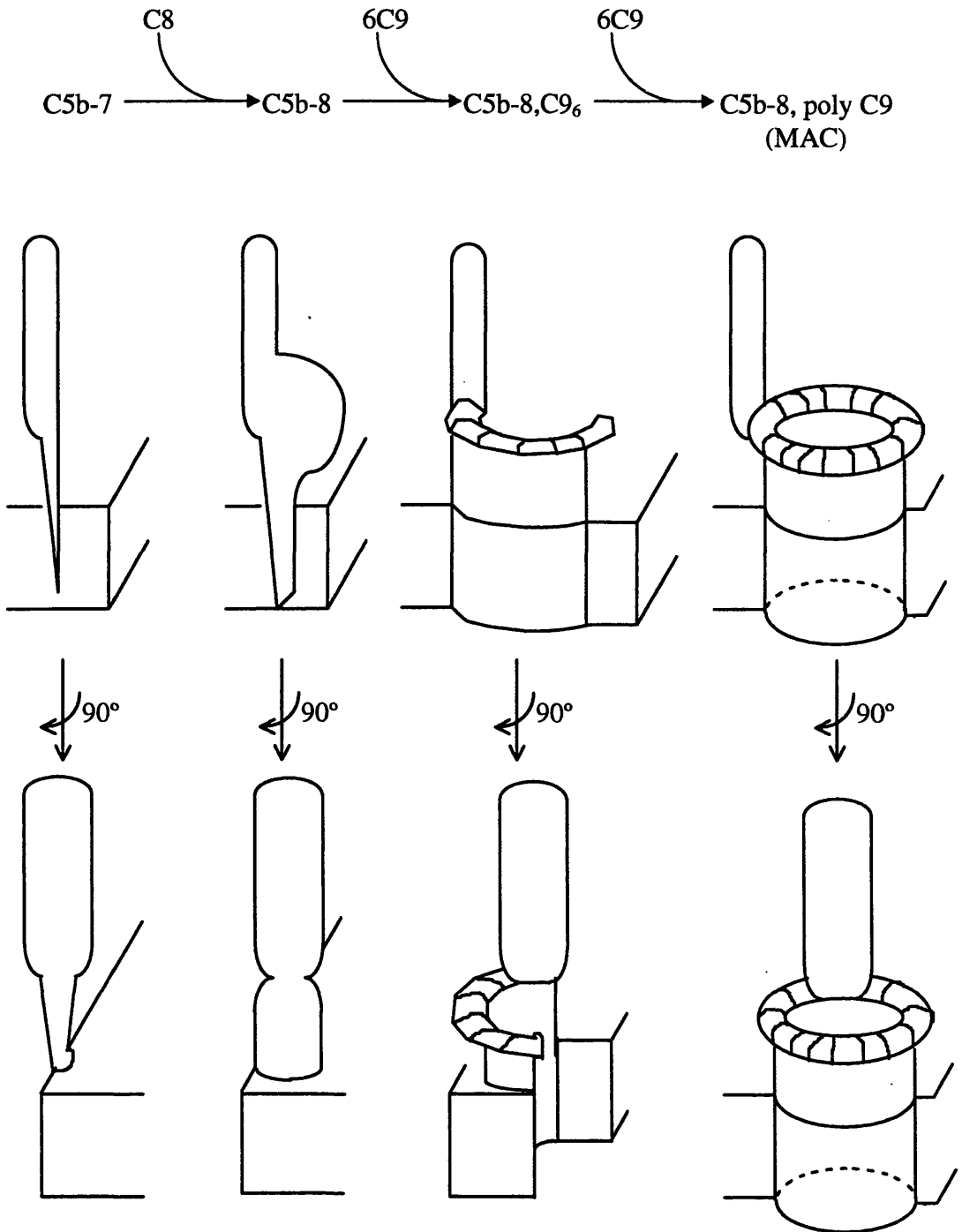
The newly activated C5b, loosely bound to C3b, binds C6 to form a C5b-6 complex and then to C7 to form a C5b-7 complex. The binding of C7 to C5b-6 complex causes an irreversible transition of the hydrophilic precursor proteins to the amphiphilic C5b-7 complex. This transition is accompanied by an increase of  $\beta$  pleated sheet structure which exposes the previously internal hydrophobic domains. Since this change occurs adjacent to the surface of C3b bearing membranes, the forming C5b-7 complex is probably in a steric configuration that allows the newly formed amphiphilic domain of C5-7 to insert itself immediately into the lipid bilayer membrane.

If the activation surface is not part of a phospholipid membrane, such as the surface of immune complexes, the membrane binding domain of C5b-7 has no substrate for hydrophobic insertion and the complex is released into the serum. This released C5b-7 complex presents a potential hazard to the host, because it has the capability of inserting itself into the membranes of neighbouring cells. Lysis of host cells is prevented by several C5b-7 inhibitors found in the plasma. These inhibitors bind to the released C5b-7 complexes and render them incapable of inserting into membranes. The C5b-7 complex bound to single bilayer phospholipid vesicles is visualised in the electron microscope as a 250 - 300Å long rod. The hydrophobic membrane binding site consists of a narrow domain and appears inserted into the lipid bilayer. C5b is located distant to the membrane binding site, where as both C6 and C7 appear to participate in the hydrophobic domain.

The binding of the three chain C8 molecule to C5b-7 takes place via a specific C5b recognition site on C8 $\beta$ . The  $\alpha$  and  $\gamma$  chains are covalently linked by disulphide bonds and non covalently associated with the  $\beta$  chain. Upon C8 binding to the C5b-7 complex the C8 $\alpha$ - $\gamma$  chain inserts into the hydrophobic core of the membrane. It is probable, therefore, that the hydrophilic to amphiphilic transition of C8 is restricted to its  $\alpha$ - $\gamma$  chain. The C5b-8

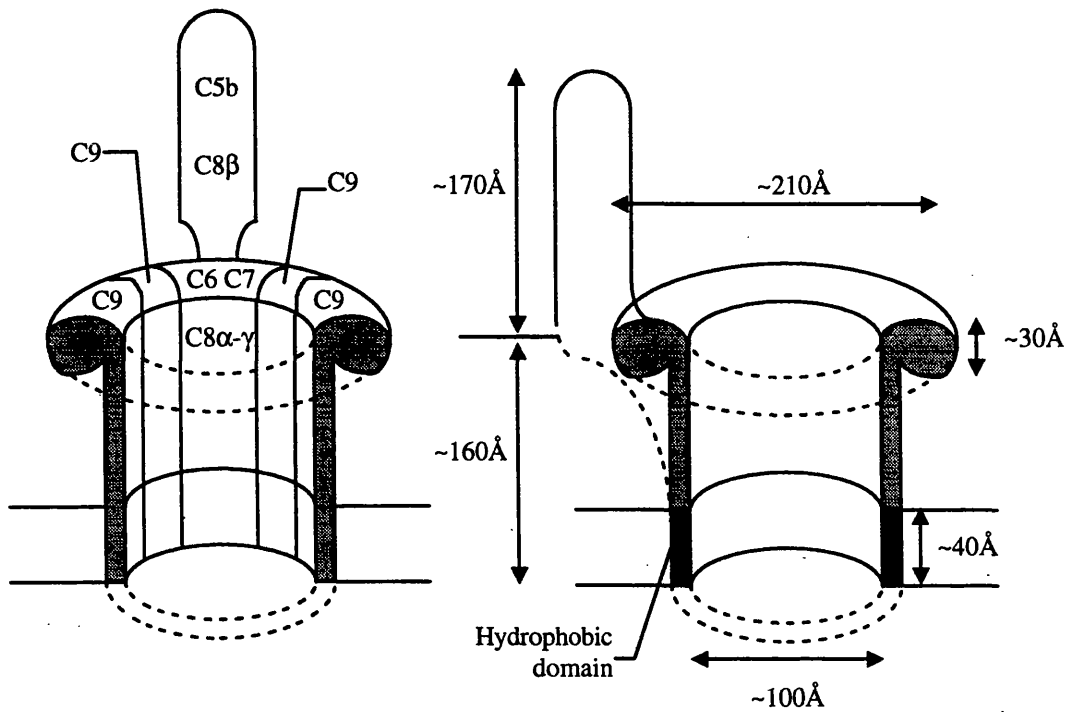
complex appears as a rod structure similar to C5b-7 but with a slightly increased width of the complex adjacent to the membrane surface indicating the location of the C8 in this area. Functionally, C5b-8 creates a small membrane pore with an effective parameter of  $\sim 10\text{\AA}$ . Therefore, the C5b-8 complex is capable of slowly lysing cells, but its principle role is to act as a receptor for C9 and to behave as a catalyst in C9 polymerisation to yield the highly effective C5b-9 cytolytic complex, see Diagram 2-5 on page 44.

C9 circulates in the plasma as a single chain, hydrophilic, globular protein of 50 - 80 $\text{\AA}$  in diameter. It binds and rapidly polymerises in the presence of the C5b-8 complex. As many as ten to sixteen molecules of C9 may be bound and polymerised by a single C5b-8 complex. As C9 polymerises it unfolds, giving rise to an increase in the  $\beta$  sheet structure and the expression of hydrophobic domains. The C9 polymerises to an amphiphilic, tubular complex. C5b-8 facilitates the insertion of polymerising C9 into biological membranes. This structure is known as the membrane attack complex (MAC). The tubular complex has the structure of a 160 $\text{\AA}$  long and 100 $\text{\AA}$  wide hollow tubule. The 40 $\text{\AA}$  long hydrophobic domain of the tubule is located on the outer surface of one end of the polymeric complex. The other end of the tubule is rimmed by a hydrophilic taurus with a 200 $\text{\AA}$  outer diameter. The inner surface of the tubule is hydrophilic. The C5b-8 complex is detectable as a 150 - 160 $\text{\AA}$  long appendage attached to the taurus of the poly C9 structure. The hydrophobic outer surface of poly C9 is inserted into the lipid bilayer to create a structural and functional lesion with an effective diameter of  $\sim 100\text{\AA}$ , see Diagram 2-6 on page 45.



**Diagram 2-5: Schematic drawing of C5b-7, C5b-8, C5b-8,6C9 and C5b-8,poly C9 bound to a single bilayer phospholipid. The drawings show two side views of the same membrane inserted complex rotated by 90°**

Each MAC structure occupies an area of  $\sim 10,000\text{\AA}^2$  in the membrane. This causes displacement of membrane constituents. The result of inserting large numbers of MACs, for example, into bacterial outer membranes causes an increase in the total surface area. The surface area of the membrane can increase by more than twofold when large numbers of MACs are inserted into it. This dramatic surface expansion may cause the loss of the structural integrity of attacked membranes. The displacement of the membrane constituents by MACs and the consequent physical alteration and surface expansion of attacked membranes may cause cell death independently of the effects caused by transmembrane channels.



**Diagram 2-6: 1)The subunit architecture of the MAC. 2) Dimensions of the MAC**

The pores created by the MACs in the outer membrane of bacteria allows access to and degradation of the peptidoglycan layer by lysozyme. Membrane pores also allow the entry



of  $\text{Ca}^{2+}$  ions into the intracellular space of cells which triggers indiscriminately a variety of cellular pathways. Pore formation is quickly accompanied by the breakdown of the membrane potential and by an efflux of  $\text{K}^+$  ions and an influx of  $\text{Na}^+$  ions. Rapid depletion of ATP and other high energy phosphates also occurs, but this may be due to compensatory ion pumping mechanisms along with cell activation by  $\text{Ca}^{2+}$  entry. All these effects may contribute to target cell death.

## **3 The Serine Proteases**

### **3.1 Enzymes In General**

One of the most important functions of proteins is to act as enzymes that catalyse specific chemical reactions. The binding of the substrate molecule to the enzyme is an essential prelude to the chemical reaction. Extremely high rates of chemical reactions are achieved by enzymes. This efficiency is attributed to several factors. First, the enzyme serves to increase the local concentration of the substrate molecules at the catalytic site and to hold the appropriate atoms in the correct orientation for the reaction that is to follow. Secondly, but more importantly, the free energy of the intermediate stages of the catalysed reaction are greatly reduced when these intermediate species are bound to the enzyme. This is especially the case for the most unstable transition states. Enzymes normally have a much greater affinity for the unstable transition states of the reaction than for their stable forms. By using the energy available in this highly favourable binding interaction, enzymes help their substrates attain a particular transition state, and thus greatly accelerate one particular reaction.

No matter how sophisticated an enzyme becomes, it cannot make the chemical reaction that it catalyses either more or less energetically favourable. As is true for any catalyst, natural or man made, a catalyst reduces the activation energy of a reaction. This increases the rate of both the forward and reverse reaction, making the reaction reach its equilibrium point much faster than if the catalyst was not present. The catalyst does not and can not shift the equilibrium point of a reaction.

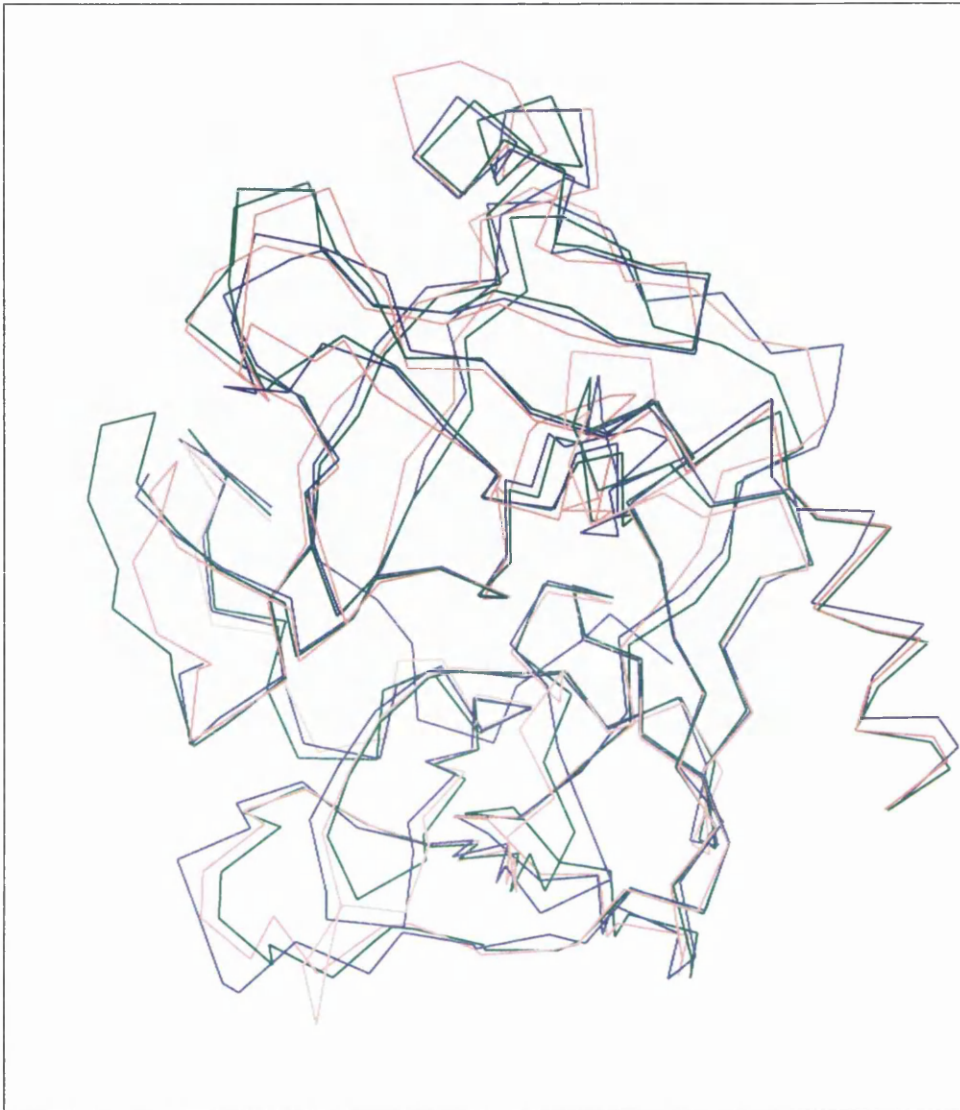
#### **3.1.1 Evolution**

There are over 1500 known enzymes which raises two main questions. The first is how many different types of structures do the different enzymes have and secondly how did such a large number evolve so quickly in the relatively short time there has been life on the planet.

##### **3.1.1.1 Divergent Evolution**

Divergent evolution can be seen most closely in the serine protease family of proteins. After the initial excitement of the discovery that the two oxygen binding proteins haemoglobin and myoglobin having a common tertiary structure as well as a common function the excitement was rekindled when the same was found out to be true of the mammalian serine proteases. The major pancreatic enzymes - trypsin, chymotrypsin and elastase are kinetically very similar, catalysing the hydrolysis of peptides and synthetic ester substrates. Their activities peak around pH 7.8 and fall off at low pH with a  $pK_a$  of around 6.8. In all three cases the reaction forms an "acylenzyme" through esterification of the hydroxyl of the reactive serine by the carboxyl portion of the substrate.

The major difference between the three enzymes is the substrate specificity. trypsin is specific for the peptides and esters of the amino acids Lys and Arg; chymotrypsin for the large hydrophobic side chains of Phe, Tyr and Trp; and elastase for the small hydrophobic amino acids such as Ala. When the crystalline structure of the enzymes were solved, it was found that the polypeptide backbones of all three are essentially superimposable (Diagram 3-1 on page 49), apart from some small additions and deletions in the chain. The difference in their specificities is due to just a few changes in a pocket that binds the amino acid side chain (section 3.2.3 on page 62).



**Diagram 3-1: Superimposition of porcine elastase (1esa) (red), bovine chymotrypsin (1gct) (blue) and bovine trypsin (1tpo) (green). The plot is a trace on the C alpha atom for each protein.**

The remarkable similarity of all three tertiary structures could not have been guessed in advance from a comparison of their primary sequences. There is extensive homology between the primary sequences, but the sequence identity and homology between elastase, trypsin and chymotrypsin is about 50% (Table 3-1 on page 50 showing sequence homology between trypsin, chymotrypsin and elastase). Closer examination of the sequence

homologies shows that 60% of the amino acids in the interior are conserved, but only 10% of the surface residues. The major differences occur in exposed areas and external loops. With such strong similarity between members of the serine protease family it is obvious the serine proteases arose from divergent evolution with a common structure which was duplicated by gene duplication which then specialised to catalyse a specific reaction.

<b>Enzyme</b>	<b>% Homology</b>
<b>Pancreas</b>	
Trypsin	100
$\alpha$ -Chymotrypsin	53
$\beta$ -Chymotrypsin	49
Elastase	48
<b>Plasma</b>	
Thrombin	38
Factor Xa	50

**Table 3-1: Sequence homologies in mammalian serine proteases**

Subsequently, some non-mammalian serine proteases were shown to be 20 to 50% identical with their mammalian counterparts (see Table 3-2 on page 51). It is now known that this suggests a very similar tertiary structure. For example the crystal structure of the elastase - like protease from *Streptomyces griseus* has been solved and despite having only 186 amino acids in its sequence, compared with 245 in  $\alpha$ -chymotrypsin it is found to have two thirds of the residues in a comparable conformation to those in the mammalian enzymes<sup>13</sup>. The possibility of an ancient ancestor between these bacterial enzymes and the pancreatic serine proteases exists but the evolutionary relationships are not clear.

Enzyme	Species	% Homology
Trypsin	Cow	100
	Dogfish	69
	Streptomyces griseus	43
Elastase	Pig	48
	Myxobacter sorangium	26
	S. griseus	~20
Subtilisin	Bacillus subtilis	0
	Bacillus amyloliquifaciens	0

**Table 3-2: Species differences in serine proteases.**

### 3.1.1.2 Convergent Evolution

The first crystal structure of a bacterial serine protease to be solved, subtilisin from *Bacillus amyloliquifaciens*, revealed an enzyme of apparently totally different construction from the mammalian serine proteases<sup>14</sup>. This was not unexpected, since there is no sequence similarity between them. But closer examination shows that they are functionally identical as far as substrate binding and catalysis are concerned. Subtilisin has the same charge relay system, the same system of hydrogen bonds for binding the carbonyl oxygen and the acylamido NH of the substrate and the same series of subsites for binding the acyl portion of the substrate as have the mammalian serine proteases (Diagram 3-7 on page 66). This appears to be a case of convergent evolution. Different organisms, starting from different tertiary structures, have evolved a common mechanism.

Another example of convergent evolution is that of the endopeptidase thermolysin from *Bacillus thermoproteolyticus* and the carboxypeptidases<sup>15,16,17</sup>. There is no sequence or structural similarities except that the active sites are very similar, containing in each case a catalytically important  $Zn^{2+}$  ion. The enzymes consequently appear to have similar catalytic mechanisms.

### 3.1.1.3 Convergence or Divergence

All the evidence points to the fact that the mammalian serine proteases have evolved through divergence, but that their common catalytic mechanism with subtilisin has developed through convergence. Other cases are not so clear cut. The accepted procedure for distinguishing between convergence and divergence is to count the number of common characteristics. If there are many then divergence is more likely, if there are few then convergence is more likely. These are<sup>18</sup>:

1. The DNA sequences of their genes are similar.
2. Their amino acid sequences are similar.
3. Their three dimensional structures are similar
4. Their enzyme - substrate interactions are similar
5. Their catalytic mechanisms are similar
6. The segments of polypeptide chain essential for catalysis are in the same sequence (i.e. not transposed)

These criteria are in descending order of strength. If 1 and 2 hold, the rest will follow, in most - but not all - cases. Lysozyme from hen egg and lysozyme produced by the bacteriophage T4 have no detectable similarities in their amino acid sequence. Yet, by showing that criteria 3 - 6 hold, a strong case has been made for divergent evolution<sup>19,20</sup>. Finally sometimes structure has been conserved through evolution but function has changed, that is criteria 3 and 4 do not hold. for example, the binding protein haptoglobin appears to have diverged from the serine proteases.

## 3.2 The Serine Proteases

### 3.2.1 Scope of the family

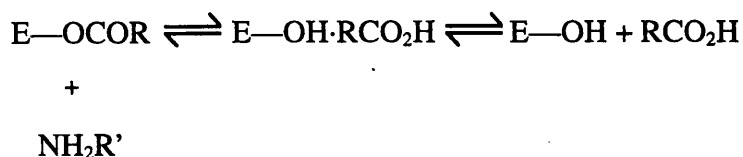
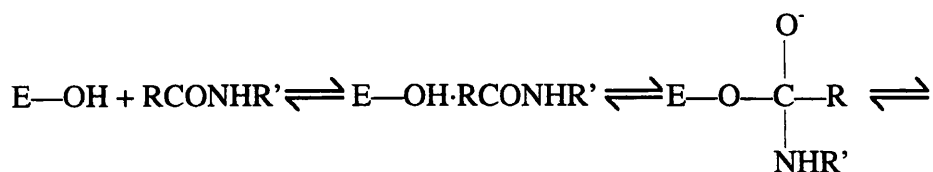
Of the many ways available to control the biological activities of proteins, e.g. induction or repression of their synthesis at the translational or transcriptional level, specific modification or destruction are the most direct. Many biological systems are controlled by such methods as these, and the serine protease family of enzymes plays a major role in many of these systems. The most widely studied of the serine proteases are the gastrointestinal serine proteases in higher animals, but serine proteases also play an essential role in blood coagulation<sup>21</sup>, in the Complement system, bacterial sporulation<sup>22</sup>, and fertilisation<sup>23</sup>.

Intrinsic to the process of digestion in mammals is the breakdown of dietary protein by the pancreatic serine proteases. These pancreatic digestive enzymes are among the most thoroughly studied of all enzymes, principally because they are extracellular enzymes that are easily separated and purified in large quantities.

### 3.2.2 The Reaction Mechanism

Peptide and synthetic ester substrates are hydrolysed by the serine proteases by the acylenzyme mechanism<sup>24</sup> shown in Diagram 3-2 on page 54:



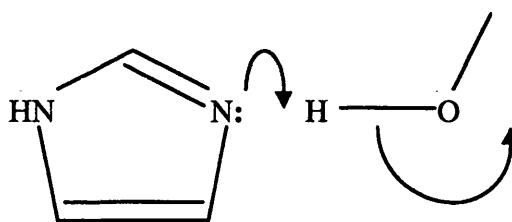


**Diagram 3-2: Acylation mechanism for the hydrolysis of a peptide or synthetic ester by a serine protease.**

The enzyme and substrate first associate to form a non-covalent enzyme - substrate complex held together by physical forces of attraction. This is followed by the attack of the hydroxyl of Ser 195 on the substrate to give the tetrahedral intermediate. The intermediate then collapses to give the acylenzyme, releasing the amine or alcohol. The acylenzyme then is hydrolysed to form the enzyme - product complex.

### 3.2.2.1 Mechanism before crystal structure

Before the crystal structure of chymotrypsin was solved it was known from solution studies of the serine proteases that the imidazole ring of His 57 increases the reactivity of Ser 195. The imidazole base of His 57 increases the nucleophilicity of the hydroxyl of Ser 195 by acting as a general base catalyst (see Diagram 3-3 on page 55). The activity falls off at low pH according to the ionisation of a base of  $\sim pK_a$  7, a characteristic value for a Histidine residue.



**Diagram 3-3: imidazole ring of His 57 acting as a base to increase the nucleophilicity of the hydroxyl group on Ser 195.**

### 3.2.2.2 Mechanism afterwards - The Charge Relay system

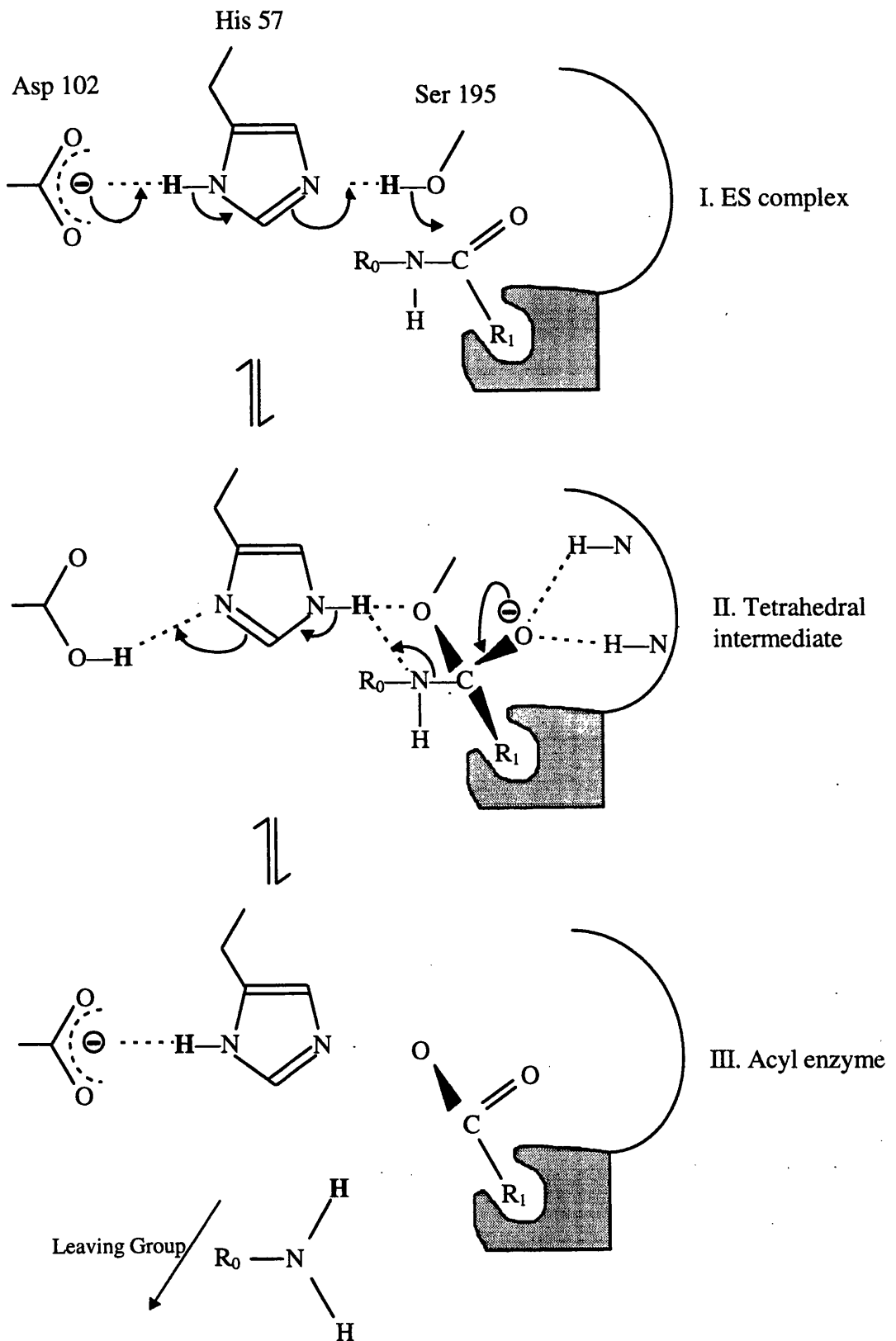
When the crystal structure of chymotrypsin was solved what was not expected was that Asp 102 played an important role in the catalytic reaction. This catalytic triad of Asp 102, His 57 and Ser 195 is now called the “charge relay system”<sup>25</sup>. This charge relay system is found in all serine proteases. Although the carboxyl group is completely buried in the interior of the protein, it is surrounded by polar residues and buried water molecules.

In this discussion it is assumed that the  $pK_a$  of Asp 102 is 6.8 and that the imidazole of His 57 is essentially neutral above pH 4.0. This leads to the ionisation of the active centre around pH 7. The mechanistic importance of these assignments is that the aspartate ion of residue 102 can act as a chemical base which can readily accept a proton from the histidine side chain during catalysis<sup>26</sup>. Together, Asp 102 and His 57 shuttle protons (charge) back and forth from enzyme to substrate, and so the mechanism can be best described as nucleophilic attack with general base catalysis by His 57<sup>27</sup> and Asp 102.

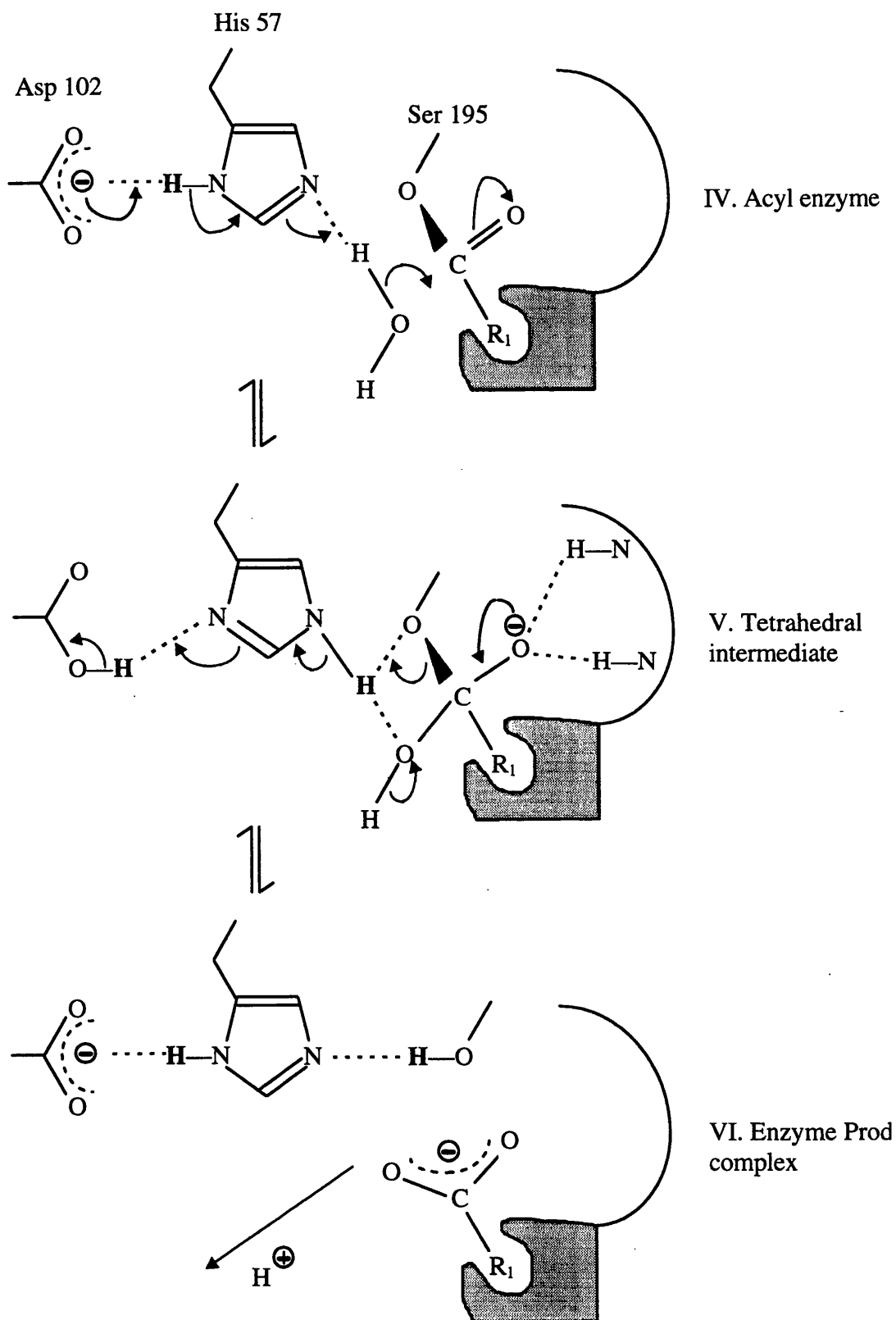
The mechanism scheme shown in Diagram 3-4 on page 57 and Diagram 3-5 on page 58 is consistent with most experimental data relating to the hydrolysis of peptides esters or amides by trypsin or chymotrypsin. In the first step (I), substrate and enzyme form a Michaelis complex. Nucleophilic attack by the hydroxyl group of Ser 195 follows. As the reaction proceeds, the hydroxyl twists around the  $C_\alpha-C_\beta$  bond and forms a covalent bond to the substrate carbon at step I-II. In concert with this, a proton is transferred from the Ser hydroxyl group to the  $N_{\epsilon 2}$  of His 57. From there it is eventually delivered to the N of the

peptide bond in the substrate. As a result of this proton transfer, the proton previously bound to the N<sub>δ1</sub> of His 57 is transferred to the carboxyl group of Asp 102, which acts as a base in this reaction.

Whether the Asp - His - Ser proton shuttle is concerted or stepwise remains in question. If the mechanism is concerted the negative charge of Asp 102 would be neutralised while negative charge develops on the carbonyl oxygen of the substrate. The imidazole ring would remain neutral throughout the reaction; thus unstable intermediates due to charge separation would be avoided<sup>28</sup>. In fact charge development in the transition state in chymotrypsin catalysis does appear to be small. The shuttle may be stepwise if the energy requirements of charge separation (negative charges on the substrate and Asp 102 and a positive charge on His 57) are offset by a more favourable entropy of activation in a two step process<sup>29</sup>.



**Diagram 3-4: The first three of six steps for serine protease hydrolysis of peptides or amides. In this representation the proton shuttle is concerted.**



**Diagram 3-5: The final three of six steps for the mechanism for serine protease hydrolysis of peptides or amides. In this representation the proton shuttle is concerted.**

One might favour the concerted mechanism because it might be expected that the precise alignment of the shuttle, which has been observed in all serine protease structures, evolved so that the entropic advantage of the two step process over the concerted process was minimised. Thus the enzyme could exploit for increased reaction rate the energy saved in eliminating charge separation. If this were not the case it would seem unnecessary to use both an Asp and an His for the general base catalysis. The Asp could be eliminated and the His could act as the base.

After the attack by Ser 195 on the substrate, a short lived tetrahedral intermediate is formed (II). This intermediate is stabilised by the covalent bond to the enzyme and by a number of hydrogen bonds. The following structural features of the tetrahedral intermediate are primarily from the crystallographic determination of many different protease - inhibitor structures.

The negatively charged substrate oxygen in the tetrahedral intermediate is stabilised by the hydrogen bonds from the amide Ns of residues 195 and 193<sup>30</sup>. Another hydrogen bond forms between the carbonyl group of Ser 214 and the  $\alpha$ -N of the substrate<sup>31</sup>. Comparisons of the kinetics of hydrolysis of specific trypsin and chymotrypsin substrates with and without the hydrogen bonding capacity of the  $\alpha$ -N suggests that the Ser 214— $\alpha$ -N bond may not form in the Michaelis complex. These results show, however, that this hydrogen bond does play a role in the transition states between the intermediates and possibly in the tetrahedral and acyl enzyme intermediates.

One explanation for the exceptional catalytic powers of enzymes is that the enzymes have evolved so that they optimally bind the transition - state structures in the reaction they catalyse rather than the substrates themselves. The hydrogen bonded structure in the serine protease - substrate transition state is an example of transition state stabilisation, for the oriented hydrogen bonds can help to speed up the reaction by smoothing down the highest barriers between the intermediate states. It is also known from studies on the serine proteases that the active site of the enzyme is complementary in structure to the transition state of the reaction, a structure that is very close to the tetrahedral adduct of Ser 195 and the carbonyl carbon of the substrate. Furthermore, the structure of the enzyme is not distorted when it binds to the substrate.

At step II - III, the now unstable C—N bond is broken and the first product of hydrolysis, an amine, is free to diffuse away taking with it a proton from the enzyme. At the same time, the bound part of the substrate rearranges to a chemically modified acyl enzyme intermediate (III). At pH 8,  $N^{14}/N^{15}$  kinetic isotope effects<sup>32</sup> show that the C—N bond rupture is partially rate determining for the hydrolysis of acetyl tryptophanamide by chymotrypsin. The rate determining step for amide hydrolysis, however, may vary from the formation of the tetrahedral intermediate to its breakdown, depending on the pH and the structure of the substrates.

The breakdown of the acyl intermediate (IV - VI) is the microscopic reverse of steps I - III, this time water is the attacking group. At step V - VI, the second product is formed. It is an acid which loses a proton to the solution and becomes negatively charged. For the first time (if the proton shuttle is concerted), there are two charges in the system. These two negative charges repel each other and help to dissociate the second product from the enzyme, regenerating free enzyme.

The presence of a carboxyl group of high  $pK_a$  and a neutral side chain of His 57 with a low  $pK_a$  would suggest two compelling evolutionary reasons why the Asp - His - Ser arrangement should be universal to serine proteases. First, by neutralising a negative charge on Asp 102 rather than generating a positive charge on His 57, during formation of the tetrahedral intermediate, there would be no unfavourable charge separation. This would contribute to reducing transition state internal energies, and thus to rate enhancement. Second, if the charged Asp 102 is to be a proton acceptor at physiological pH values its  $pK_a$  must be raised and it must have access to a proton donor. The imidazole of His 57 is ideally suited both to insulate Asp 102 from solvent (so raising the  $pK_a$  of the buried carboxyl group) and to serve as a proton conductor, transferring charge from the carboxyl group to the substrate. It is also important to note that both the reverse separation of the  $pK_a$  values of Asp 102 and His 57 and the structure of trypsin at pH 7 and pH 8, which shows a symmetric interaction between the charge on Asp 102 and His 57, are unlike the situation expected in aqueous solution and reflect a unique micro-environment for these groups.

### 3.2.2.3 Other possible mechanisms

Recent theoretical studies<sup>33</sup> point to a different mechanism than the one described above. The above mechanism involves the concerted transfer of two protons (step I in Diagram 3-4 on page 57) from Ser 195 to His 57 and from His 57 to Asp 102. The acceptance of this mechanism in the chemical community might have been motivated by the recognition that ion pairs are not stable in non polar environments so that Asp<sup>-</sup>102 His<sup>+</sup>57 must be less stable than Asp 102 His 57 in non polar active sites. This, however overlooks the fact that the active site around Asp 102 is very polar.

Most quantum mechanical calculations that considered only the catalytic triad without its surrounding protein supported the concerted charge transfer<sup>34</sup>. Calculations that attempted to include the effect of the protein active site on the catalytic triad<sup>35</sup> contradicted the concerted reaction mechanism, but these latter calculations did not include the key effect of the solvent around the protein and did not calibrate the intrinsic energy of the ion pair on reliable experimental information. This prevented a quantitative assessment of the feasibility of the concerted charge relay mechanism.

An alternative mechanism to the concerted double proton transfer mechanism is electrostatic catalysis. Instead of the proton shuttle of the concerted mechanism only the one proton moves position. It is the proton from the hydroxyl group on Ser 195 that is transferred onto His 57. This causes His 57 to develop a positive charge while both Asp 102 and the substrate carbonyl each develop a negative charge.

In order to clarify the role of Asp 102 the following three points have to be considered:

1. Asp 102 is left in its ionised form and serves to stabilise the ionic transition state.
2. Asp 102 is used to accept a proton from His 57 in the transition state.
3. Asp 102 helps in orientating His 57 into a proper position to interact with the substrate contributing entropically to the rate acceleration.

The results of using a microscopic model of the catalytic site shows the replacement of Asp 102 by a neutral residue results in destabilisation of the transition state by more than 4 kcalmol<sup>-1</sup>. This is because the replacement of Asp 102 by a non charged residue leads to a major reduction in the negative potential on His 57, destabilising the transition state.



Whether or not Asp 102 is used to accept a proton from His 57 in the transition state in the concerted proton shuttle mechanism could not be explored quantitatively by early quantum mechanical calculations<sup>36</sup> since the actual difference between the two options is smaller than the error associated with the neglect of the surrounding water molecules and with the errors associated with calculations of the intrinsic gas phase energy of the reacting fragments. However it is possible to calculate the difference between the two mechanisms without the uncertainty associated with quantum mechanical calculations of large systems. The calculations indicate that the concerted proton shuttle is strongly unfavourable as compared to the electrostatic model. To realise this from a simplified point of view, it is important to recognise that the ionised form of Asp 102 is even more stable in the protein active site than in water, as is apparent from its observed pK<sub>a</sub> being equal to 3 in chymotrypsin and the fact that this group is stabilised by three hydrogen bonds<sup>37</sup>. The more stable the negative charge on Asp 102, the less advantageous a proton transfer from His 57 to Asp 102 would be.

From mutations studies and free energy calculations on them it is not possible to rule out that the role of Asp 102 is to fix His 57 in the correct conformation for the catalysis. Mutating Asp 102 to Asn in trypsin gives a mutant where, in the crystal structure, His 57 adopts two different conformations. One of these conformations has the His rotated out of the catalytic site where it cannot interact with Ser 195. The population ratio of this rotated conformation to the one where His 57 is inside the active site is 1:2. However, the configuration with His 57 in the correct position appears to stabilise a tautomer that is unable to accept a proton from Ser 195<sup>38</sup>. The free energy associated with the 180° inversion of His 57 in the active site of the Asn 102 mutant although significant (2 kcalmol<sup>-1</sup>) is not entropic in nature but simply the free energy required to move an incorrect orientated group to the correct orientation.

### 3.2.3 The 3D Crystal Structure Of The Serine Proteases

Comparative modelling works best when there are several experimental structures. In the serine protease family there are 16 trypsin like serine proteases which have their crystal structure solved. References to the solved three dimensional crystal structures for members of the serine protease family can be found in Table 3-3 on page 63 and Table 3-4 on page

64. With so many crystal structures solved, there is a large amount of real structural information available to help in the understanding of the serine protease enzyme mechanism and substrate specificity.

Protein	Source	Resolution (Å)	Reference
Trypsin	Human	2.2	39
	Bovine	1.5	40
	Porcine	1.8	41
Streptomycin Griseus Trypsin	S. Griseus	1.7	42
$\alpha$ Chymotrypsin	Bovine	1.68	43
$\gamma$ Chymotrypsin	Bovine	1.6	44
Human Leukocyte Elastase	Human	2.3	45
Elastase	Porcine	2.5	46
	Salmon	1.61	47
$\alpha$ - Lytic protease	Lysobacter	2.0	48
Protease A & B	S. Grseus	1.8	49, 50
Urokinase	Human	2.5	51
Kallikrein	Porcine	2.05	52
Tonin	Rat	1.8	53
$\gamma$ Thrombin	Human	2.5	54
Serine Protease II	S. Fridiae	1.6	55

**Table 3-3: Experimentally known three dimensional structures of trypsin like Serine Proteases.**

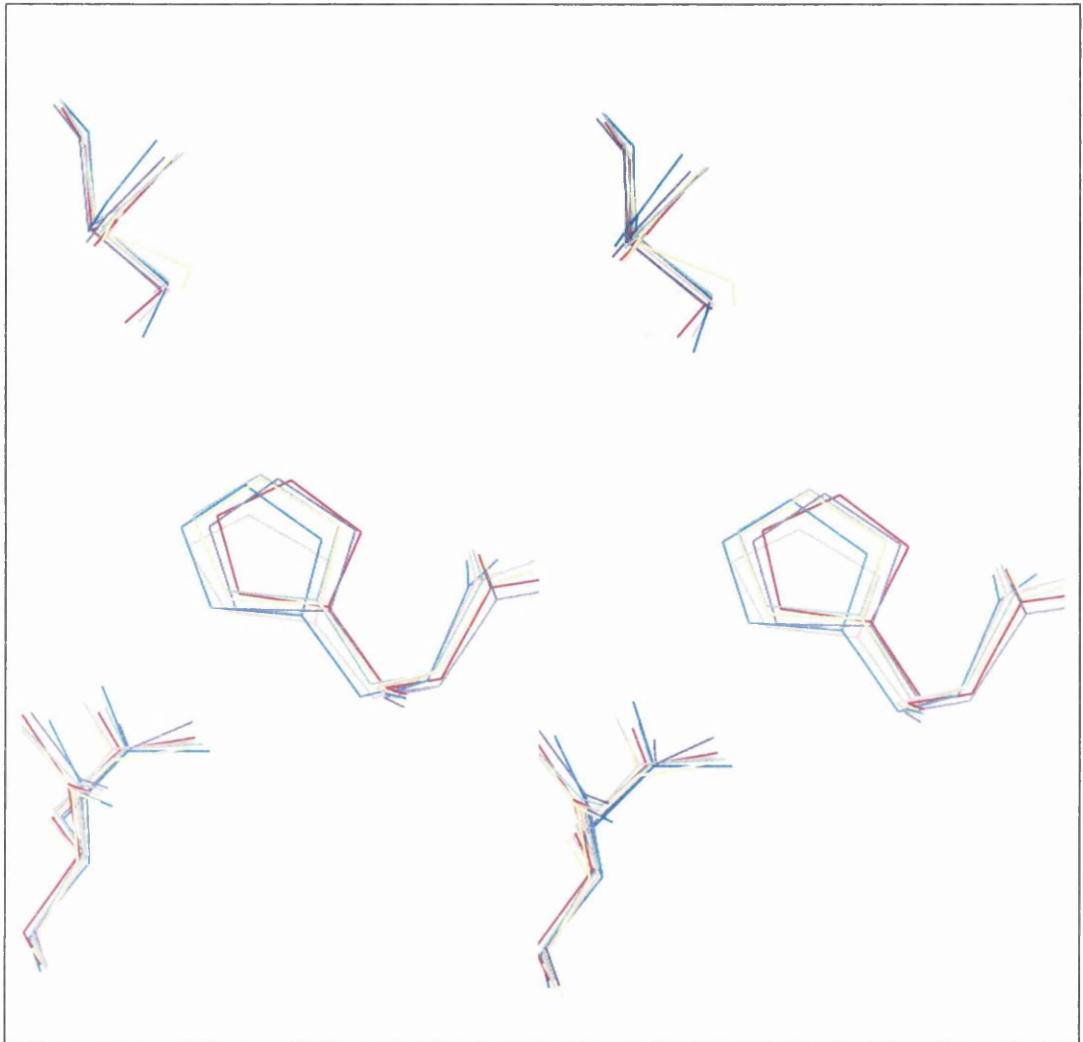
Protein	Source	Resolution (Å)	Reference
---------	--------	----------------	-----------

Protein	Source	Resolution (Å)	Reference
Subtilisin	Bacillus SP.	2.4	56
M - Protease	Tritirachium Album Lumber	2.4	57
Endopeptidase	Tritirachium Album Limbur	1.5	58
Thermitase	Thermoactinomyces Vulgaris	1.37	59

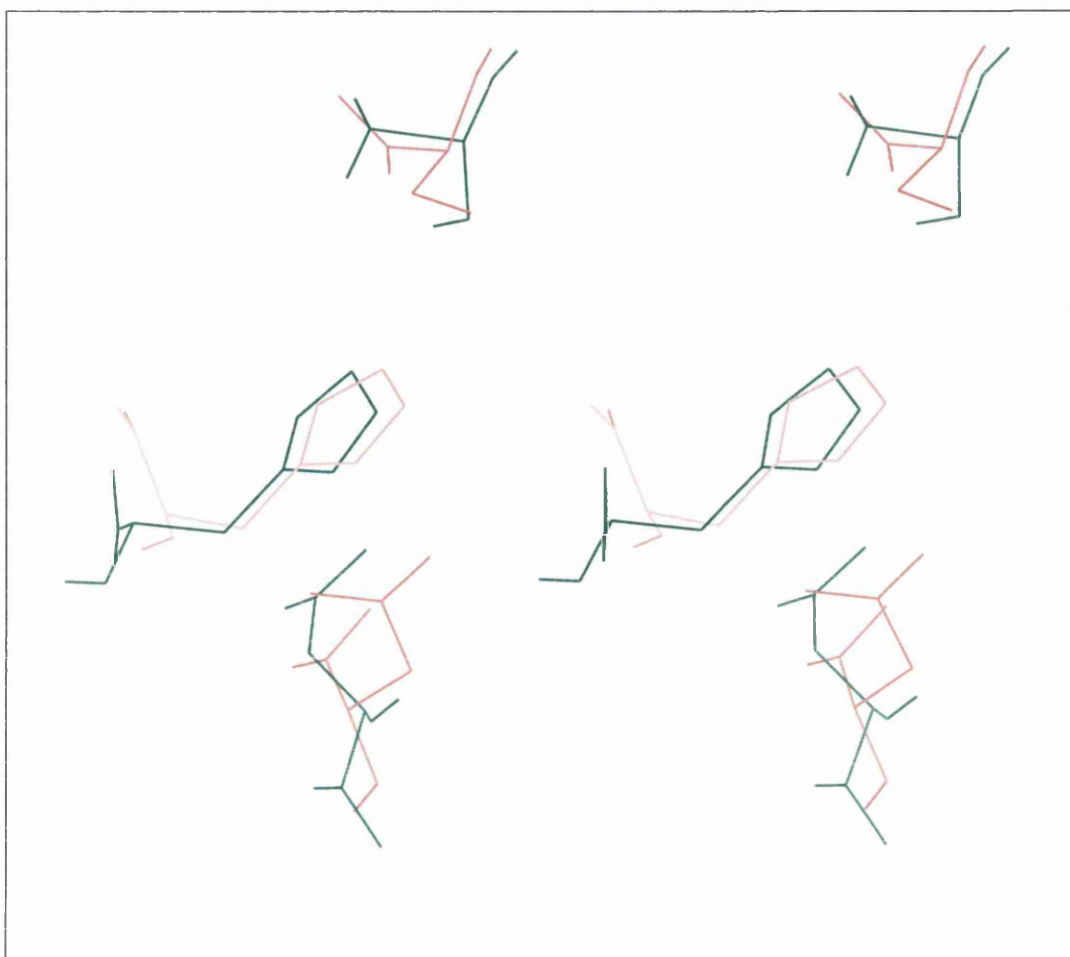
**Table 3-4: Experimentally known three dimensional structures of Subtilisin like Serine Proteases.**

### 3.2.3.1 The Catalytic Triad

The catalytic triad is common to all serine proteases (trypsin like and subtilisin like). Within the trypsin like serine proteases the sequence and structure around His 57, Asp 102 and Ser 195 are the most highly conserved regions. The catalytic triad is in the exact same spatial arrangement for each member of the trypsin like serine proteases, see Diagram 3-6 on page 65. The remarkable spatial similarity of the catalytic triad in human trypsin and Subtilisin (see Diagram 3-7 on page 66) show the importance of the spatial relationship between the residues involved in the catalytic mechanism. Any deviation from the optimal arrangement significantly affects the enzymes catalytic effectiveness.



**Diagram 3-6: Stereo view of the catalytic triad from several trypsin like serine proteases superimposed. 1trn (red), 1ept (green), 2sfa (blue), 1sgt (yellow), 2hnt (cyan), 1esa (magenta).**



**Diagram 3-7: Stereo view of the catalytic triad for human trypsin (1trn) (red) and subtilisin (1sbt) (green) superimposed.**

### 3.2.3.2 The substrate binding cleft

Nature has provided a rare opportunity for determining the structures of the enzyme - substrate complexes of trypsin and chymotrypsin with polypeptides. There are many naturally occurring polypeptide inhibitors that bind to trypsin and chymotrypsin very tightly because they are locked into the conformation that a normal flexible substrate takes upon binding<sup>60</sup>. They do not hydrolyse under normal physiological conditions because the amino acid that is released on the cleavage of the peptide is constrained and cannot diffuse away from the active site of the enzyme. On removing the constraints in the pancreatic trypsin inhibitor (by reducing an —S—S— bridge in its polypeptide chain) the peptide



### 3.2.3.2.2 The Primary Binding Site ( $S_1$ )

The binding pocket for the aromatic side chains of the specific substrates of chymotrypsin is a well defined slit in the enzyme 10 to 12Å deep and 3.5 to 4.0 by 5.5 to 6.5Å in cross section<sup>62</sup>. This gives a tight fit since an aromatic ring is ~6.0Å wide and 3.5Å thick. A methylene group is ~4.0Å in diameter, so the side chain of Lys or Arg is bound nicely by the same shaped slit in trypsin. More significant than the tight fit of the side chain into the pocket, in trypsin, there is a carboxylate group at the bottom of the pocket from Asp 189. This Asp at residue position 189 in trypsin forms a salt linkage with the positively charged ammonium or guanidinium groups in Lys and Arg respectively. In the elastase structure the two Gly at the mouth of the pocket in chymotrypsin and trypsin are replaced by bulky Val (Val 216) and Thr (Thr 226). This prevents the entry of large side chain into the pocket, and provides a way of binding the small side chain of Ala.

There are certain important hydrogen bonds found in all the enzymes. The  $O_{\text{carbonyl}}$  of the reactive bond has a binding site between the backbone NH group of Ser 195 and Gly 193. The hydrogen bonds made there are very important because this oxygen becomes negatively charged during the reaction. There is also a hydrogen bond between the NH part of the N-acetylamino group of the substrate and the  $O_{\text{carbonyl}}$  of Ser 214.

There is a vast literature of site directed mutagenesis studies carried out on the serine proteases<sup>63</sup>. These studies have generally concentrated on the origins of the specificity at the primary binding site  $S_1$ . Two such studies using trypsin variants D189K<sup>\*64</sup> and D189S<sup>65</sup> indicate that the presence of a negative charge at the base of the binding pocket is essential to high level catalysis by trypsin. Further studies reveal that it is not the number of direct contacts made with the charged residue but the accessibility of the negative charge to the Lys / Arg substrate which the binding affinity depends on in the trypsin  $S_1$  binding pocket. Both these studies indicate that the role of Asp 189 in trypsin is twofold, it provides both tight binding affinity as well as a high acylation rate<sup>66</sup>. Other studies point to the fact that for trypsin homologs, if the charge residue is any other residue that forms part of the  $S_1$  binding pocket other than position 189 then the charge is partially sequestered.

---

\* This notation indicates that in this particular Trypsin variant the expected aspartic acid residue found in wild type Trypsin at position 189 is mutated to lysine. The same notation is used for all other variants of the wild type enzymes.

This is the probable explanation why Asp 189 is so strongly conserved in the trypsin homologs.

Another important residue which plays a role in the high specificity of trypsin for Lys / Arg residues at the  $P_1$  position of the substrate is the backbone conformation of Gly 216. In all available crystal structures of trypsin and chymotrypsin, two hydrogen bonds are formed in an antiparallel  $\beta$  sheet fashion with the backbone amide group of Gly 216. The backbone conformation at Gly 216 differs between trypsin and chymotrypsin. Mutagenesis studies were carried out where trypsin was made to have chymotrypsin specificity by changing the residues at certain positions in and around the  $S_1$  binding pocket. These studies showed that the backbone conformation which Gly 216 adopts can affect the preference of the  $S_1$  site by a factor of  $10^4$ . The mechanism by which Gly 216 functions is likely to be through promoting accurate scissile bond positioning<sup>67</sup>. Because Asp 189 of trypsin also plays a crucial role in this function, it appears that the identity of the amino acid at position 189, and the backbone conformation at Gly 216, must be matched in order to permit efficient and specific catalysis by trypsin and chymotrypsin.

#### **3.2.3.2.3 Sites $S_1$ — $S_2$ — $S_3$**

The hydrogen bond between the N-acylamino NH and the  $O_{\text{carbonyl}}$  of Ser 214 initiates a short region of antiparallel  $\beta$  sheet between the residues Ser 214, Trp 215 and Gly 216 of the enzyme and the amino acids  $P_1$ ,  $P_2$  and  $P_3$  of the substrate.

#### **3.2.3.2.4 Site $S_1'$ - The Leaving Group Site**

There is a leaving group site that is constructed to fit L-amino acids<sup>68</sup>. The contacts with the enzyme are predominately hydrophobic, which accounts for the lack of exopeptidase activity with the enzyme, since this would require binding a  $-\text{CO}_2^-$  in a non polar region.



## 4 Molecular Modelling

The term molecular modelling covers an extensive area of topics ranging from simple ball and stick representations of small molecules on a computer screen to complicated molecular dynamic simulation studies of complex multi molecule systems. The common factor is that the molecule or molecules under investigation are considered as a set of coordinates to which the appropriate equations are applied and that the relevance of any results is dependent on how accurately the equations used simulate ( i.e. model ) the appropriate effects, in the real world, on the real system.

### 4.1 COMMET

'COMMET' is the name given to the molecular modelling package which was used throughout the project. It was developed in house over a number of years by Dr D.N.J. White and J.N. Ruddock. It runs on a transputer which usually resides within a Personal Computer (PC) or clone.

Although physical models are still used at times, the vast majority of molecular modelling carried out today is performed using computers. These modelling programs can be divided into two groups. The first of these groups are the single dedicated programs<sup>69</sup>. In these cases the program will usually start by loading a file containing the structural data of a molecule. The program will then proceed to perform a single function such as an energy minimisation or calculation of partial charges before terminating, writing any required results to the screen or file. An example of the above group of modelling packages is the Amber4 suite of routines<sup>70</sup>. This force field and suite of routines was developed to build models of proteins and then to carry out energy minimisation calculations or molecular dynamic simulations of the protein models. This is accomplished by an individual programme to carry out the different steps of the modelling. There are three separate programmes which must be run to simply build the protein model from the amino acid sequence. Then the main programme can be run to carry out either energy minimisation calculation or molecular mechanic simulations on the model built

The other approach is to have a package that integrates many different options<sup>75</sup>. This means that it is possible to stay within the single package while performing a wide variety of operations. With the increase in computer power and memory size the integrated packages are becoming more useful and therefore more used. An example of an integrated package is the 'COMMET' system ( COncurrent Molecular Modelling Environment on Transputers) which has been developed within the laboratory. This package was originally developed from the VAX program COGS<sup>76</sup>. COMMET is a graphics based, menu driven package. It is useful to examine the various functions within such a package to obtain an overview of the options covered by the general heading of 'Molecular Modelling'.

### 4.1.1 List Of Functions Available Within COMMET

#### 4.1.1.1 Files

This menu contains all the file operations available from within the system. These include reading, writing, inspecting, and erasing of files, changing the current directory and the ability to choose if a new structure being loaded will replace the currently resident model or be loaded alongside the current structure.

#### 4.1.1.2 Build

This menu contains all the options available for constructing molecules. Building can occur in several ways. Either by the addition or deletion of individual atoms, or by adding amino acid residues to build up a polypeptide chain. There is also a 'sketch molecule' option where the user can sketch the required molecule on the computer screen with the mouse and indicate which atoms are above or below the plane. This option will then invoke an energy minimiser from which the final structure can be obtained. The 'protein build' option allows the user to edit the amino acid sequence and the backbone structure of a polypeptide chain. For smaller molecules the option 'assemble fragment' allows the user to join two molecules by the elimination of two hydrogen atoms.

#### 4.1.1.3 Edit

The edit menu is used for the removal of all or part of a molecule, for joining or breaking bonds, and for selecting a section of a polypeptide chain to be viewed.

#### 4.1.1.4 Change

There are two parts of this menu, the first involves changing certain properties relating to the molecule itself i.e. such things as bond length, bond angle, atom types and atom charges. The other menu items are for making changes to the default programme parameters i.e. the colour of the backbone, whether the depth cue option is on or off, and the hydrogen visibility.

#### 4.1.1.5 Display

This menu allows the user to change the way that the molecule is displayed on the screen. Initially the model is displayed as a simple wire framed model. The following options are available:

'Simple Space Filled'	displays the atoms of a molecule as simple circles
'CPK Surfaces'	draws the molecule as intersecting spheres of van der Waals radii
'Dot Surface'	as CPK surfaces but the surface of the spheres are represented by dots rather than a solid surface
'Stereo'	red and green images are displayed, where one of the molecules is rotated several degrees (nomally 6°) around the screen's y axis to produce a stereo image when viewed with the appropriate coloured glasses
'Alpha carbon backbone'	single line joining each C <sub>α</sub> with its neighbouring C <sub>α</sub> in a protein or polypeptide
'Beta Spline'	as for alpha carbon backbone but using a beta spline smoothing technique to produce a smooth curve
'Ribbon'	as for beta spline but with several splines side by side to produce a ribbon effect

**'Ball and Stick'**

draws the atoms as small balls and the bonds between atoms represented by cylinders

It is possible in this menu to produce screen displays that are suitable for being photographed and to highlight selected areas of the molecule.

#### **4.1.1.6 Show**

In the show menu the user can highlight various selected groups, display several different atomic properties, and determine which atoms are likely to be sterically crowded or involved in hydrogen bonding.

#### **4.1.1.7 Compare**

This menu allows the superimposition of molecules and then, if required to pulse between them to help show any differences and similarities.

#### **4.1.1.8 Calculate**

The menu contains the options for calculating the 'steric congest'<sup>77</sup> at any atom or pair of atoms where it is possible for a stereoscopic reaction to take place. 'Delre Charge'<sup>71</sup> will calculate the charge on a molecule using the Delre method. The menu also contains options to calculate both the surface area and volume of any selected molecule.

#### **4.1.1.9 Search**

This menu contains the routines that attempt to find the global minimum energy position of molecules using different methods. SITAR<sup>76</sup> is used to find the global minimum of amino acid side chains in polypeptides. The 'Global Minimum'<sup>76,77</sup> option is used for cyclic polypeptides. The 'Monte Carlo'<sup>72</sup> routine will attempt to use the Monte Carlo technique and can be applied to any molecular situation. The 'Z axis Permutation' is used for small molecules that contain one or more rings.

#### **4.1.1.10 Energy**

This menu contains most of the other options that are required to calculate the energy of the system. The options are as follows:

‘Torsion Angle Profile’ and ‘Ramachandran Map’<sup>73</sup> calculate the energy profile plots for rotation about one and two torsion angles respectively.

‘Pattern Search’<sup>74</sup> and ‘Newton Raphson’<sup>74</sup> minimisations are energy minimisation routines that use specified techniques for minimisation.

‘Auto Docker’ attempts to dock a molecule into a specific site on another molecule.

‘Molecular Dynamics’<sup>74</sup> is the simulation of the movement of the atoms in the molecule above absolute zero so that the atoms will have vibrational motion.

‘Chain Annealer’ is used when an alteration has been made to a polypeptide chain and involves minimisation along only the section the backbone chain where the user specifies.(i.e. where the alteration took place)

#### **4.1.1.11 Macromolecules**

This menu contains the options that deal with the analysis and properties of the primary sequence of a protein. The options allow you to align sequences, make secondary structure predictions, and calculate the hydrophobic character of the chain along its length.

Also within this menu are the options to generate structures for loops within a polypeptide structure. A search of the Brookhaven database is possible to find a fragment of another protein with a similar sequence to the loop being built, or build the structure of the loop by examining all possible conformations the loop can adopt.

#### **4.1.1.12 Geometry**

This menu allows the user to measure simple geometric properties such as bond length, valence angle and torsion angle.

#### **4.1.1.13 Transformations**

The options in this menu allow the user to manipulate the position of any model by translation or rotation, or to scale the size of the model on the computer screen.

There is also a rotation bar at the bottom of the screen which enables rotation about the x, y or z axis. This option is available in any menu or submenu and is very useful when selecting individual atoms from large molecules.

#### **4.1.1.14 Summary**

Within such a modelling package as 'COMMET' there is a wide range of operations that can be applied to a range of different molecular situations. It is advantageous at this stage to try and make a degree of distinction between different sorts of operations.

The operations can be split into actual calculations and graphical operations. The graphical presentation of the molecule is an important part of the molecular modelling package. Looking at different views and different representations of the model can give a greater increase in understanding of the model. The data produced by calculations can more easily be interpreted if they are displayed in a graphical context. All this can help the scientist in deciphering what is happening to the model under certain conditions.

The other main area is the actual calculations. These can be subdivided depending on the degree of parameterisation which is required for the function. For the majority of the calculations there is either no requirement for extra parameters or the number of extra parameters required will be small and relatively quick to work out. These calculations include the calculation of the distance between two atoms, Delre charge etc. The other extreme is where the number of parameters required for a calculation becomes a major factor in the accuracy and speed of the calculation. This is the case for molecular

mechanics calculations. In the case of molecular mechanics the number of parameters required determines the range of atom types that the molecular mechanics routine can cope with and the accuracy of the final answer is dependent on the quality and number of the parameters in the force field.

## 4.2 Molecular Mechanics

### 4.2.1 Introduction

Molecular mechanics can be considered as a technique for calculating many properties of molecular systems based on the ability to determine an estimation of the energy for such systems in any atomic configuration. When calculating the energy it is assumed that all the interactions within a molecule can be treated in an empirical manner.

Even though molecular mechanics is empirical in nature it can be justified, to some extent, from quantum mechanics. This is done by examining the Born - Oppenheimer approximation. This states that in quantum mechanics it is possible to separate the motion of the nuclei in a molecule from the motion of their associated electrons, with very little effect on the calculated results. This is used to find the electronic structure of a molecule by considering the nuclei to be fixed in a given configuration. It is equally valid though, to investigate the motion of the nuclei. Here it can be considered that the surrounding electron density leads to the various interactions represented in molecular mechanics by the potential functions.

The other main supposition in molecular modelling is that the interactions between atoms within a molecule can be divided into various distinct types. The terms used for each of these interaction types and the empirical parameters that these terms require are what is known collectively as the force field.

The total structural energy of a molecule is a simple sum of the energy calculated for each of the terms. This is known as the steric or strain energy. Thus the total steric energy  $E_s$  of a molecule can be given by:

$$E_s = E_b + E_\theta + E_\omega + E_{nb} + E_c + E_{oopb} (+E_x)$$

Where  $E_b$  is the energy relating to bond stretching or compression,  $E_\theta$  is the energy for valence angle distortions,  $E_\omega$  is the component relating to the energy from torsional barriers,  $E_{nb}$  is the non bonded contribution from the van der Waals potential,  $E_c$  is the coulombic energy arising from charge interactions, and  $E_{oopb}$  is a summation term over all of the nominally trigonal planar atoms to account for the increase in potential energy due to pyramidisation.  $E_x$  represents the possibility that other terms may be required such as cross terms that could be added to the force field to increase accuracy.

For each term in the force field there are usually several possible equations that could be used and in each of these cases there are usually several different ways that the required parameters can be selected.

For the majority of these terms the parameters consist of a 'natural' or 'strain free' value and one or more force constants that determine how difficult it is to deform the property from this 'strain free' value. The interaction types not using this approach are the torsional angle twist where periodicities and barrier heights are used and the non bonded interactions which use separate terms to cover attraction and repulsion plus a single term for coulombic interactions.

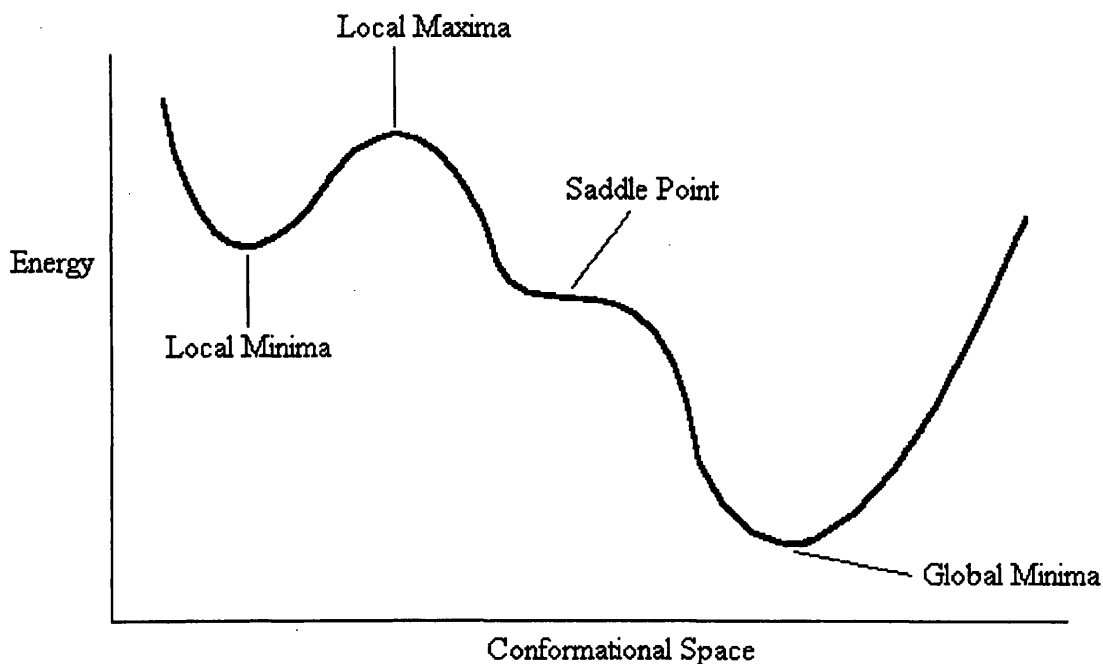
The basic idea of using an empirical force fields had been proposed as early as the 1930's<sup>69</sup> but serious attempts to use molecular mechanics were not made until 1946<sup>75,76,77</sup> Due to the lack of reliable information on which to base the parameters and the difficulties involved in carrying out large calculations at that time, the more wide spread use of molecular mechanics did not start until the 1950's<sup>71</sup>. Since then its importance has increased steadily with the growth in computer power and with the increase in the number of computational approaches that have been devised and implemented.

For a molecule with  $N$  atoms it is possible to imagine a  $3N$  dimensional surface that describes the energy of the molecule in its electronic ground state as it is affected by the values of the  $3N$  co-ordinates ( $x, y, z$  for each of the  $N$  atoms). This surface is usually called the Born - Oppenheimer surface or the potential energy hyper surface. Depending on



the symmetry of the molecule under investigation there will be one or more locations on this surface that will correspond to the lowest energy position of the molecule. Thus, as well as a method to determine the energy of the molecule in any conformation there is also a need to manipulate the molecular co-ordinates to allow for a search of the potential energy hyper surface so that the position or positions of minimum energy can be found.

This is done using a minimisation technique that will usually examine the forces on the atoms and then apply transformations to each of the co-ordinates in an attempt to obtain a configuration with smaller average forces. This process is repeated iteratively until a stable conformation is reached. One of the main drawbacks of these optimisation methods is that the minimum is found by heading towards the nearest local energy minima and this is generally not the global minimum energy for the given structure. As well as this, some minimisation techniques can terminate at an energy maxima or get caught at a saddle point in the potential energy surface, see Diagram 4-1 on page 79.



**Diagram 4-1: Likely elements on a Potential Energy Surface**

Some of the various methods of energy minimisation are described later. In most cases the Newton - Raphson technique is used as it provides a rapid method of reaching the local minima. It is not, though, the optimum technique when the initial structure is highly distorted from the minimum energy position. Because of this some molecular mechanics programmes initially use a different method such as steepest descent to get closer to the minimum before using the Newton - Raphson method.

One of the problems with the Newton - Raphson technique arises with large deformations and in some cases it can result in a situation of increasing oscillation for the affected atomic positions. To counteract this a variation of the line search technique has been implemented which recognises the start of such an oscillation and prevents it by reducing

the maximum atomic shift until a reduction rather than an increase in the energy is obtained.

#### 4.2.2 Why use Molecular Mechanics

With all the problems associated with the production and use of a molecular mechanics force field there is a need to consider why molecular mechanics should be attempted. Results, after all, can be obtained from various experimental techniques and from *ab initio* quantum mechanic calculations which do not require any empirical parameters.

The main reason that molecular mechanics has become so popular is the speed and convenience with which it is possible to produce the required results. This is especially true because of the large range of data that can be obtained from molecular mechanics calculations. For example, if an investigation of a new molecule is required then to get the structure alone experimentally would require that the molecule under investigation be synthesised, followed by the need to determine the molecular conformation using one of several methods such as x ray diffraction.

Thus each result obtained from molecular mechanics calculations, such as an estimate for the heat of formation, would, if determined experimentally, require its own experimental procedure. So in many cases such calculations can save a considerable amount of time, effort and money. This is, however, not to say that such experimentation should cease. The whole basis of molecular mechanics is that it is an empirical technique and so will never totally replace the more accurate experimental approaches. Indeed it is the expanding database of such experimental data that allows for the continually improving quality of molecular mechanics force fields.

#### 4.2.3 Quantum Mechanics

What then of quantum mechanical calculations. They too can be used to find the minimum energy of a molecule by computational methods alone. The principal problem here is that when carrying out a quantum mechanics calculation there is always a trade off in accuracy against calculation time, depending on which basis of atomic wave functions is chosen.

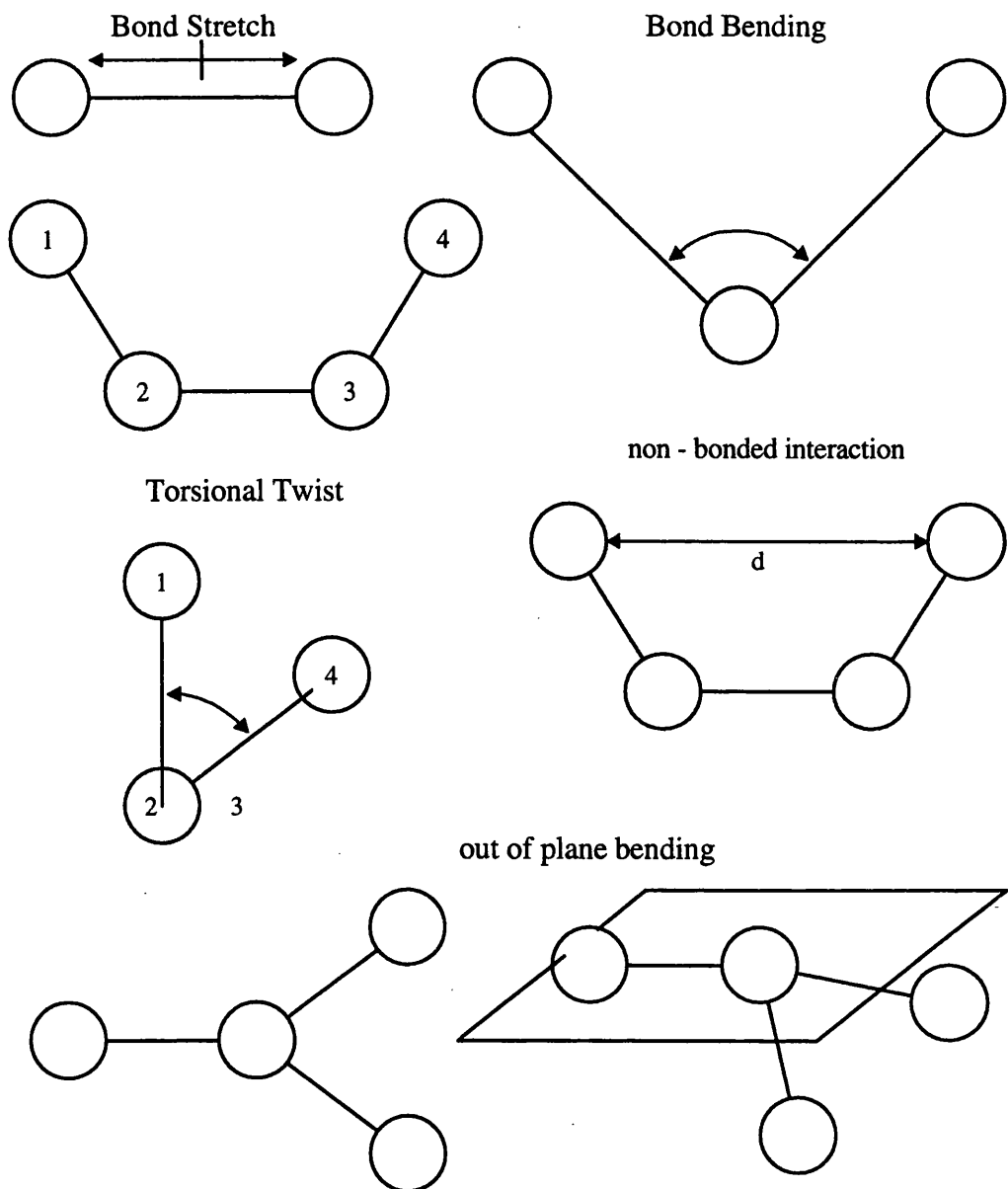
With the simpler wave functions the calculations are completed more rapidly, if still orders of magnitudes slower than the molecular mechanics calculation for the same molecule. Even a simple basis set is in itself an approximation and so the results themselves are also approximate<sup>72</sup>. The use of a more accurate basis set will improve the accuracy of the results but this will also substantially increase the computational time required for the calculation. Another advantage with molecular mechanics is that the computation time increases with roughly the square of the number of atoms involved, whereas, with quantum mechanics, it increases with the fourth power of the number of atoms. This means that, although molecular mechanics calculations are regularly performed on systems of hundreds of thousands of atoms, such systems are still out of the applicable range of *ab initio* quantum mechanics calculations, even on the fastest supercomputers.

There are also several popular quantum mechanical techniques that neglect specific orbital overlaps to speed up the calculations but these bring with them the need to use some empirical parameters in an attempt to make up for the resulting deficiencies, hence the term semi - empirical which is given to these methods. Even with the resultant increase in speed none of these methods come close to the accuracy of molecular mechanics within the same computational time scale.

Quantum mechanics calculations are still very useful, especially when investigating situations that are not reliably parameterised to molecular mechanics or when studying reactions, which are difficult or impossible to simulate using molecular mechanics.

#### 4.2.4 Force Fields

As previously stated the force field consists of both the form of the potential functions used to calculate each energy component and the related parameters. Diagram 4-2 on page 82 shows the main interaction types for which potential energy functions will be required.



**Diagram 4-2: Interaction Types Considered In A Force Field**

#### 4.2.4.1 Bond Stretching

The Morse curve<sup>73</sup> describes how the molecular potential energy of a diatomic molecule varies as a function of the internuclear separation. The function has a minimum at the distance corresponding to the equilibrium bond length ( $l_0$ ) of the molecule. The region of

the curve close to the minimum fits a parabola ( general equation  $y = x^2 / 4a$ ) so that for an individual bond we can write:

$$v_1 = \frac{1}{2}k_1(l-l_0)^2$$

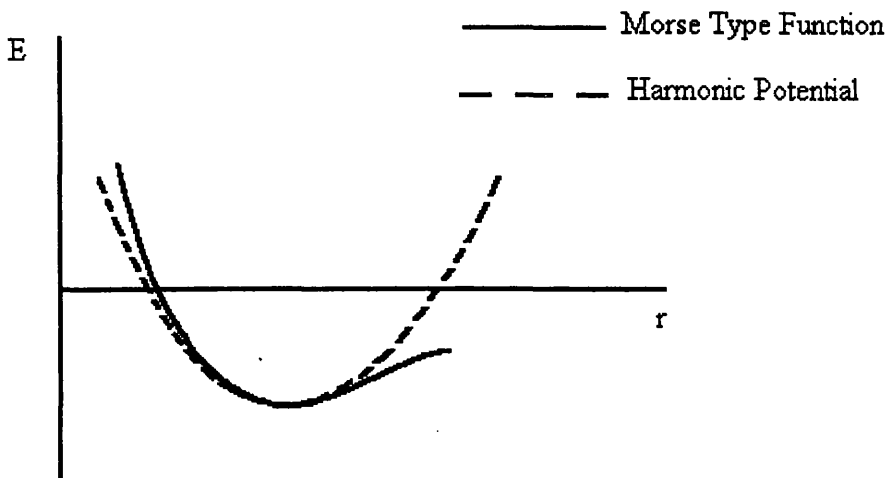
where  $v_1$  is the potential energy.  $k_1$  is the force constant and  $l$  is the current bond length. Therefore  $l-l_0$  is the deviation of the actual bond length from the equilibrium value.

Alternatively the diatomic molecule can be considered as two masses joined together by a Hookean spring where the force between the two masses is proportional to the extension.

The restoring force is given by  $-\frac{dv}{dl}$ , the extension is given by  $(l-l_0)$ , and the force constant is  $k_1$ . Integrating  $-\frac{dv}{dl} = -k_1(l-l_0)$  with the boundary condition that  $v_1 = 0$  when  $l = l_0$  gives:

$$v_1 = \frac{1}{2}k_1(l-l_0)^2$$

The value of the force constant is very large for all combinations of bonded atoms and this ensures that  $l$  is never very different from  $l_0$  so that the Hookean, harmonic, or parabolic approximations hold for all chemically sensible values of  $l$ , see Diagram 4-3 on page 84.



**Diagram 4-3: Morse and Harmonic Curves For Bond Length**

The expression for  $E_1$  given above is applied to all pairs of bonded atoms in a complex molecule, with a different value of  $k_1$  for each unique combination of bonded atom types, and the individual values summed to give  $E_b$ :

$$E_b = \sum_1 \frac{1}{2} k_1 (l - l_0)^2$$

#### 4.2.4.2 Valence Angle Bending

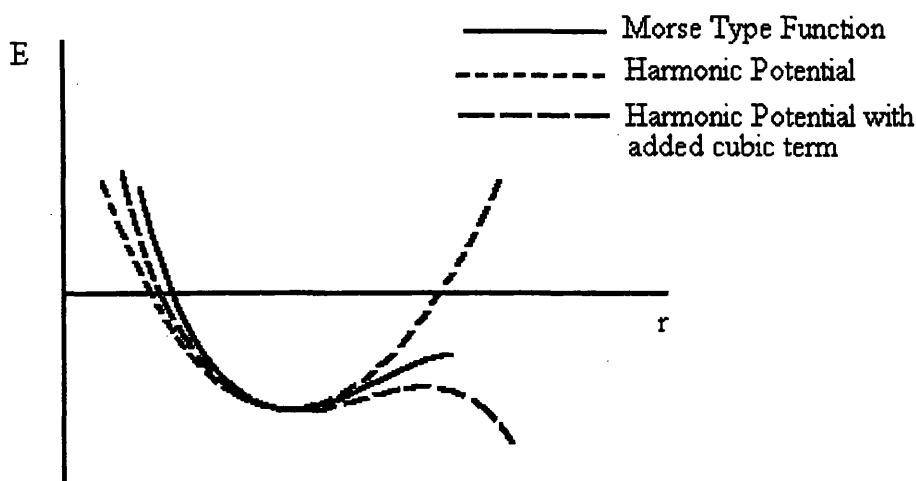
A similar expression to that for  $E_b$  can be derived for valence angle bending:

$$v_\theta = \frac{1}{2} k_\theta (\theta - \theta_0)^2$$

The force constants  $k_\theta$  are sufficiently small that in strained molecules the angle  $\theta$  deviates from  $\theta_0$  to such an extent that the harmonic approximation is no longer valid, see Diagram 4-4 on page 85. The situation can be improved by adding a single anharmonic correction term<sup>78</sup> to give the following equation:

$$E_\theta = \sum_\theta \frac{1}{2} [k_\theta (\theta - \theta_0)^2 - k_\theta' (\theta - \theta_0)^3]$$

where  $k_{\theta}'$  is a supplementary constant for anharmonicity correction. However, although there is a different value for the force constant  $k_{\theta}$  for each unique grouping of three bonded atom types, the values of  $k_{\theta}'$  is the same for all atom type triplets.



**Diagram 4-4: Morse, Harmonic and Harmonic with Cubic Curves**

This works well except in the few cases where the angle starts off being greatly deformed from the strain free value. A situation where this could easily occur is during the energy minimisation of a molecule that has been sketched into a molecular modelling package. In these cases the cubic correction term can become dominant over the squared term. Because of this the total energy will decrease if the angle is further deformed away from the strain free value, an obviously unrealistic situation. To prevent this from occurring an extra term can be added which will, in these extreme cases have a greater effect than the cubic term and so force the angle back towards a more reasonable value. The fifth power is often chosen because it does not drastically increase the calculation required as it can be produced from the product of the square and cubic terms:

$$E_{\theta} = \sum_{\theta} \frac{1}{2} [k_{\theta}(\theta - \theta_0)^2 - k_{\theta}'(\theta - \theta_0)^3 + k_{\theta}''(\theta - \theta_0)^5]$$

#### 4.2.4.3 Torsional Strain



The next term to be considered is that relating to the torsional energy,  $E_\omega$ . It had long been thought that the need for a torsional energy term in a force field was only due to the deficiencies in the other parameters, most specifically the non bonded interactions, and that if these could be optimised properly then this term would not be required. There have been, however, several attempts to devise force fields that do not contain a torsional term but these have failed even in the simplest of cases. They could not, for example, even reproduce the correct internal rotational barrier height for ethane<sup>79</sup> without giving incorrect answers for other properties such as angles.

The variation of potential energy with bond torsion for a simple molecule such as ethane is given by the expression:

$$v_\omega = \frac{1}{2} k_\omega (1 - \cos 3\omega)$$

where  $\omega$  is any one of the H—C—C—H torsion angles about the central C—C bond (it does not matter which torsion angle is chosen as they are all equivalent by symmetry) and  $k_\omega$  is the barrier to free rotation in ethane ( $\sim 3 \text{ kcal mol}^{-1}$ ). This equation can be generalised to:

$$v_\omega = \frac{1}{2} k_\omega [1 + s \cos(n\omega)]$$

where  $s$ , the phase of the barrier, is either +1 for a torsion angle where the minimum energy conformation corresponds to a staggered arrangement of bonds (e.g. ethane), or -1 where the minimum energy conformation is an eclipse arrangement of bonds (e.g. ethene). The periodicity of the barrier,  $n$ , depends on the nature of the two central atoms comprising the torsional angle;  $n=3$  for the threefold periodic barrier in ethane, and  $n=2$  for the twofold periodic barrier in ethene.

It cannot be assumed in general that the torsion angles around a bond will be symmetrically equivalent, so it is necessary to calculate a contribution to  $v_\omega$  from each individual torsion angle around the central bond (i.e. there are nine individual torsion angles around any  $\text{C}_{\text{sp}^3} - \text{C}_{\text{sp}^3}$  bond). For any one torsion angle therefore:

$$v_{\omega} = \frac{1}{2} \frac{k_{\omega}}{N_{\omega}} [1 + s \cos(n\omega)]$$

where  $N_{\omega}$  is the total number of unique torsion angles around the bond concerned. For example, there are four torsion angles around any  $C_{sp}^2 = C_{sp}^2$  double bond so that the previous equation would have to be evaluated four times, once for each torsion angle, in order to get the molecular potential energy due to the  $C_{sp}^2 = C_{sp}^2$  bond torsion.

Even this picture is a little simplistic for practical purposes; consider the torsion around the central C—C bond in n-butane. There are three types of torsion angles :H—C—C—H, C—C—C—H and C—C—C—C. While the periodicities of the first two are essentially threefold the major component of C—C—C—C is onefold (any C—C—C—C starting arrangement is not repeated during a 360° torsion rotation and  $v_{\omega}$  is much larger for 2,3-dimethylbutane than for ethane), with a minor threefold addition. In such cases it is necessary to use a short Fourier series for calculating  $v_{\omega}$ . In general therefore  $v_{\omega}$  can be written as:

$$E_{\omega} = \sum_{\omega} \frac{1}{2} \left\{ \frac{k_{\omega}}{N_{\omega}} [1 + s \cos(n\omega)] + k_{\omega}' [1 + s \cos(\omega)] \right\}$$

In most cases  $k_{\omega}'$  is zero but exceptions include  $C_{sp}^3 - C_{sp}^3 - C_{sp}^3 - C_{sp}^3$  and  $C_{sp}^3 - C_{sp}^2 - N_{amide} - C_{sp}^3$  (peptide bond).

#### 4.2.4.4 Non Bonded Interactions

There are two types of non bonded interactions, the Coulombic interactions between the charges on the atoms, and the van der Waals interactions between the atoms themselves.

##### 4.2.4.4.1 Coloumbic Interactions

A real molecule consists of the positively charged atomic nuclei surrounded by the negatively charged electrons in their appropriate orbitals. This results in a charge distribution that extends throughout the volume of the molecule. The calculation of the

total coulombic energy is very computationally intensive and as such is not particularly feasible for the time scale of molecular mechanics calculations.

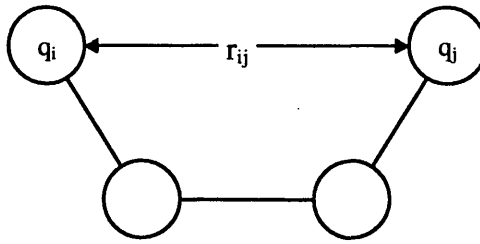
Instead force fields containing a charge term use either a system where the charge is distributed as point charges on the atom nuclei or one that assign dipoles along each bond. With the dipole approach the usual method is for each pair of bonded atoms to be given a dipole depending on the atoms in the bond. This means that their values are easy to assign but require relevant values for the dipole to be known. The dipole values also do not take into account the effect that any other atoms bonded to either of the relevant atoms might have.

The calculation of point charges will usually take these effects into account but the calculation of a charge distribution is itself not a simple problem and can require a substantial amount of computation. This point charge distribution, though, needs only be calculated once as the charges are usually calculated just from the connectivities and the atom types and thus should be equally valid whatever the atomic positions.

In the point charge case the energy can be obtained from the sum of the pairwise interactions between all the possible combinations of monopoles i.e.

$$E_c = 332 \sum_{ij} \frac{q_i q_j}{D r_{ij}}$$

where  $q_i$  and  $q_j$  are the charges (in units of electrostatic charge), on the atoms  $i$  and  $j$  separated by the distance  $r_{ij}$  and  $D$  is the dielectric constant. The scaling factor of 332 converts the units of the energy to  $\text{kcalmol}^{-1}$ . Diagram 4-5 on page 89 shows a single pairwise monopole interaction.

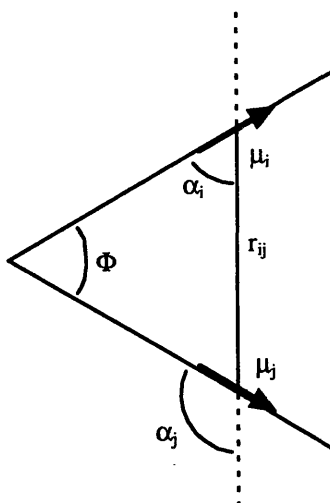


**Diagram 4-5: Single Pairwise Interaction**

In the dipole case the total energy for the charge interaction is obtained from the sum of the interactions of the pairs of dipoles<sup>80</sup> i.e.

$$E_c = \sum_r \frac{\mu_i \mu_j}{D r_{ij}} [\cos \Phi - 3 \cos \alpha_i \cos \alpha_j]$$

where  $D$  is the dielectric constant,  $r_{ij}$  is the separation of the two dipoles,  $\Phi$  is the angle between the dipoles,  $\mu_i$  and  $\mu_j$  are the values of the dipole, and  $\alpha_i$  and  $\alpha_j$  are the angles each dipole makes to a line connecting them. Diagram 4-6 on page 90 shows a single dipole interaction



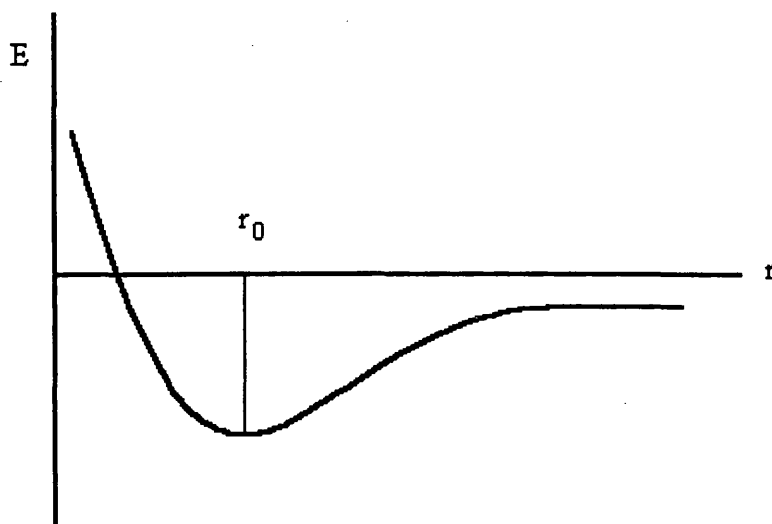
**Diagram 4-6: Single Dipole Interaction**

The 'intramolecular' dielectric constant is not the same as the dielectric constant of the bulk material but is at present impossible to measure. Consequently  $D$  is often set to a constant value between 1<sup>81</sup> and 5, an action which has undesirable computational consequences. This is because of the term  $r_{ij}$  in the denominator of the above expressions. Computing  $r_{ij}$  means taking the square root of  $r_{ij}^2$ , which would normally be computed first - a relatively time consuming process best avoided if possible.

A solution to the problem outlined above is to use a 'distance dependent' dielectric where  $D$  is set equal to  $r_{ij}$  so that the denominator becomes  $r_{ij}^2$ , removing the requirement to calculate a square root. There is also a physical justification for this procedure. Allowing molecules to diffuse into an evacuated volume leads to an increase in the dielectric constant of the fluid, and up to a point the more molecules present the higher the dielectric constant. If two point atomic charges in a molecule are very close together there is little possibility of other atoms or molecules interposing themselves between the charges and the dielectric constant of the intervening space will be low. On the other hand if the charges are widely separated the chances of matter appearing between them is high and this would lead to a higher dielectric constant. Therefore the distance dependent dielectric is not only computationally efficient but physically justified.

#### 4.2.4.4.2 Van Der Waals Interactions

Whereas with all the other terms the choice of potential function is quite often just between the level of complexity of a particular form, the van der Waals interaction is different in that there are a number of alternative forms for the equations. All of these equations attempt to reproduce the typical shape of the van der Waals interaction as seen in Diagram 4-7 on page 91.



**Diagram 4-7: Typical Van Der Waals Interaction where  $r_0$  is the equilibrium distance.**

Lennard Jones Potential<sup>82</sup>

$$E_{nb} = \sum_{ij} [A_{ij}r_{ij}^{-12} - B_{ij}r_{ij}^{-6}]$$

Buckingham Potential<sup>83</sup>

$$E_{nb} = \sum_{ij} [A_{ij} \exp(-B_{ij}r_{ij}) - C_{ij}r_{ij}^{-6}]$$

In both equations A, B and C are constants peculiar to each unique pair of atom types (e.g.  $C_{sp^3-N_{amide}}$ ,  $C_{sp^2-O_{sp^3}}$  etc.), and  $r_{ij}$  is the internuclear separation of atoms i and j. The non

bonded potential energy is the summation of the individual energy over all unique pairs of non-bonded atoms in the molecule.

Although the Buckingham potential is used in some molecular mechanics programmes it has two major drawbacks. It is assumed in a molecular mechanics calculation that non-bonded atoms  $i$  and  $j$  will always remain non-bonded, so that  $v_{ij}$  will be a minimum at the equilibrium separation  $r_0$  of  $i$  and  $j$  and will increase ever more steeply as  $r_{ij}/r_0$  becomes smaller than 1. The Buckingham potential turns over when  $r_{ij} \cong 1.5\text{\AA}$ , and the energy starts to decrease. While this may be physically realistic in some instances (i.e. a bond has been formed), it is not what is required for molecular mechanics calculations. The behaviour of the Buckingham potential is particularly troublesome, unless precautions are taken, when molecular mechanics calculations are used to “three dimensionalise” atomic co-ordinates taken from a two dimensional model of a molecule.

The second disadvantage of the Buckingham potential is a purely computational one. The ‘exponential’ function is computationally very expensive to calculate, whereas the  $r^{-12}$  term in the Lennard Jones potential can be evaluated by squaring the already calculated  $r^{-6}$ . This is an important consideration when dealing with large molecules.

The use of a simple pairwise potential has been criticised<sup>84</sup> because it neglects the many body effects and that the van der Waals interactions would be affected by the electron density of any other atoms that happened to be between the two atoms in question, in a similar way to the coulombic interaction. Despite this it is still used as these effects appear to be insignificant compared to the advantages obtained in calculation time.

A greater problem with the van der Waals interaction is that in both of the above equations the atoms are assumed to be spherical. This generalisation results in two main problems.

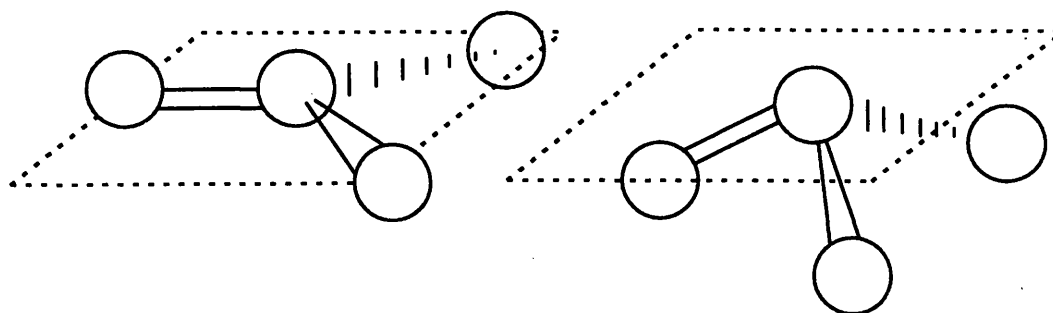
The first is that some types of atoms, such as oxygen, have lone pairs of electrons with the result that, in the real molecule, the repulsion between such a lone pair bearing atom type and another atom will vary depending on the location of the lone pair electrons with respect to the atom centres. That is, the repulsion will be greater when a lone pair is between the atoms’ centres than when they are not. It is possible to simulate this situation by adding

lone pairs of electrons as pseudo atoms but this will necessarily result in an increase in the number of parameters required in the force field.

A second problem is that bonded hydrogen is found to have an electron density centred not at the hydrogen nucleus but instead is shifted along the bond towards the other atom in the bond. Some force fields attempt to produce this effect by moving the centre of the hydrogen atom about 10% along the bond when calculating the non-bonded interactions<sup>8586</sup>. This has been found in some cases to reproduce the crystal packing of hydrocarbons better than when the centres are taken at their normal positions. This approach will increase the computational time as the “new” atomic centre for each hydrogen will need to be recalculated at every point required.

#### 4.2.4.5 Out of Plane Bending

Nominally trigonal planar atoms, such as  $C_{sp^2}$  or  $N_{amide}$ , can deform under strain in such a way as to convert the trigonal planar arrangement into a trigonal pyramid, with the trigonal atom at the apex of the pyramid and the three substituents in a plane either above or below the trigonal atom, Diagram 4-8 on page 93.



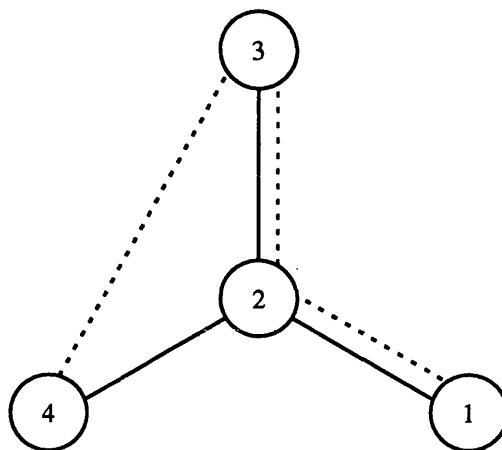
**Diagram 4-8: Out Of Plane Bending**

Pyramidalisation obviously leads to an increase in the potential energy and this is accounted for as follows:

$$E_{oopb} = \sum_{\chi} \frac{1}{2} k_{\chi} (\chi - 180)^2$$



where  $k_\chi$  is the force constant for out of plane bending and  $\chi$  is the “improper” torsion angle. If the substituents around the trigonal atom are labelled 1, 3, 4 and the trigonal atom labelled as 2,  $\chi$  is defined by 1—2—3...4 (Diagram 4-9 on page 94). Obviously for a minimum energy trigonal planar arrangement the 1—2—3...4 torsion angle is  $180^\circ$ , while for a trigonal pyramid it deviates from  $180^\circ$  by an amount related to the height of the trigonal pyramid.



**Diagram 4-9: Improper Torsion Angle**

#### 4.2.4.6 Cross Terms

There are several cross terms that can be used to simulate the interaction between different terms in the force field. The main ones used are the stretch - bend, torsion - stretch and bend - bend interactions. These terms can be included, if required, to create a force field that is as complete as possible and are especially useful when the force field is being used to try and reproduce information such as vibrational frequencies. Ideally it would be advantageous to have all these terms in a force field. However, in the current situation where a force field is required for a large number of atom types without the need to reproduce infra red vibrational frequencies then it becomes necessary to examine these terms and decide if the increase in accuracy is justified when considering the extra difficulties in optimising the force field.

## 4.2.5 Force Field Parameters

Deciding on the form of the force field is only the first part of the process of building the force field. Next, the level at which the parameters should be specified must be decided.

For example, the strain free bond angle  $\theta_0$  can be specified at many different levels of complexity. At the simplest level it can be defined so as to have a value relating only to the central atom type (e.g. for  $C_{sp^3}$ ,  $\theta_0 = 110.5^{o87}$ ). The next level of complexity is to define an appropriate strain free bond angle for each of the possible bond angles where the types of the three atoms involved in the angle are used (e.g. for  $C_{sp^3}-C_{sp^3}-C_{sp^3}$ ,  $\theta_0 = 109.5^{o88}$ ). Finally it is possible, to some extent, to take into account the other atoms connected to the central atom by having a different reference angle for each possible degree of substitution (e.g. for  $C_{sp^3}$  with 3 H,  $H-C_{sp^3}-H$ ,  $\theta_0 = 108.2^{o89}$ ).

It can be seen that this will give rise to a large variation in the number of possible parameters and the same is also true of most of the other parameter types. For example, the torsional barrier (or barriers) can be specified for all four atom types involved or by taking just the central two atoms, a situation that will require fewer parameters. It is found, as might be expected, that the greater the complexity and number of parameters, the more accurate the optimised force field becomes. This greater accuracy is of course, made at the expense of the ease and speed of the force field optimisation.

To examine this balance further it is necessary to inspect the numbers involved. To try and optimise a full MM3<sup>74</sup> types force field for the 30 or so atom types which would be required to cover a reasonable cross section of organic molecules would require:

- ~ 300 parameters for each  $k_i$ ,  $l_0$ , A and B
- ~ 2,000 values for  $k_\theta$ ,  $k_\theta'$ , and  $\theta_0$
- ~ 10,000 values for  $v_n$

This is an impossibly large force field to optimise as not only would each parameter have to be optimised in turn but it would be necessary to make sure that there were a significant number of molecules in the optimising structure set that used each of the possible parameters. That is, to get a reasonable optimisation for a parameter it is necessary to

possess information on a sufficient number of cases where that parameter will be used. So if very specific conditions for the application of a certain parameter are set, it can be expected that the parameter will be employed less often than a less specific parameter. Thus there will be a need for more structures if each parameter is to be utilised a similar number of times compared to the less specific cases.

If such a large force field is impractical to optimise then a way must be found of reducing the number of parameters that need to be optimised. The obvious way of doing this is to use a lower degree of complexity for the parameters. For example, as stated earlier, reducing the specificity of the torsional barrier to just specifying the types of the central two atoms will have a drastic effect on the number of parameters required. A problem with this approach is that in going from the most complex but unimplementable levels to the less complex levels there is often an unacceptable drop in accuracy of the results produced by the force field. There is no point in getting the number of parameters that need to be optimised down to a workable total if it is then impossible to produce any reasonable results with them.

One possible method for overcoming this problem is that if a pattern can be observed in the parameter values it may be possible to come up with some general rules from which one can calculate the parameters from a combination of some of the already present parameters and a few new ones. For example, calculation of reasonable bond stretching force constant can be done using the bond length parameters and a few general parameters that can be used for all possible combinations of bonds<sup>90</sup>.

#### 4.2.6 Minimisation of Steric Energy

As well as producing a value for the steric energy of a system a method is also required to alter the co-ordinates of the atoms in the system until an equilibrium is reached. The minimisation can be executed in one of two co-ordinate spaces, namely internal or Cartesian.

The internal co-ordinates of a molecule are its bond lengths, valence angles, and torsion angles. All other quantities such as non-bonded distances are dependent functions of the

internal co-ordinates. It is usually computationally inconvenient to perform minimisations in internal space, the vast majority of programmes use Cartesian co-ordinates.

The Cartesian co-ordinates are the x, y, and z co-ordinates (usually in Å) of its constituent atoms. There is some redundancy built into the Cartesian co-ordinates. For any molecule of N atoms it requires 3N Cartesian co-ordinates to define the molecule in space, but only 3N-6 co-ordinates are necessary to completely define the atomic positions in internal co-ordinates.

#### 4.2.6.1 Pattern Search

A simple method of energy minimisation in Cartesian space is as follows. First the steric energy of the starting structure is calculated. Then some initial shift value is chosen (say 0.1Å), which is added to the x co-ordinate of the first atom and the steric energy is re-evaluated. If the energy has gone down the atom is left at its new position and the new steric energy becomes its current steric energy. However, if the new steric energy went up then twice the current shift value (in this case 0.2Å) is subtracted from the x co-ordinate ( $x_1+0.1$  to  $x_1-0.1$ ), the original steric energy is retained as the current value, and the new steric energy is calculated. If the new energy is lower than the current steric energy then the new atomic position is accepted and the current steric energy is updated. If, on the other hand the energy goes up again the co-ordinate is reset to its original position and the present steric energy is retained. This process is repeated for all the co-ordinates for all the atoms in the molecule. The entire process above is repeated until no further lowering of the steric energy is possible. At this point the shift is halved (in this case to 0.05Å) and the above algorithm repeated. The calculation is terminated when the shift reaches a suitably small number (e.g.  $10^{-5}$ Å). This is a robust procedure which is guaranteed to find a local energy minimum.

In the pattern search procedure as each successful atomic shift is found it is noted. When all the atoms have been sampled the programme has stored a pattern of successful moves for all atoms in the molecule. On the basis of "what was good once will be good again" the entire pattern is repeatedly applied until no lower energy can be found. The pattern of

directions is then applied with a halved shift value until no further lowering of the energy is possible. This process is repeated until the shift is sufficiently small. The shift is then reset to half of its original value, another pattern of successful moves established by trial and error, and the pattern applied as before. The entire process is repeated until either no successful moves can be found or the shift becomes too small.

#### 4.2.6.2 Newton Raphson

The total force for each co-ordinate is zero when a molecule is in a local energy minimum, but it should be noted that this condition is also true at an energy maximum or a saddle point. The force is given by minus the partial derivative of the steric energy  $E_s$  with respect to each co-ordinate and each component of this should be zero, i.e.

$$\frac{\partial V_s(x)}{\partial x_i} = 0 \quad i = 1, 3N$$

where  $N$  is the number of atoms in the molecule and  $x$  is a  $3N$  long vector of the current atomic Cartesian co-ordinates.

The aim of any minimisation technique must be to systematically alter the positions of the atoms until the partial derivative of the steric energy with respect to each co-ordinate is zero and that such a system should be both consistent and reliable in its results.

Assuming that the energy function is at a point  $x_s$  close to the minimum energy then it is possible to expand the steric energy in a Taylor Series. As  $x_s$  is close to the minimum it will be reasonable to truncate the series after the linear terms:

$$V_s(x_s + \partial x) = V_s(x_s) + \sum_{i=1}^{3N} \frac{\partial V_s(x_s)}{\partial x} x_i$$

Taking the case where the value of  $\partial x$  is such that  $x_s + \partial x$  is the location of the minimum, then at this minimum the derivatives of the energy with respect to each of the  $3N$  co-ordinates must be zero. Thus differentiating:

$$\sum_{i=1}^{3N} \frac{\partial V_s(x_s + \partial x)}{\partial x_i} = \sum_{j=1}^{3N} \frac{\partial V_s(x_s)}{\partial x_j} + \sum_{j=1}^{3N} \sum_{i=1}^{3N} \frac{\partial^2 V_s(x_s)}{\partial x_i \partial x_j} \partial x_j = 0$$

This can be more simply shown in matrix notation as:

$$\Delta V_s(x_s + \partial x) = \Delta V_s(x_s) + F_s \partial x = 0$$

where  $\Delta V_s$  is the gradient of  $V_s$ , and  $F_s$  is the matrix containing the second derivatives of  $V_s$ .  $F_s$  is a  $3N \times 3N$  square matrix of the second partial derivatives  $V_s''(x_i x_j)$  of the steric energy with respect to the Cartesian co-ordinates.  $\Delta V_s(x_s)$  is a  $3N$  long vector of the first partial derivatives  $V_s'(x_i)$  of the steric energy with respect to Cartesian co-ordinates. By subtracting  $\Delta V_s(x_s)$  from both sides and then multiplying by the inverse matrix of  $F_s$  the following equation can be obtained:

$$\partial x = -F_s^{-1} \cdot \Delta V_s(x_s)$$

where  $F_s^{-1}$  is the inverse of  $F_s$ .

The Newton Raphson algorithm calculates the elements of the matrix  $F$ . This being the matrix of second derivatives of  $V_s$ , i.e.

$$F = \frac{\partial^2 V_s}{\partial x_i \partial x_j}$$

#### 4.2.6.3 Full Matrix Newton Raphson

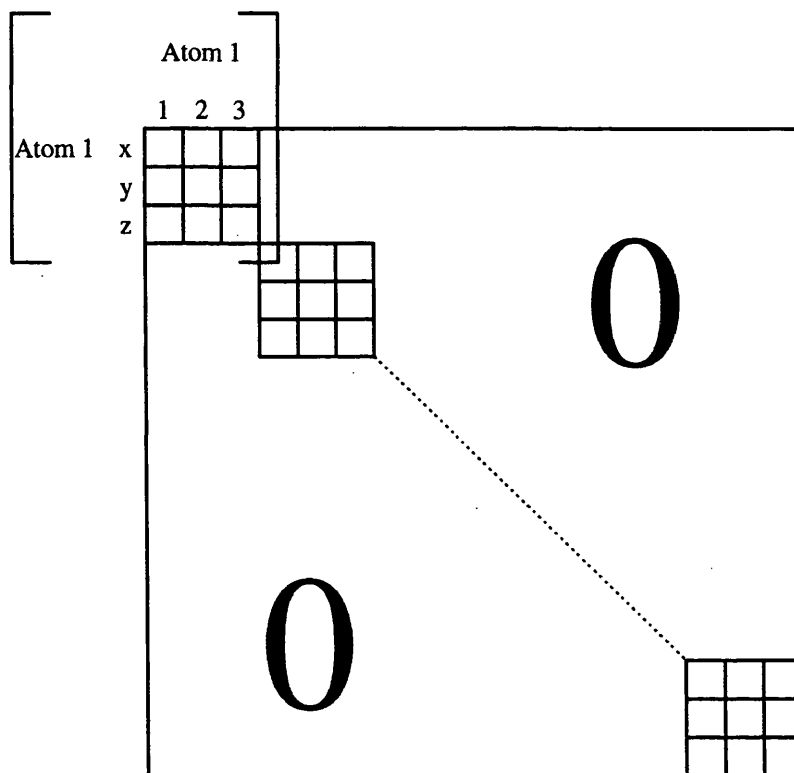
The full matrix Newton Raphson converges extremely rapidly (5 or 6 iterations at most from a point within the radius of convergence) upon a local energy minima. Ideally at an energy minimum, as stated previously, the components of  $\Delta V_s(x_i)$  should be zero but in practice 5 or 6 iterations will reduce the rms component value to around  $10^{-7}$  kcalmol<sup>-1</sup>Å<sup>-2</sup> which is close enough to zero.

The full matrix Newton Raphson algorithm has a number of disadvantages particularly for large molecules. The full matrix of second derivatives is extremely time consuming to

compute and invert, and the radius of convergence is very small, so that even calculations on carefully constructed models of molecules with apparently good geometry will not converge.

#### 4.2.6.4 Block Diagonal Newton Raphson

It is possible to trade the radius of convergence for rate of convergence by using approximations to the full matrix  $F_s$ . The most commonly used approximation results in the block diagonal Newton Raphson method. In the full matrix<sup>91,92</sup>, a derivative for each possible combination of co-ordinates would be calculated. The co-ordinates that will produce the greatest effect on each other are those relating to the same atom. If just the interaction of these co-ordinates are used the result will be a block diagonal matrix<sup>93</sup>



$$\begin{vmatrix} V_s''(x_1x_1) & V_s''(x_1y_1) & V_s''(x_1z_1) \\ 0 & V_s''(y_1y_1) & V_s''(y_1z_1) \\ 0 & 0 & V_s''(z_1z_1) \end{vmatrix}$$

**Diagram 4-10: Block Diagonal Newton - Raphson**

Only the second derivatives in the collection of 3x3 blocks along the leading diagonal are calculated, all of the others are set to zero. The matrix is symmetrical, so that within each block there are only six unique elements (N.B. there are no derivatives involving a co-ordinate of one atom with a co-ordinate of another,  $x_i$  and  $x_j$  always come from the same atom). Block diagonal Newton Raphson is popular because the radius of convergence is large, encompassing most of the crudest of models usually constructed with a molecular graphics system and convergence to the local minimum is reasonably quick; 50 - 100 iterations is typical for most calculations on small organic molecules. There are some relatively minor drawbacks to the Block Diagonal Newton Raphson (BDNR) algorithm, because of the construction of the F matrix BDNR is not very good at adjusting torsion angles in really flexible molecules. If this is a problem then BDNR can be used as a preminimiser (BDNR will optimise almost all structures so that they are within the radius of convergence of the full matrix Newton Raphson algorithm in short computation times) for full matrix Newton Raphson which does not have this problem.

#### 4.2.6.5 Calculations of Derivatives

The derivatives required can be determined in two ways. In the simplest of these, numerical derivatives, the atomic co-ordinates are moved by a small amount and the energy recalculated. These energies can then be used to estimate the required derivatives. The second method, analytical derivatives, are determined by applying calculus to the various steric energy terms.

This second method has the advantage that the minimisation is faster as it does not require the multiple energy calculations of the numerical method. For analytical derivatives the following sum of derivatives is required:



$$\sum_{i=1}^{3N} \frac{\partial E}{\partial x_i}$$

Each derivative can then be expanded to include all the contributions towards the total energy i.e.:

$$\frac{\partial E}{\partial x} = \frac{\partial E}{\partial l} \cdot \frac{\partial l}{\partial x} + \frac{\partial E}{\partial \theta} \cdot \frac{\partial \theta}{\partial x} + \frac{\partial E}{\partial \omega} \cdot \frac{\partial \omega}{\partial x} + \frac{\partial E}{\partial r} \cdot \frac{\partial r}{\partial x} + \frac{\partial E}{\partial \chi} \cdot \frac{\partial \chi}{\partial x}$$

The following are typical analytical derivatives with the original formulae from which they were determined:

- **Bond Length**

If the energy for each bond is given by:

$$\begin{aligned} E_1 &= \frac{1}{2} k_1 (l - l_0)^2 \\ &= \frac{1}{2} k_1 (l^2 - 2ll_0 + l_0^2) \end{aligned}$$

then its derivative will be:

$$\begin{aligned} \frac{dE_1}{dl} &= \frac{1}{2} k_1 (2l - 2l_0) \\ &= k_1 (l - l_0) \end{aligned}$$

- **Angle Distortion**

If the energy for each angle is given by:

$$E_\theta = \frac{1}{2} k_\theta \left[ \Delta\theta^2 - k_\theta' \left[ \left| \Delta\theta^3 \right| - .0004 \left| \Delta\theta^5 \right| \right] \right]$$

where:

$$\Delta\theta = (\theta - \theta_0)$$

then its derivative will be:

$$\frac{dE_{\theta}}{d(\Delta\theta)} = \frac{1}{2}k_{\theta} [2\Delta\theta - Sk_{\theta}'(3\Delta\theta^2 - 0.002\Delta\theta^4)]$$

where S = sign of  $\Delta\theta$

$$\begin{aligned} \frac{d(\Delta\theta)}{d\theta} &= 1 \\ \therefore \frac{dE_{\theta}}{d\theta} &= \frac{dE_{\theta}}{d(\Delta\theta)} \cdot \frac{d(\Delta\theta)}{d\theta} \end{aligned}$$

- **Torsional Twist**

If the energy for each torsion angle is given by:

$$E_{\omega} = v_1(1 + s\cos(\omega)) + v_n(1 + s\cos(n\omega))$$

then its derivative will be:

$$\frac{dE_{\omega}}{d\omega} = -v_1s\sin(\omega) - v_n(ns\sin(n\omega))$$

- **Out of Plane Bending**

If the energy for the out of plane bending is given by:

$$E_{\chi} = \frac{1}{2}sk_{\chi}(180^{\circ} - |\chi|)$$

then its derivative will be:

$$\frac{dE_{\chi}}{d\chi} = sk_{\chi}(180^{\circ} - |\chi|)$$

where s = sign of  $\chi$

- **Non Bonded Interactions**

If the energy for the non - bonded interaction is given by:

$$E_r = Ar^{-6} + Br^{-12}$$

then its derivative will be:

$$\frac{dE_r}{dr} = -6Ar^{-7} - 12Br^{-13}$$

- **Coulombic Function**

If the energy for each coulombic interaction is given by:

$$E_q = k_q \frac{q_i q_j}{r^2}$$

then its derivative will be:

$$\frac{dE_q}{dr} = -2k_q \frac{q_i q_j}{r^3}$$

As well as calculating the absolute values of the individual forces it is also necessary to determine the direction of each force so that the overall force on the molecule can be determined. In the case of the Newton Raphson method the second derivatives need to be determined

For numerical derivatives the equations used are as follows:

$$\frac{dE_s}{dx_i} = \frac{E_s(x_i + \partial) - E_s(x_i - \partial)}{2\partial}$$

The second derivative of the above equation is:

$$\frac{\partial^2 E_s}{\partial x_i^2} = \frac{E_s(x_i + \partial) + E_s(x_i - \partial) - 2E_s(x_i)}{\partial^2}$$

$$\frac{\partial^2 E_s}{\partial x_i \partial x_j} = \frac{E_s(x_i + \partial_j x_j + \partial) + E_s(x_i - \partial_j x_j) - E_s(x_i x_j + \partial) + E_s(x_i x_j)}{\partial^2}$$

#### 4.2.6.6 Reducing Oscillation During Minimisation

It is found in some situations that have bad initial structures that from one iteration to the next the root mean square (rms.) force will increase rather than decrease. The reason for this is often that the initial situation has two or more atoms whose van der Waals surfaces are slightly overlapping. This produces a large gradient on the energy surface and causes the atoms to be moved an accordingly large distance apart. In many cases this can move one or more atoms such that the result is an even greater overlap with another atom, which will lead to a steeper gradient and so the cycle will continue resulting in a larger effect every time.

When the shift on an atom is calculated it is usually compared with a maximum shift and is reduced to this value if it is found to be greater. The reason for this is that a very large steep slope would result in the atom moving an unrealistic distance. The technique used to counter the oscillation problem is to allow the programme to alter the value of this maximum shift. There are two ways this has been implemented.

The simplest case is where the energy of the system is calculated at the end of an iteration. This is compared with the lowest energy found so far and, if greater than this by an amount determined empirically from the current maximum shift, the co-ordinates are reset to those at the beginning of the iteration. The maximum shift is reduced and the minimisation is repeated again until either the energy gain is not more than the empirical level or the energy decreases. The main disadvantage with this algorithm is that when using numerical derivatives it increases the time per iteration by about 20%. A major advantage is that the energy is available at the end of each iteration which provides a far more user friendly indication of the progress of the minimisation than a record of the rms. force.

The second approach uses the fact that it is possible to get an estimate of the energy by summing up the individual energies while going through the atoms in the molecule calculating the numerical derivatives. This has the advantage that it requires little extra calculation. The drawback is that if the energy of the molecule does start to rise it will tend to appear in the iteration after the bad shift has been made. Thus, if this algorithm is used,

the programme is required to save the last two sets of atomic positions and return to the earlier of these if any substantial energy rises occurs.

While both these systems have been found to work well, especially in situations involving separate highly charged molecules, the first is preferred because in systems that do require this option there is little time advantage over the second method and this is more than compensated for by the advantage of being able to see exactly how the energy is varying during the course of the calculation.

## 5 Methods and Materials

### 5.1 Sequence Alignment

The first step to starting a homology modelling project is to find a protein of known structure, the parent, that is 'homologous' to the unknown protein structure, the target. The term homologous is used to imply that the proteins in question have a similar sequence and a similar overall fold structure.

It has been known for some time that proteins from different sources and with quite diverse function can have a similar primary sequence. These sequences are grouped into what is known as protein families. The primary sequences of these proteins are similar even over different species. The more diverse members of a protein family may have a completely unrelated function. For example some of the more diverse and distantly related members of the serine protease family have lost their enzymatic ability. The variability of different residues at different positions in the primary sequence of these families allows us to develop ideas on the likelihood of an amino acid being substituted for another amino acid.

The first is that certain amino acids will be substituted in the same position in the primary sequence of different proteins in a family more often than certain other residues. This is because of the environment the amino acids find themselves in at that particular position, i.e. polar, non-polar, buried or exposed to the solvent. If an amino acid is substituted for a different amino acid that is of similar size and chemical properties then very little change in the structure of the protein will result because of this substitution. Whereas if the amino acid was substituted for one of different size and more importantly different chemical properties then the substitution will cause strain in the area around the substitution in the protein structure. This can cause the protein to have different properties and be less well adapted at carrying out its biological role in the organism. Therefore amino acids with similar chemical properties and size will more readily be substituted in a primary sequence than amino acids with different chemical properties and size. This allows us to calculate the probability of an amino acid being substituted for another amino acid in the primary sequence.

Secondly, statistically if a certain amino acid substitution happens less often than it is likely to then that substitution must be a radical change, and substitutions that occur most frequently are conservative substitutions. The problem with this statistical approach is that comparisons of the rate of substitutions and the choice of the most favourable substitution can be affected by the data base from which the information was extracted. The statistical method relies on looking at families of proteins. The actual results you get depend on which protein family is used in the analysis. Therefore, analysis of several protein families is more likely to lead to a more true representation of the probability of certain substitutions taking place than analysis of one or two protein families.

The statistical approach leads to the development of series of mutation matrices showing the likelihood of an amino acid being substituted for another amino acid. There are several matrixes that are commonly used in sequence alignments to calculate the probability that an amino acid will be substituted by another amino acid e.g. the PAM<sup>94</sup> and Blosum<sup>95</sup> series of matrices.

Sequence alignment routines look to maximise the score obtained using a mutation matrix while aligning two or more protein sequences. The routine searches for the optimal alignment of the sequences by trying to maximise the score between each pair of residues at each position in the alignment. Penalties are added to the score if a gap is introduced to the alignment and there is a penalty for increasing the gap size. There are a few different algorithms used in producing sequence alignments but the most popular alignment algorithm is the Needleman and Wunsch algorithm<sup>96</sup>. All the alignments depend on which mutation matrix is used to score substitutions. Different matrices produce quite different results, especially over less closely related sequences. This is due to the different statistical methods and different data sets used in producing the mutation matrix. Another major factor in producing sequence alignments is the gap creation and gap extension penalties that are used in the alignment algorithm.

The sequence alignment routines can produce good results while aligning two sequences together, but they are not so successful at multiple sequence alignments. When carrying out multiple sequence alignments it is not a feasible option to carry out a pair wise score for

every position but only over a selected range. It is therefore very possible for multiple sequence alignment routines not to give the best alignments possible. In general the multiple sequence alignment produced can be improved upon by manual alteration to the alignment.

A new improvement to sequence alignment routines is to take the secondary structure of the protein into consideration while carrying out sequence alignment. Studies of protein families show that secondary structure features are conserved more strongly in members of a protein family than the primary sequence is. It is also known that insertions and deletions are much less likely to occur in an  $\alpha$  helix or  $\beta$  strand structure. Unfortunately this comes into its own only if the crystal structure of a protein which is a member of the family is known so that an accurate description of the secondary structure can be calculated. This procedure is much better at aligning distantly related sequences than the procedures that do not take secondary structure into consideration.

### 5.1.1 Homology scan of Brookhaven Database

To model a sequence with unknown structure it must be homologous to a protein with known structure. Therefore a search of the database of proteins with known crystal structures is required. The largest most up to date database of known protein structures is the Brookhaven database<sup>97</sup>. This database is available on line over the internet at several sites, as well as CD-ROM and tape that are updated quarterly. The on-line services over the internet offer a more up to date version of the database and so offer the latest structures to be released in to the database.

To find the optimal alignment of a sequence to every sequence in a database, even one as small as the Brookhaven database, is extremely time consuming. The only way to carry this out in a reasonable time scale is to use a processor designed specifically to carry out sequence alignments very rapidly. The "MasPar" processor is one such processor. The circuitry in the processor has been developed specifically to run the Smith and Waterman alignment algorithm<sup>98</sup>. What this means is that the algorithm has been hard wired into the silicon chip and the processor has been specially parallelised to speed up the calculation. This is a very specialised and expensive method of carrying out a database homology



search which relies on the fact that the algorithm used is correct and the best algorithm for finding homology between sequences. Having a specific machine to carry out database homology searches can only be justified if the database has to be thoroughly searched frequently with many different sequences so that it is kept busy. Therefore these machines are only found in large pharmaceutical companies or research departments which specialise in offering the service to other users (Seqnet).

If no access to such a processor is possible a scanning routine can be used to find protein sequences in the database that are similar to the unknown sequence. These scanning routines are less exact than the sequence alignment routines but offer a quicker method of searching the database to find possible candidates which can then be extracted from the database and a full sequence alignment carried out.

The output from the database scanning routines is a list of the top scoring sequences found in the database in descending order. Only the top fifty or one hundred sequences are normally printed out as these are the sequences with the most likely chance of being homologous to the target sequence. It is common to find several members of the same protein family near the top of the list. This can indicate that the target sequence is related to that family of proteins.

If no good match is found in the Brookhaven database it is then necessary to use protein sequence databases to see if the target sequence is related to any known sequences. Doing this enables the protein to be categorised into a protein family and hopefully a member of this family will have known structure and will be present in the Brookhaven database. This method relies on the fact that the fold pattern a protein adopts is more strongly conserved than the primary sequence. It is possible for two proteins with only very weak homology to have the same fold pattern.

## **5.2 Homology Modelling**

Homology modelling is an extension to the Comparative modelling<sup>99,100,101</sup> methods that can be used to construct a three dimensional model structure of a new protein from knowledge

of its sequence, and the crystallographic structures and sequences of other members of its homologous family.

It has long been apparent that proteins from very different sources and sometimes diverse functions can have homologous sequences and consequently a very similar three dimensional structure<sup>102</sup>. The structure, or pattern of folds, that a protein adopts appears even more conserved than primary sequence<sup>103,104</sup>. This fact is the basis of comparative modelling methods, which permit extrapolation from the experimentally determined structure of one or more members of a homologous family to a new member of this family whose sequence has been determined but whose structure is as yet unknown. A number of factors combine to increase greatly the application of comparative modelling techniques today. The large number of protein structures and the exploding number of sequences that are being submitted to the computerised databases provide the basic structural and sequence data needed to apply the method. The proteins in which we are interested in are often available in only small quantities, too little for structural studies unless the gene or mRNA is cloned or synthesised and expressed. Even when this latter effort is deemed worthwhile and is initiated, the comparative modelling studies can be performed in the meantime, providing an approximate view of the structure of the molecule until sufficient protein can be obtained and the experimental structure determined.

The first stage in both comparative and homology modelling is to align the sequence of the protein of unknown structure (target protein) to the family of proteins that it belongs to. In comparative modelling the sequence alignment is carried out on a structural basis rather on sequence identity. This is accomplished by examining the members of the family with known structures. These known structures are superimposed in three dimensions to obtain a maximal overlap of the structures. Once superimposed, there are parts that overlap very well, indicating the structures are well preserved in these regions, they are called "structurally conserved regions", SCRs. These SCRs are usually composed of secondary structure elements, the immediate active site, and other essential structural framework residues of the molecule. Between these conserved elements are highly variable regions which differ significantly from one member of the family to the next. These are called "variable regions", VRs. They are almost always loops that lie on the external surface of

the protein, and they contain all the additions and deletions between different protein sequences. The alignment is primarily concerned with the SCRs, since these are the portions of the structure that are the same in all protein members of the family.

An alternative, when not so much structural information is available, is to align the sequence of the unknown structure with the sequence which has the highest homology with the unknown in the structural database. The first stage is to scan the structural database with the target sequence for protein sequences in the database with which it is homologous. The highest scoring sequence from the database is used as the starting structure for the modelling study. The known and target structures' primary sequences are aligned using a sequence homology scoring function. Once the two sequences are aligned it is possible to examine where the deletions and insertions occur on modifying the starting structure to the model of the target structure. Most of the insertions and deletions, particularly the larger alterations in the primary sequence occur in loops that lie on the external surface of the protein.

Many of the modelling methods employed in comparative and homology modelling for substitutions, deletions and insertions are the same. The only difference is in modelling the deletions and insertions. With comparative modelling there are many more known structures from the same protein family. This means that occasionally the primary sequence of a variable region of one of the known structures is homologous to the corresponding section in the sequence of the unknown. This VR can be added to the model of the unknown structure directly as it is likely that they will have the same conformation. The more known structures in protein family the better the chance that the VR of the unknown will already exist in one of the known structures. This tends to happen a lot less in homology modelling where the sequences are less similar to members of the protein family.

### **5.3 Substitutions**

Amino acid substitutions are more likely to occur in areas of the primary sequence that correspond to non critical regions in the protein structure. These regions tend very much to be at or near the surface of the protein in regions with no secondary structure. A

substitution will cause less strain if the side chain of the substituted amino acid points towards the surrounding solvent. If a substitution occurred in the core of the protein it is highly probable that it will cause a large amount of steric stress. This is because the core region of globular proteins is tightly packed with very little room for any extra atoms. If a substitution were to decrease the number of atoms then a hole in the core region of the protein would result. This increases the energy of the protein. If on the other hand the substitution were to increase the number of atoms then there would be no space for the extra atoms as the hydrophobic core of globular proteins is a close packed structure with no holes. The atoms surrounding the substitution would move to compensate for the increased number of atoms causing high energy strain around the substitution. The core of globular proteins is also highly hydrophobic so all possible hydrogen bonds between polar atoms within the core are formed to reduce the strain of having these polar atoms in a hydrophobic region. If a substitution were to take place that removed a polar atom or introduced a polar atom into the hydrophobic area then it would cause a polar atom to be in an unfavourable position (surrounded by non polar atoms). This would lead to an increase in energy of the protein about this area.

This allows sections of the primary sequence that are involved in the core packing to be associated with regions of low substitution. Regions of secondary structure can also be identified by looking at the alignment of a protein family as again substitutions in the primary sequence are more likely to cause steric strain in sections of repeating secondary structure than they are in a region of coil. Secondary structures are tight packed structures with hydrogen bonds forming between the polar atoms of the peptide chain. Any disruption in this order will cause an increase in the energy of the surrounding residues and will make the secondary structure less stable.

### 5.3.1 Conservative and non - Conservative Substitutions

Substitutions are the simplest modification to model as they cause the least amount of disruption to the model. There are two types of substitution in protein modelling, conservative substitutions and non-conservative substitutions. Conservative substitutions are when a residue is substituted for another residue of similar size and structure e.g.

valine for leucine. This causes minimal additional steric strain to the protein model and polar atoms in the substituted residue make the same contacts as the residue that was removed. With conservative substitutions the side chain of the new residue follows the same path as the removed residue's side chain and there is very little disturbance to the surrounding atoms.

Non-conservative substitutions can be more disruptive, particularly if changing a non polar residue for a polar residue or a small side chained residue for a long side chain residue. If a polar residue is substituted for a non-polar residue in the core of the protein then a polar atom is introduced into a hydrophobic environment increasing the strain in the protein. The reverse of this, substituting a non-polar residue for a polar residue can leave the hydrophobic side chain in a hydrophilic environment, or if the residue is buried can leave an unpaired polar atom from another residue in a hydrophobic environment. Substantially altering the size of the residue when substituting residues can cause steric strain in the protein model. If a large side chain residue is substituted for a small side chain residue this can cause severe steric clashes as there is not enough space for the new larger side chain to fit in. This causes the surrounding atoms to adjust to accommodate the larger side chain. For substitutions where a large sidechain residue is substituted for a small sidechain residue then a gap can be left in the protein structure. The core of the protein is a close packed structure with only very few and small gaps found in the structure. If a large gap is formed the surrounding atoms will move around and try to minimise the size of the hole disrupting the close pack order of the protein core. Over all non-conservative substitutions are more disruptive than conservative substitutions and can involve careful monitoring. As non-conservative substitutions are more disruptive they have a tendency to occur at or near the protein surface where any disruption can be minimised.

Substitutions are the simplest modification to a model as they cause the least amount of disruption to the model compared to deletions and insertions. As described above there are two types of substitution, conservative and non conservative, both are treated in a similar manner when modelling the changes from going to the parent protein to the target protein. The only difference will be that a conserved substitution should be less disruptive than a non conserved substitution.

Modifying the primary sequence of a protein is an easy process using functions that are present within COMMET. There are routines available that allow a protein model that has already been loaded into the system to be manipulated and altered. Substitutions can quickly and easily be carried out within the COMMET package. A very similar method of substituting one amino acid for another is used for conservative and non - conservative substitutions. There is an extra step in the placement of non - conservative substitutions to find the global minimum of the side chain.

As the only difference in the 20 amino acids is the side chain atoms the routines that are used in the substitution do not alter the conformation of the protein backbone in any way. In both conservative and non- conservative substitutions the substitution routine is called to change one amino acid into another. The routine starts by removing the atoms in the old side chain. It then places the new side chain atoms along the same conformation as the old conformation until there is a difference in the structure between the old and new side chains. The reasoning behind this is that the old side chain is in its preferred conformation which should be the lowest energy conformation. Also the substitution should cause the least amount of disruption to neighbouring atoms and the most effective way of minimising the disruption is to try and follow the conformation of the old side chain. In conservative substitutions where the new side chain is the same length or shorter than the original side chain the new side chain's atoms should hopefully be close to the global minimum in which case the minimiser will find the global conformation. In the case of non - conservative substitutions the atoms in the new side chain that have no corresponding atom in the old side chain are left in an extended chain conformation by the substitution routine. To define the rest of the atoms in the new side chain another routine called SITAR must be used.

### 5.3.2 SITAR

'SITAR' stands for Sequential Iterative Torsion Angle Refinement. It works by rotating about the  $\chi$  torsion angles one at a time out along the length of the side chain until the lowest energy conformation is found. It starts at the  $N_{\text{amide}}-C_{\alpha}-C_{\beta}-C_{\delta}$  ( $\chi_1$ ) torsion angle and rotates it a set amount usually specified by the user. At each step the energy of the residue is calculated taking into account the rest of the protein and if the energy is lower at

this step than the previous lowest energy step this new conformation for the  $\chi_1$  is stored. The torsion angle is rotated a full  $360^\circ$ , at this point the conformation of the  $\chi_1$  torsion angle is set to the lowest energy conformation found.

The routine now goes to the next torsion angle, which is the  $C_\alpha-C_\beta-C_\gamma-C_\delta$  ( $\chi_2$ ) torsion angle and carries out the same procedure. That is it steps through the torsion angle the user specified step calculating the energy of the amino acid at each step. If the energy is lower for this step than so far found the conformation is stored. The angle is rotated in steps for  $360^\circ$  when the  $\chi_2$  torsion angle is set to the conformation which gave the amino acid side chain its lowest potential energy. This is continued until the routine gets to the end of the side chain.

At this point the routine goes back the first torsion angle in the side chain,  $\chi_1$  torsion angle and again rotates around this torsion angle in steps specified by the user calculating the potential energy of the amino acid at every step. If the energy is lower at any step the conformation is stored. Again after the torsion angle has been rotated  $360^\circ$  the  $\chi_1$  torsion angle is set to the conformation with the lowest energy. The routine carries on out along the side chain in this manner. The routine keeps repeating the search until the energy of the amino acid cannot be improved for an entire search of the length of the side chain. When this occurs the routine ends setting the amino acid's side chain to the conformation of lowest energy which it found.

## 5.4 Deletions

Deletions are the next least disruptive mutation to occur in proteins. Along with insertions they are most commonly found to occur on the surface of proteins in loop regions. This is because the deletion cause less disruption to the packing of the protein and does not disrupt any secondary structure if it occurs on the surface of the protein in a loop structure.

What can be seen happening in deletions is a loop or bulge section of the protein is lost. When this happens the remaining peptide smoothes over this area becoming more extended with the deviation of the backbone atoms from their original position quickly becoming

smaller the further from the deletion the backbone is. The effect of a deletion can be fairly localised to about five or six residues from the deletion. Residues further than this distance from deletions can be seen to be unaffected by the deletion in the conformation they adopt in the altered model.

#### 5.4.1 Single amino acid Deletions

COMMET allows single amino acid deletions to be carried out automatically. There is a single routine the user can use which removes all the atoms in the deleted residue. It then connects the two free atoms in the adjacent residues ( $N_{amide}$  and  $C'$  atoms) to form a new long peptide bond. A section of polypeptide chain is allowed to relax using an energy minimiser. Five residues either side of the deletion are allowed to move while carrying out the energy minimisation over this section of the polypeptide chain. This is enough to relieve the protein of strain caused by the deletion. The peptide bond rapidly closes and because of the force field adopts the proper trans conformation.

During the energy minimisations carried out to close the gap in the polypeptide backbone only selected atoms in the protein model are free to move. These atoms are the atoms in the five residues either side of the insertion and the atoms in the surrounding vicinity of the removed side chain. Only the residues that were within  $5\text{\AA}$  of the deleted residue are allowed to move in the energy minimisation calculation. The rest of the atoms in the model are held rigid. This set up of holding most of the atoms in the model rigid is a compromise in the accuracy of the energy minimisation while allowing the calculation to proceed at a reasonable rate. The atoms in the calculation that are free to move can still feel the effect of the surrounding rigid atoms during the energy minimisation. This takes into some consideration the environment around the deletion without making the calculation too complicated and time consuming. This also has the effect that the backbone atoms are effectively fixed in position  $i+5$  and  $i-5$  residues from the deleted residue  $i$ . As the backbone moves from the point of deletion to the fixed atoms of the backbone the atoms of the backbone deviate less from their initial starting position in agreement with crystallographic evidence.



This method by no means perfectly recreates the effect of having all the atoms of the protein included in the minimisation but allows the minimisation to be carried out at a reasonable rate. It is often the case that the complexity of the model that the user wants to simulate is severely limited by the time available to run the simulation. Compromises have to be made to reduce the complexity of the model and hence speed up the simulation at the expense of accuracy or more importantly reliability.

This also affects when hydrogen atoms are explicitly modelled. Adding explicit hydrogen atoms to the model of a protein can treble the number of atoms in the simulation. Therefore a great amount of effort has been put into developing force fields that can model the effects of hydrogen atoms implicitly without the need of having to explicitly position the hydrogen atoms in the model. Although a simplification in the model, for the modelling of substitutions, deletions and insertions the approximations of the hydrogen atom effects are a close enough simulation of the hydrogen atom as other forces and approximations affect the model to a much greater degree. Ignoring the hydrogen atoms while modelling the substitutions, deletions and insertions allows the models to be much simpler and simulations to run faster without compromising the accuracy of the model. The hydrogen atoms are added at a later stage of the modelling when the primary sequence of the new model is correct and global minimisation is being carried out.

#### 5.4.2 Deletions up to 3 residues

For the slightly larger deletions of up to three residues the above method remains a quick and reliable method for modelling. Again it is common to find these small amino acid deletions in loops or where the backbone is not in an extended conformation. This leads to a smaller gap being formed than might be expected on the removal of three residues from the model.

After the removal of the three residues from the model the two free atoms are joined to form an exceptionally stretched peptide bond. The same procedure is followed in minimising the backbone fragment. Five residues either side of the deletion are free to move and the surrounding amino acids less than 5Å from the deletion are included in the simulation. The rest of the model is kept frozen in the same position. Most of the steric

energy is from the extremely long and artificial peptide bond either side of the deletion. As this distance closes to its normal distance the energy of the model quickly drops. The peptide remains in a trans conformation due to the potential in the force field.

The small deletions are easily modelled as the remaining atoms in the model quickly fill in any gaps left in the structure by the removed atoms. Also because the deletions happen on the protein surface the atoms which are removed are likely to be at least partly surrounded in solvent and not other atoms from the model. This means that when a deletion occurs at or near the surface of a protein a rearrangement of the surrounding atoms is enough for the protein to find its global minimum again. It is not necessary to carry out conformational searches of the side chains of the surrounding amino acids as there is no major disruption.

### 5.4.3 Larger deletions

Larger deletions usually involve entire sections of a loop to be removed and the remaining loop to take on a different conformation. This involves remodelling the loop and treating the smaller sized loop as an insertion. The alignment of Bb with BT shows that no loop was larger than three residues therefore no modelling of larger deletions was carried out.

Having to model the large deletion as an insertion of the smaller loop region involves much more computationally intensive routines to try and calculate the conformation the loop will have in the model. As will be described in the next section the modelling of insertions is much more complicated than modelling small deletions and substitutions.

## 5.5 Insertions

Insertions are the most disruptive of the three alterations that can happen to the sequence of a protein. This is because space for the additional atoms has to be made in the structure and the backbone conformation has to change. As with the other two alterations to the primary sequence, insertions are much more common in loop regions at or near the surface of the protein. Small insertions of one or two amino acids do occasionally happen in the  $\beta$  sheets as a  $\beta$  bulge. This is where a residue bulges out of the sheet and takes no part in the hydrogen bonding between the strands of the  $\beta$  sheet but again this only occurs when the

strand is at the surface and the bulge heads out to the solvent rather than into the core of the protein.

### 5.5.1 Where insertions tend to be

Even more so than substitutions and deletions, insertions are found in the surface loop regions of proteins. This is because insertions are so very traumatic. As the loops point into the solvent they can more easily change their conformation to allow an insertion to occur without causing the entire structure of a protein to become too unstable for biological activity. If a large insertion were to occur in an internal region of the structure the steric strain would be such that the conformation would no longer be feasible and the protein would fold into another conformation losing its biological function in the process. Therefore it is normal to find the larger insertions on surface loops.

Large insertions of four and more amino acid residues can make a difference to the conformation an entire loop will have in a protein. The surface loop regions of a protein do not have a well-defined structure but the conformation changes over time. The conformation of the loop can vary quite dramatically in the crystals of proteins, so much so that particularly loops with large movements in the residues cannot have the structure of the loop determined by x-ray diffraction techniques. In solution the loop regions will have an even larger range of movement, so that by their very nature the loop regions often have no well defined conformation.

### 5.5.2 Small Insertions Of Amino Acids

Within COMMET there is a function that can be used to model insertions of a few amino acid residues in length quickly and accurately. This is because for small insertions up to 3 residues in length the backbone conformation only diverges from the original protein over a very limited number of amino acid residues. Like deletions, small insertions, as in one, two and three amino acids in length, cause very localised disruption to the protein structure. The distortion to the backbone from the original position only lasts for four or five residues in either direction of the insertion. It is therefore possible to carry out these small insertions with a fairly high degree of confidence and in a reasonable time scale.

To carry out these small insertions within COMMET the user invokes the insertion routine. This routine carries out a one residue insertion at a time. The routine asks for the amino acid that is to be inserted and the residue in the structure the new residue is to be inserted immediately after in the structure. When the routine has this information it first superimposes the new residue directly on top of where the insertion point was specified. It then breaks the amide bond between the two residues where the insertion is to occur and forms two new peptide bonds between the inserted residue and the two residues either side of the insertion.

The steric strain is too much for the normal Newton Raphson energy minimiser to handle at this stage therefore a constrained pattern search minimiser is used over a range of residues four amino acids either side of the point of insertion. The  $\phi$  and  $\psi$  torsion angles of the new residue are forced in turn to the 7 low energy  $\phi$   $\psi$  torsion angle combinations found in the Ramachandran map<sup>105</sup> (the A, C, D, E, F, G, A\* regions of the Ramachandran map). Minimisation calculations are carried out forcing the inserted residue's  $\phi$  and  $\psi$  torsion angles to each of the low energy regions of the Ramachandran map in turn. The conformation with the lowest energy after the minimisation is taken to be the correct conformation for the backbone at this position.

The pattern search minimiser used to carry out the minimisation of the highly strained region around the insertion point uses internal co-ordinates of the residue rather than the Cartesian co-ordinates. This is to give the best geometry and energy to the region around the chain break consistent with a minimum change in the position of the existing residues. This restricts the search to the limited number of torsion angles in the 'molten' zone since the bond lengths and valence angles are constrained. For small regions of a protein like this internal co-ordinates provide a quick and useful method to carry out minimisation as they ensure good geometry of the protein in the section allowed to move. Using Cartesian co-ordinates in such a strained situation can sometimes cause problems with the geometry. For large sections and whole proteins using internal co-ordinates to carry out an energy minimisation becomes more difficult as a small change in a torsion angle in one part of the peptide chain backbone can cause massive changes in the conformation in other parts of the model. This is because altering a torsion angle in the backbone of a polypeptide model

can cause sections of the chain to swing about by a large amount. Using Cartesian coordinates for polypeptide energy minimisations prevents these small changes in one section amplifying into large movements elsewhere in the structure.

## 5.6 Large insertions

To deal with large insertions is very expensive in computer time. As a result of this there are many different methods available in the literature to carry out the modelling of large insertions with varying degrees of complexity. The method available within COMMET modelling package is known as the loop conformation generator. Other methods exist for modelling large insertions but all are very computationally intensive. The reason for the problems involved in the modelling of large insertions is the large number of conformations an insertion can adopt due to flexible nature of the backbone. The more residues that a loop contains the more flexible it becomes and the theoretical number of conformations that it can adopt increases exponentially with the number of residues. Very quickly the sheer number of possible conformations of a loop becomes enormous.

### 5.6.1 Conformation Loop Generator

This is a method where all possible conformations that a loop can have are systematically built and tested with the 'best' loop being used in the final model. In this project the 'best' loop was the one with the lowest steric energy after energy minimisation. This method although very computationally intensive does guarantee that all possible conformations that the loop can adopt will be checked.

#### 5.6.1.1 Algorithm used

As discussed in an earlier chapter the peptide bond is planar and for this study can be considered to adopt a trans conformation. This leaves two torsion angles able to rotate, called the  $\phi$  and  $\psi$  torsion angles. The one exception to this is the proline amino acid. For proline the  $\phi$  angle is fixed due the five member ring that includes the  $N_{amide}$  and  $C_{\alpha}$  atoms. The only torsion angle free to rotate in the proline amino acid is the  $\psi$  torsion angle.

The algorithm generates all possible conformations a loop can theoretically have in an ordered manner. One way of generating the conformations is by starting at the first residue in the loop and rotating its  $\phi$  and  $\psi$  torsion angles by a set amount through  $360^\circ$ . The loop generator then goes to the second residue in the loop and rotates its  $\phi$  and  $\psi$  torsion angles by one step. The loop generator then goes back to the first residue and rotates its  $\phi$  and  $\psi$  angles in steps until the torsion angles have been rotated  $360^\circ$ . The above steps are continually repeated until the second residue's  $\phi$  and  $\psi$  angles have been rotated a full  $360^\circ$ . At this point the  $\phi$  and  $\psi$  torsion angles of the third residue in the loop are then started to be rotated by the loop generator. The number of conformations generated the algorithm grows exponentially as the length of the loop (insertion) grows. Therefore it is essential to try and limit the number of conformations generated.

From energy studies of dipeptides containing two amino acids of the same type (i.e. Gly - Gly) local energy minima can be seen in the Ramachandran map. For each of 19 dipeptides there are clearly 8 or 9 distinct local energy minima. The Pro-Pro dipeptide is a special case as the  $\phi$  torsion angle is fixed into one conformation and hence only 3 energy minima are seen. These local minima can be represented as discrete  $\phi$   $\psi$  angle pairs. It is these 8 or 9 distinct  $\phi$   $\psi$  angles that are tried in turn to represent stepping through the  $\phi$   $\psi$  angles a full  $360^\circ$ . The reason is because only a backbone conformation containing these low energy  $\phi$   $\psi$  angles will produce conformations of low energy. It becomes pointless generating other  $\phi$   $\psi$  angles as these will undoubtedly lead to conformations with higher energies. Therefore for every residue longer a loop is, the number of conformations generated and hence the time taken to generate them, is increased by a factor of approximately 8.

Even using the selected  $\phi$   $\psi$  torsion angle pairs, the conformation loop generator is a very computationally intense procedure. Any way to further cut down the number of conformations generated would greatly speed up the time taken to work out the best conformation. Another method used to speed up the calculation is to check for 'suicide conformations'. A suicide conformation is one where at a certain point there are not enough amino acid residues left to be added to the loop for it to be able to reach across the gap to the other side of the insertion point. The conformation sequence is terminated when

the chain is no longer long enough to reach to the other side of the gap. This happens when the loop is growing in the wrong direction and there are too few residues left in the chain to define their  $\phi$   $\psi$  angles such that the loop can rejoin the polypeptide backbone. This saves a lot of computational time by not having to define the remaining  $\phi$   $\psi$  torsion angles.

The conformations produced by the loop generator at this stage are still too numerous to be processed manually. Even with the suicide sequences taken out, filters are required to be added to the loop conformation generator to remove as many loops as possible automatically. The order the filters are applied to remove generated conformations is important due to the large numbers of conformations generated by the routine. Filters that are quick to compute are applied before filters that are more computationally intensive. This allows as many conformations as possible to be removed as quickly and as computationally inexpensively as possible so that the slower more computationally intensive filters are called on less often.

It is even better if a quick filter can be found which removes the majority of the conformations generated. This is the case for the suicide sequence filter as it reduces over 99% of the loops generated, or is possible to generate. It is also relatively computationally quick as only the distance from the C' atom of the end of the growing chain to the first N<sub>amide</sub> atom of the first amino acid residue after the insertion needs to be calculated. The maximum remaining length of the loop still to be modelled is the number of residues left to be modelled in the chain times 3.9Å (the length of an amino acid in the extended conformation). If the length to close the gap is larger than the length of the remaining chain then the conformation is terminated.

Other computationally quick filters are also used to further reduce the number of loop conformations to manageable levels:

- Checking that the final amide bond angle (C'—N<sub>amide</sub>—C<sub>α</sub>) is of a reasonable geometry.
- The peptide bond ( $\Omega$  torsion angle) is near planar
- Final  $\psi$  torsion angle is a reasonable geometry close to a minimum in the Ramachandran map.

These three filters remove the vast majority of the remaining loop conformations generated.

More time consuming filters can now be applied as the numbers of conformations that make it this far are of a more manageable amount. The total number of short non-bonded interactions gives an indication of the amount of steric crowding caused by a particular conformation. If the number of non - bonded interactions is too high then there is too much steric crowding and the local energy minima for this conformation will be significantly higher than other conformations even after carrying out energy minimisations on the segment. Therefore by rejecting loop conformations with a large number of short non - bonded interactions reduces the number of conformations further.

Counting the non-bonded interactions is a computationally expensive task, especially for proteins as the calculation has to be carried out over every atom in the protein for each atom in the loop. It is therefore one of the last filters to use so that as few loop conformations as possible make it through to this filter.

The other time consuming test is the number of sphere violations a given conformation has. This is how many atoms in this particular conformation of the loop are closer than a predetermined distance from the centre of the protein. As all large insertions occur on the surface of the protein they extend into the surrounding solvent or lie flat on the surface of the protein. They should not extend down into the core of the protein. This filter removes conformations that bury the loop into the core of the structure being modelled.

Using all these filters reduces the number of conformations that have to be manually checked to manageable numbers. At this stage to select the best conformation for the insertion has to be done by inserting each conformation of the insertion into the polypeptide model one at a time and an energy minimisation carried out on the model to find which loop has the gives the lowest energy conformation. As before at this stage it would take too long to carry out an energy minimisation on the entire protein model. Instead a section of the model around the insertion is taken and held rigid while the insertion and four residues either side of the insertion are allowed to move in the energy minimiser.



## 5.7 Addition of Hydrogens

Once the amino acid substitution, deletions and insertions are completed it is only energy minimisation that is left to be carried out on the model. Before energy minimisation is carried out hydrogens are added to the model. Throughout the previous modelling stages the hydrogen atoms were not included in the model so that the modelling could be carried out quicker. The effects hydrogen atoms have on the structure of a protein can be incorporated into the calculations carried out on the model up to now. The main effect hydrogen atoms have on the protein structure is the hydrogen bond that acts to stabilise certain backbone conformation, mainly the  $\alpha$  helix and  $\beta$  sheet.

The stabilising effect of the hydrogen bond is small and can easily be swamped out by other forces such as atom clashes, bond angle deformation, bond stretches and the stronger electrostatic effects. It is therefore not practical to add hydrogens until these stronger forces are reduced and the effect of hydrogen bonding can make a contribution to the conformation of the model. Also including hydrogen atoms dramatically increases the number of atoms in the model which means it will require more memory and slow down the calculations as many more atom - atom interactions have to be calculated. With all these disadvantages it is only at the later stages of building a polypeptide model that adding hydrogens is a reasonable option.

The addition of hydrogens is a quick procedure that is automatically carried out by the modelling package COMMET. Hydrogen atoms are added to a heavy atom such that it will cause the minimum amount of steric clash with other atoms in the model

## 5.8 Final Energy Minimisation

The final minimisation stage is carried out in discrete steps. This is to minimise the deviation from the backbone structure of the original model as much as possible. The steric strain in the new model is the overwhelming energy at this stage in the modelling. If the entire model were allowed to move freely during the first steps of the energy minimisation then the atomic overlap between the atoms of the model could easily be such that the model becomes unstable during the minimisation and comes apart. This can easily happen if there is a particularly bad overlap in the position of two atoms. What occurs during the

minimisation of a highly strained section of the model is that the atoms involved in the clash can move too great a distance in one step causing unrealistic conformations to be adopted in that section of the model. It is a much better consideration to slowly release this excessive steric energy.

The method employed to do this is to keep all but the newly added hydrogen atoms in the model to be held fixed in position. This allows the hydrogen atoms to quickly settle down in the model and to remove any bad steric contacts that may arise between a heavy atom and a hydrogen atom when the hydrogen atoms were added to the model. Energy minimiser routines have options allowing the user to fix in position certain types, groups or ranges of atoms in a model while allowing the remaining atoms to be minimised. The atoms allowed to move, in this case the hydrogen atoms, still feel the complete effect of the rigid atoms so are minimised while taking the positions of the fixed atoms into consideration.

After the energy of the hydrogen atoms has been minimised the side chain atoms of the amino acid residues in the model are allowed to move during the next step of energy minimisation. The backbone atoms of the residues remain fixed to prevent the backbone from deviating too far from the original model. This allows any clashes between side chain atoms to be resolved without affecting the position of the backbone atoms.

While the energy minimisation is taking place a periodic check in the bond lengths and bond angles of the model is carried out to make sure the model is not stuck in a strained conformation at any part. This is easy to spot by running a simple routine that calculates the bond lengths and bond angles and displays any that are out of a specified range from the default value used in the force field. This gives a good measure of any regions in the model where there are problems with the conformation. The energy minimiser programme within the COMMET molecular modelling package allows the above information to be part of the output if it is so required. Optional parameters can be set that will print an individual energy value higher than the inputted value. Modifying the model at this time allows these high energy conformation to be rectified before the model becomes too fixed into a low energy conformation.

The strained bond length or valence angle which is causing the problem can be easily fixed by closer examination of the model around the region of the energy strain. The high energy in the bond angle or valence angle is normally down to a bad steric contact which has caused an atom into the highly strained conformation. Resetting the bad bond length or valence angle to its default setting normally highlights the reason for the highly strained conformation. By examining the surrounding atoms it is possible to rearrange just one or at the most two atoms to remove the cause of the bad conformation. Moving the offending atoms can be done manually or by using an automatic procedure within the molecular modelling package. If an entire amino acid residue is to blame for the clash then selecting this residue to go through the 'SITAR' routine within COMMET will move the residue to a more favourable conformation where it will not clash with the surrounding atoms. Once the cause of the strain has been removed then the energy minimisation can proceed.

Once the side chain atoms of the model have been through the energy minimisation routine and all serious atom clashes have been removed then it is possible to allow the entire model to move freely in the energy minimiser routine. Waiting until the serious clashes caused by the side chain atoms have been removed keeps the backbone atoms closer to the original position in the starting model. Studies have shown that the backbone atoms deviate the least in homologous proteins. The side chain atoms show much greater deviation from protein to protein in a family of homologous proteins. Carrying out the energy minimisation as described above ensures that the worst of the bad clashes have been removed and the model is in a reasonably stable conformation before the backbone atoms are allowed to move in the energy minimisation routine.

To decrease the total energy of the model further water molecules are added to the model. Many crystallographic models of proteins from the Brookhaven database contain coordinates of some water molecules. To be seen by the x-ray crystallography technique the water molecules have to remain in the same position for a significant amount of the time. These waters are often buried in the protein structure or are important for stabilising a certain conformation of the protein at the surface. Adding these waters to the target model at the same relative position as they are found in the parent structure can stabilise the conformation of the target model.

## 5.9 Use of the Transputers

The processing speed of computers has increased dramatically since the 1950's. The increase in the speed of computer arithmetic has been roughly tenfold every five years. From the beginning in the 1950's to 1975 the performance of computers had increased by a factor of nearly 100,000. Today it is the normal to have access to computers with speeds in the order of gigaflops ( $10^9$  floating point operations per second).

### 5.9.1 The Need For Computing Power

No matter how much computing power is available it is not long before scientists have devised problems that exceed the current ability of even the most powerful computers. Quite often a problem will have been simplified so that it will be able to run on the current computer. Therefore when more powerful computers come along the first thing that is done is to add more complexity to the initial problem so that it better models the real life process. A good example of this is the current state of protein modelling and its need of the most powerful computers possible. Initially only small molecules of a few tens of atoms could be represented in a simple manner on the screen (stick representation) and it took a considerable time to calculate the minimum energy of the molecule. As computers became faster and more powerful the complexity of the molecules that could be modelled was increased. Molecular mechanic calculations were speeded up so that larger molecules containing hundreds of atoms could be put through the energy minimiser. Increasing the speed of the computer further meant it became possible to minimise the structure of entire proteins containing a few thousand atoms. At this point energy calculations started to become more complex in being able to have the protein in a more natural environment. With the most powerful machines today it is now possible to carry out energy minimisation calculations and molecular dynamic simulations of proteins surrounded by shells or boxes of water molecules. Even at this stage approximations are made about the water molecules and the number of atoms and molecules that can be modelled in the simulations is limited by the processing power and memory of the computer. There are still approximations used in the modelling of the protein itself. Therefore protein modelling is still limited not by the size and complexity of the simulations that can be visualised but by the speed and computational power of the modern computers.

As the power and speed of computers increased the power of its graphics capability increased. From simple unmoving ball and stick representations of small molecules the complexity of the graphics displayed on a computer terminal's screen increased until nowadays, where entire proteins can be displayed on the screen and rotated in three dimensions in real time. Graphical algorithms have been developed to show the atoms of the model as three dimensional spheres where a light source can be added to give realistic shading and a feeling of depth to the molecule. This is where the most powerful of today's computers reach their limit but the need for more powerful graphics continues. Further developments involve the manipulation of complex graphics in real time such as molecular dynamic simulations or even the simpler case of docking a small molecule into a cleft on the protein surface can still be developed further by increasing the graphics capability of the computer.

A new development has been the advent of virtual reality. This is where the user is submersed in a three dimensional world with which he can interact in real time. Molecular modelling is seen as one of the areas where virtual reality can help. It will allow the user to see around and through the molecule model; to manipulate the model and see the alterations occur in real time; to get some feed back on the alterations to the model either visually or physically through touch and pressure. One of the dreams of virtual reality in the molecular modelling field is for the user to pick up a molecule in his hand and to dock it manually to a protein. While the user is doing this there is feedback to the glove the user is wearing so that he can feel the molecule in his grasp and feel the forces as he pushes, twists and rotates the molecule to find the best docking position and conformation of the molecule with the protein.

The computing power to allow the molecular modeller to do what he wishes can not feasibly be accomplished with a single processor. Already the speed and complexity of the single processor is reaching its limits. As the performance of the single processor reaches its limit, in order to increase the processing power new approaches have to be used. One common method is to use some form of parallel processing. There are two ways parallelism has been implemented. It is possible to modify the architecture of a single processor or increase the number of processors.

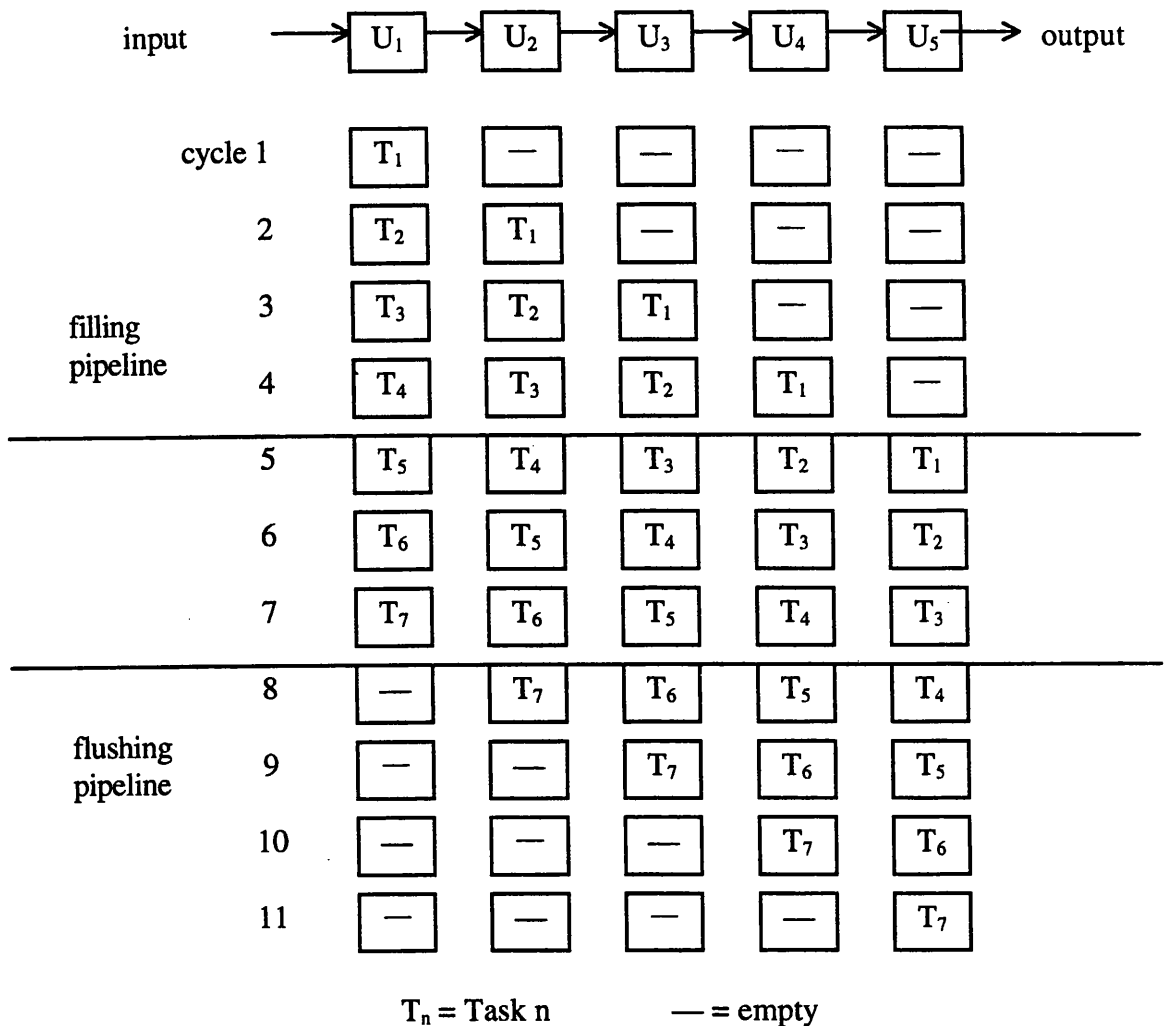
## 5.9.2 Single Processor Systems

Parallelism has existed for quite some time in the microprocessors used in the supercomputers. There are two forms of parallelism employed by modern supercomputer processors to maximise the number crunching power of the processor. The first is array processors the other vector processors

### 5.9.2.1 Array Processors

The array processor has its architecture designed to optimise the efficiency in dealing with arrays of data. At the heart of the array processor is its pipeline of functional units. Floating point multiplication and addition can be subdivided into several smaller processes. For example multiplication can be split into several steps: add the exponents of each number, multiply the mantissas, normalise the result. In an array processor each of these steps are executed independently of the other. The set up of these small processes is in a pipeline where data is put in at one end and the result comes out the other end of the pipe. The subunits of the pipeline are connected in a serial fashion and the data flows through the subunits in a serial fashion with the output of one unit being the input of the next. In the above multiplication example each subunit would correspond to each broken down step of the multiplication.

The pipeline works most efficiently when it is kept full. If each subunit in the pipeline takes one clock cycle of the CPU to process its result, then, when the pipeline has to be filled it takes the number of clock cycles of the CPU as there are subunits before the first results are produced. After this fill time a result comes out of the pipeline every clock cycle as long as data is continually being put into the pipeline. At the end of the calculation, when input stops flowing through the start of the pipeline, it takes the same number of clock cycles to flush the last of the results out of the pipeline as it did to fill the pipeline, see Diagram 5-1 on page 132 for a schematic view. Clever, well written programs designed for array processors can keep all the pipelines continuously full to achieve maximum efficiency, however the algorithms that allow this are not too common.



**Diagram 5-1: A schematic representation of a pipeline with 5 units. It shows how 7 tasks move through the pipeline, how several cycles pass before any results are used, and how the pipeline requires to be flushed to get the remaining results.**

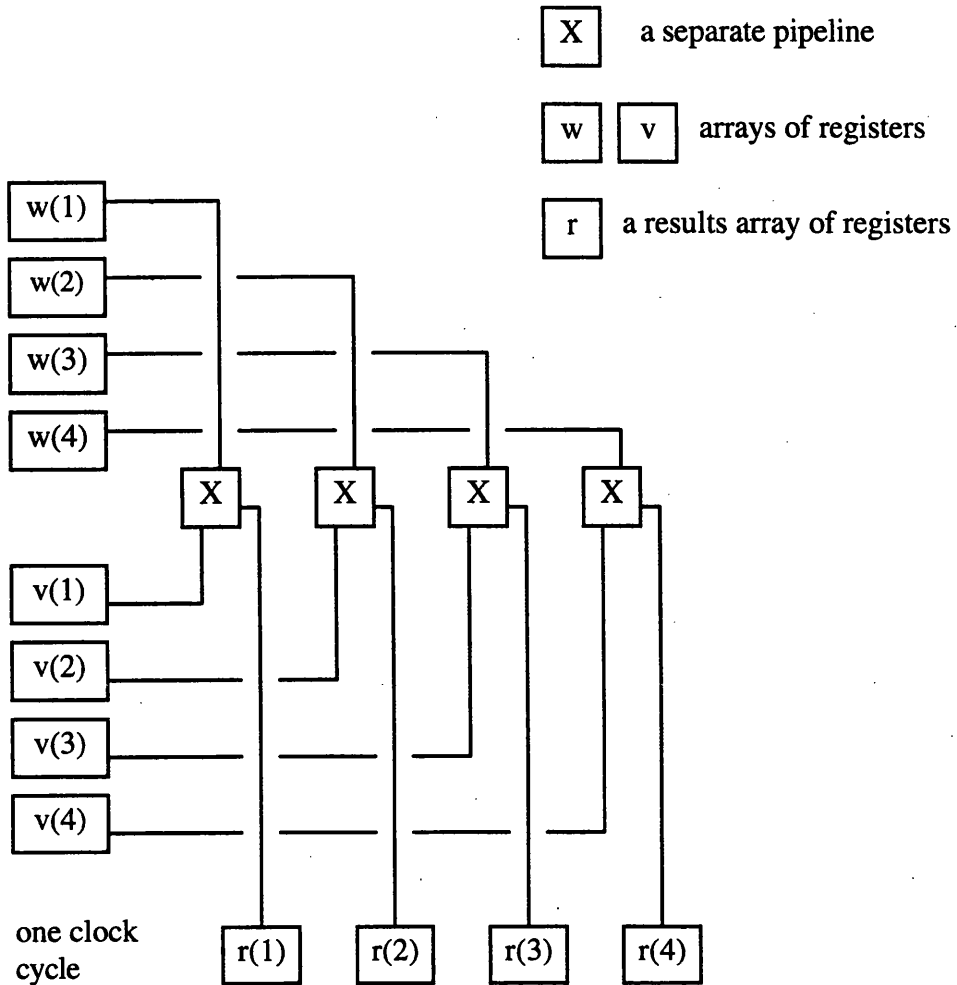
Provided that the processor is restricted to operating on large vectors or matrices i.e. with more than 1,000 elements, the array processor can be improved by lengthening the pipeline, while making each step simpler and consequently faster. The draw back is that the fill time and flush time of the pipeline is an increased number of clock cycles. This is of

little significance if the processor is always working on very large arrays of data but if on the other hand the calculation involved mainly very short vectors or small matrices then the number of clock cycles wasted on filling and flushing the pipeline can become significant. Therefore there has to be a balance in the length of the pipelines implemented on the array processors.

### **5.9.2.2 Vector Processors**

Large computers used mainly for scientific calculations tend to be vector processors. The vector processor is pipelined like the array processor but can operate on entire vectors of data in a single step. For example the Cray - 1 has eight 64 element vector registers, so that a 64 element vector can be loaded into the  $V_1$  register and another 64 element vector loaded into another register say  $V_2$ . An instruction to multiply  $V_1$  and  $V_2$  together would result in 64 multiplication operations in one clock cycle of 12.5 ns, see Diagram 5-2 on page 134. Therefore the Cray - 1 vector processor has 64 multiplication pipelines operating in parallel and offers a higher degree of parallelism than the array processor does.





**Diagram 5-2: Schematic representation of a vector process. The pipelines run in parallel so each pipeline can produce a result every clock cycle. In this example 5 results are produced per clock cycle.**

### 5.9.3 Multiprocessor Systems

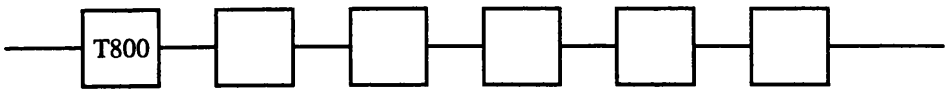
The afore mentioned array and vector processors are examples of fine grain parallelism where individual operations such as addition and multiplication are subdivided into smaller tasks. It is these smaller subtasks that run in parallel. Coarse grain parallelism is where individual instructions, subroutines of the same programme, or copies of the same programme are run simultaneously on different processors.

#### 5.9.4 The Transputer

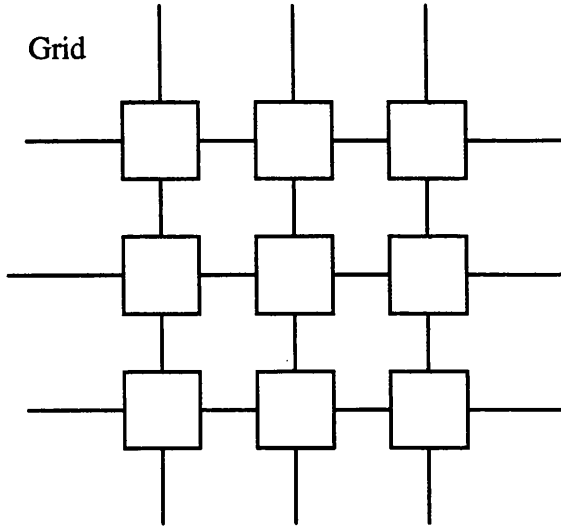
The transputer series of processors (i.e. T212, T400, T800, T9000) are specifically designed for coarse grain parallelism. They have specially designed communication links that allow fast communication between Transputers. Ideally if  $n$  processors were working on the same job then the job will finish  $n$  times quicker than when using one processor. Unfortunately this is not the case as each processor must spend some time communicating to other processors so that it knows what the other processors are doing. This communication between processors means there is an overhead which has to be paid. Because of this overhead in communications the best a parallel system can hope to be is 85-90% efficient.

The communications bottle neck can be reduced if different connection topologies are used. Different topologies are suited to certain applications. Some common topologies can be found in Diagram 5-3 on page 137.

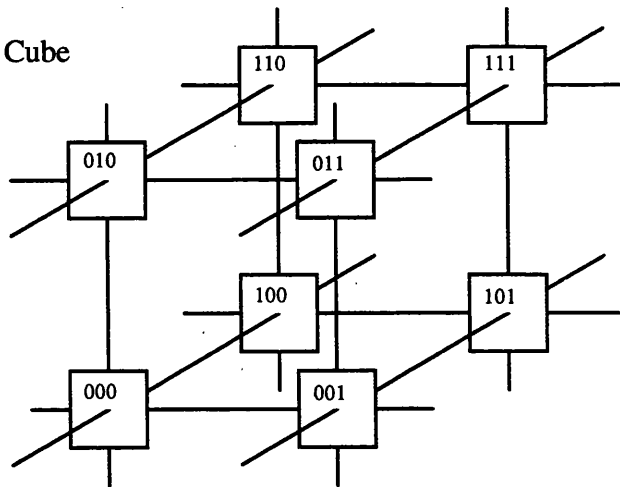
Pipeline



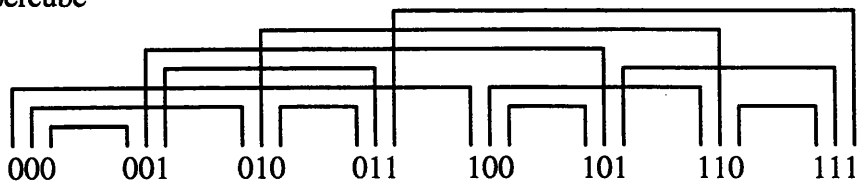
Grid



Hyper Cube



2Dimensional projection of a hypercube



**Diagram 5-3: Schematic representations of some common topologies used to connect processors in a parallel computer.**

In some systems the topology is hard wired into the machine and it can not be altered. Some more advanced systems have software programmable switches.

The molecular modelling package “COMMET” was developed to run on the T800 Transputer. The T800 has some unique performance enhancing architecture that makes it suited to scientific programming. Most scientific programmes involve a significant amount of numerical calculation to be carried out. The T800 transputer is suited to number crunching as it has a 64 bit floating point unit (FPU) integral with the central processing unit (CPU) rather than the normal situation where the FPU is a separate co-processor on a separate chip. When the FPU is a separate co-processor the CPU has to send the data to do the calculation to the FPU which takes time, and then get the result of the operation sent back to it. With the FPU on the same chip as the CPU it can operate concurrently (at the same time) with and under the control of the CPU.

Other benefits of using the T800 Transputer are that parallel code can be written and tested and debugged using just one T800 processor. The hardware links of the actual multi processor machine are simulated by the software and the computational processes’ time shared on the single processor rather than distributed across all the processors. The single transputer programme being developed behaves exactly the same as it would on the network of Transputers and can be transferred onto the network of T800 Transputers simply by changing a few lines in the configuration section of the programme. This means that the large array of Transputers can be solely reserved for work while programme development takes place on a multiple host computer or individual transputer set-up containing the software development tools.

The molecular modelling package “COMMET” runs on a single transputer that resides within a host machine. The host machine in this case is a PC compatible clone. The host machine is used only for booting up the transputer with the modelling package and as a file server. The transputer running COMMET can be connected to a network of eight or twenty-four T800 Transputers. This network of Transputers makes realistic the very long

run times of the loop conformation generator programme and allows very large models (i.e. the entire protein model) to be put through the energy minimiser routines in a reasonable time. The eight Transputers in the network have a modified tree pattern that reduces to a minimum the distance any message has to travel through the network. In the twenty four transputer partition the T800s are connected in a pipeline.

The network is used to speed up the running time of the Newton - Raphson energy minimiser and the loop conformation generator. These two routines are very computationally intensive and can have a long run time for large jobs. Using the network allows molecules with more than 1000 atoms to be run through the energy minimiser and up to ten residues to be modelled with the loop conformation generator.

## 6 Results

### 6.1 Sequence Alignment of B Fraction of Factor B (Bb) with bovine trypsin, low homology

Factor B in its zymogen form is a serum glycoprotein of approximately 90,000 Da. Activation of Factor B requires the cleavage of a single peptide bond resulting in two non-covalently linked fragments. The N - terminal Ba fragment has a molecular mass of about 30,000 Da. The C - terminal Bb fragment has a molecular mass of about 60,000 Da and is responsible for the catalytic activity. In association with C3 it forms the complex proteinase C3 convertase (C3bBb) and C5 convertase ((C3b)<sub>3</sub>Bb) (see chapter 2).

The sequence of Bb was first obtained by protein chemistry<sup>106,107</sup>. The sequence of Ba was derived mainly from the nucleic acid sequence of a cDNA clone<sup>107</sup>, except for the short amino and carboxyl terminal sequences of Ba that were obtained by protein chemistry<sup>106</sup>. Nucleic acid sequencing of the gene coding for factor B and is in good agreement with the sequence obtained by protein chemistry. The only exception is that the DNA sequence indicates the presence of an isoleucine residue at position 272 instead of a threonine residue.

The complete amino acid sequence of Factor B consists of 739 residues. Activation by Factor D results in the cleavage of the peptide bond between Arg 234 and Lys 235 forming Ba which is 234 amino acid residues long and the catalytic chain Bb consisting of 505 amino acid residues.

Partial sequence studies of Bb had previously shown factor B to be an unusual serine protease<sup>108</sup>. This serine protease domain is located in the C - terminal domain Bb. It comprises residues 457 - 739. The serine proteases is a large family of proteins that contains members with diverse functionality. Carrying out sequence alignments of Bb with members of the serine protease family with known crystal structure showed that bovine trypsin gave the best alignment.

C2, the equivalent protein in the classical pathway as FB in the alternative pathway, is highly homologous with FB. This means that C2 and FB recently in evolution terms share a common ancestor that recently duplicated itself to give rise to the two different genes coding for C2 and FB<sup>109</sup>. Their similar roles and function in the two pathways also highlight the fact that there is probably little difference in the structures between C2 and FB. Most of the differences will be centred on the different specificity towards their differing methods of activation and control.

Although bovine trypsin gave the best sequence alignment with Bb the alignment is poor. Both FB and C2 are only distantly related to the serine protease family of proteins. It is normally considered that a protein is homologous to another protein only if the sequence alignment shows that their primary sequences are above 25% identical to one another. For the case of FB and bovine trypsin the two proteins have a score of 17 - 21% identity. This very low score means that there will be some major deviations in the structures between the two proteins. This can be seen in the large insertions that are to be found in the alignment between FB and bovine trypsin (Diagram 6-1 on page 147).

	10
C2	A P S C P Q N V N I S - G G
FB	T P W S L A R P Q G S C S L E G V E I K G G
	20
	30
C2	T F T L S H G - W A P G S L L T Y S C P Q G
FB	S F R - - - - L L Q E G Q A L E Y V C P S G
	40
	50
C2	L Y P S P A - S R L C K S S G Q W Q T P G A
FB	F Y P Y P V Q T R T C R S T G S W S T L K T

60

70

C2 T R S L S - - K A V C K P V R C P A P V S F  
FB Q D Q K T V R K A E C R A I H C P R P H D F

80

90

C2 E N G I Y T P R L G S Y P V G G N V S F E C  
FB E N G E Y W P R S P Y Y N S V D E I S F H C

100

110

C2 E D G F I L R G S P V R Q C R P N G M W D G  
FB Y D G Y T L R G S A N R T C Q V N G R W S G

120

130

140

C2 E T A V C D N G A G H C P N P G I S L G A V  
FB Q T A I C D N G A G Y C S N P G I P I G T R

150

160

C2 R T G F R F G H G D K V R Y R C S S N L V L  
FB K V G S Q Y R L E D S V T Y H C S R G L T L

170

180

C2 T G S S E R E C Q G N G V W S G T E P I C R  
FB R G S Q R R T C Q E G G S W S G T E P S C Q



190 200

C2 Q P Y S Y D F P E D V A P A L G T S F S H M

FB D S F M Y D T P Q E V A E A F L S S L T E T

210 220

C2 L G A T N P T Q K T K - - E S L G R K I Q I

FB I E G V D A E D G H G P G E Q Q K R K I V L

C2b C2a

Ba Bb

230 240

C2 Q R S G H L N L Y L L L D C S Q S V S E N D

FB D P S G S M N I Y L V L D G S D S I G A S N

250 260 270

C2 F L I F K E S A S L M V D R I F S F E I N V

FB F T G A K K C L V N L T E K V A S Y G V K P

280 290

C2 S V A I I T F A S E P K V L M S V L N D N S

FB R Y G L V T Y A T Y P K I W V K V S E A D S

300 310

C2 R D M T E V I S S L E N A N Y K D H E N G T

FB S N A D W V T K Q L N E I N Y E D H K L K S

320

330

C2 G T N T Y A A L N S V Y L M M N N Q M R L L  
FB G T N T K K A L Q A V Y S M M S Y P D D V -

340

350

C2 G M E T M A W Q E I R H A I I L L T D G K S  
FB - - P P E G W N R T R H V I I L M T D G L M

360

370

380

C2 N M G G S P K T A V D H I R E I L N I N Q K  
FB N M G G D P I T V I D E I R D L L Y I G K D

390

C2 - - - - R N D Y L D I Y A I G V G K L D V D  
FB R K N P R E D Y L D V Y V F G V G P L - V N

400

410

420

C2 W R E L N E L G S K K D G E R H A F I L Q D  
FB Q V N I N A L A S K K D N E Q H V F K V K D

430

440

C2 T K A L H Q V F E H M L D V S K L T D T I C  
FB M E N L E D V F F Q M I D E S - Q S L S L C

450

460

C2 G V G N M S A N A S D Q E R T P W H - - - V  
 FB G M V W E H R K G T D Y H K Q P W Q A K I S  
 TB I V G G Y T C G A N T V P Y Q - - - V  
 | \* |

470

480

C2 T I K P - K S Q E T C R G A L I S D Q W V L  
 FB V I R P S K G H E S C M G A V V S F Y F V L  
 TB S L N - - S G Y H F C G G S L I N S Q W V V  
 \* | | | \* \* \* | \*

490

500

C2 T A A H C F R D G - - - - - N D H S L W R V  
 FB T A A H C F T V D D - - - - - K E H S I - K V  
 TB S A A H C Y K S G I Q V R S G Q D N I - - -  
 \* | | | | \* |

510

520

C2 N V G D P K S Q W G K E L L I E K A V I S P  
 FB S V G G E K - - - - - R D L E I E V V L F H P  
 TB N V V E G N Q Q - - - - - F I S A S K S I V H P  
 | \* \* \* \* \* | |

530

540

C2 G F D V F A K K N Q G I L E F Y G D D I A L  
 FB N Y N I N G K K E A G I P E F Y D Y D V A L  
 TB S Y N S N T L N N - - - - - - - - - - D I M L  
 | | | | | \* |

550

560

C2 L K L A Q K V K M S T H A R P I C L P C T M  
 FB I K L K N K L K Y G Q T I R P I C L P C T E  
 TB I K L K S A A S L N S R V A S I S L P T S C  
 | | | | \* \* | | | \*

570

580

C2 E A N L A L R R P Q G S T C R D M E N E L L  
 FB G T T R A L R L P P T T T C Q Q Q K E E L L  
 TB A - - - - - S A G T Q C L I S G W G N -  
 \* | |

590

600

C2 N K Q S V P A H F V A L N G S K L N - - - I  
 FB P A Q D I K A L F V S E E E K K L T R K E V  
 TB - - - - - T K S S G T S Y P D V L K C  
 | \*

610

620

C2 N L K M G V E W T S C A E V V S Q E K T M F  
 FB Y I K N G D K K G S C E R D A - Q Y A P G Y  
 TB L K A P I L S N S S C K S - - - - -  
 | | \*

630

640

650

C2 P N L T D V R E V V T D Q F L C S G T Q E -  
 FB D K V K D I S E V V T P R F L C T G G V S P  
 TB - - - - A Y P G Q I T S N M F C A G Y L E G  
 \* | | | \* \*

660670

C2 - - D E S P C K G E S G G A V F L E R R F R  
 FB Y A D P N T C R G D S G G P L I V H K R S R  
 TB G K D - - S C Q G D S G G P V V C S G K - -  
           |       \* |   | | | | | | \* \*       \*

680690

C2 F F Q V G L V S W G L Y N P C L G S A D K N  
 FB F I Q V G V I S W G V V D V C K N Q K R Q K  
 TB - - L Q G I V S W G S - - G C A Q K N K P G  
           | \* \* | | |                   | \*       \*

700710

C2 S R K R A P R S K V P P P R D F H I N L F R  
 FB Q V P A H - - - - - - - A R D F H I N L F Q  
 TB V Y T K V - - - - - - - - - - - - - - - C N  
\*

720730

C2 M Q P W L R Q H L G D - V L N F L P L  
 FB V L P W L K E K L Q D E D L G F L  
 TB Y V S W I K Q T I A S N  
           \*   | \* |       \*

Homologous Residues

1. D E K R
2. G A V
3. A V L I
4. V L I M
5. F Y W

6. S T
7. Q N
8. G P

**Diagram 6-1: Sequence Alignment Between C2, Factor B (FB) and bovine trypsin (TB). The \* mark homologous residues between FB and TB, and a | denotes identical residues between FB and TB.**

## 6.2 Substitutions

As the similarity between the two protein sequences was so weak there were numerous substitutions that had to be carried out along the primary sequence of the parent sequence (TB) as it was changed to the sequence of the target structure (Bb). The method described in Chapter 4 for conservative and non - conservative substitutions was followed. There were no problems involved in the substitutions. This was to be expected as substitutions are the least disruptive of the three modelling techniques in going from the parent to the target structure.

## 6.3 Deletions In Bb Sequence

There are only two deletion modifications that have to be made in the modelling the Bb protein from the parent model, TB. The first is a four residue deletion at residue position Asp 62\*. The deleted residues are Gln, Val, Arg and Ser. This deletion was modelled as for the short deletion method described in the previous chapter (Chapter 4). Although slightly longer than the normal three residues this method is limited to, the deletion was easily modelled by this method. There was no problem in closing such a large gap in the polypeptide chain left after the removal of the four residues using the energy minimisation routine. The extended bond closed rapidly and was of normal length within a few iterations of the energy minimiser.

The second of the deletions is a single residue deletion modelled at residue Lys 75. The residue removed from the model was Gln. There were no problems encountered in

---

\* Chymotrypsin numbering

removing the residue and running the protein fragment containing the extended peptide bond through the energy minimiser. The extended bond closed rapidly to the normal length.

### 6.3.1 Small Insertions Not Modelled Using The Conformation Generator

Insertions at Arg 34 and Val 217 are both two residues in length. These two insertions were modelled using COMMET's insertion routine. This routine inserts one amino acid into the polypeptide chain. Therefore to model a two residue insertion involved calling the insert function twice. The insert routine is adequate for modelling small insertions but can not handle the larger insertions.

### 6.3.2 Large Insertions 3 Residues In Length and Longer

There are seven insertions three residues in length or longer which required to be modelled.

They are at:

- Gln 30: 3 residues in length
- Ser 186: 3 residues in length
- Gly 129: 7 residues in length
- His 231: 8 residues in length
- Glu 101: 9 residues in length
- Leu 143: 9 residues in length
- Arg 170: 13 residues in length

A variety of methods were used to build each insertion depending on its length, its relative position in the primary structure and the local secondary structure of the parent model. Each of the large insertions are discussed separately below.

#### 6.3.2.1 Insertion at Residue Position Gln 30

Two residues either side of the 3 residue insertion were included to be modelled as part of the loop. The reasoning behind this is to open a reasonable sized gap in the polypeptide chain so that the extra backbone atoms have some room to be accommodated. This

increases the number of residues modelled in the loop conformation generator to seven. The primary sequence either side of the insertion is:

His	Lys	Gln	Pro	<b>Trp*</b>	<b>Gln</b>	<i>Ala†</i>	<i>Lys</i>	<i>Ile</i>	<b>Ser</b>	<b>Val</b>
<b>Ile</b>	<b>Ser</b>	<b>Val</b>	<b>Ile</b>	<b>Arg</b>	<b>Pro</b>					

The default setting in the loop conformation generator routine were used except for short non - bonded cutoff which was set at 10Å. The results showing the minimisation of the conformations generated by the loop conformation generator can be found in Table 6-5 on page 175

As the energy of the lowest energy conformation (conf 8) was still quite high the insertion was put through the loop conformation generator a second time. This time three residues either side of the insertion was included as part of the loop. This was done in order to increase the size of the gap in the polypeptide backbone of the model into which the insertion is to be inserted. This increased the number of residues which are modelled in the loop conformation generator to nine. Again the primary sequence either side of the insertion is:

His	Lys	Gln	<b>Pro</b>	<b>Trp</b>	<b>Gln</b>	<i>Ala</i>	<i>Lys</i>	<i>Ile</i>	<b>Ser</b>	<b>Val</b>
<b>Ile</b>	<b>Ser</b>	<b>Val</b>	<b>Ile</b>	<b>Arg</b>	<b>Pro</b>					

The default setting in the loop conformation generator routine were used except for short non-bonded cut-off which was set at 10Å. The results showing the minimisation of the conformations generated by the loop conformation generator can be found in Table 6-6 on page 178

The longer insertion did not produce a conformation for the insertion of lower or comparable energy as conformation 8 from the short insertion run. Therefore conformation 8 from the short insertion was used in the final model.

---

\* The residues shown in bold were modelled in the loop conformation generator.

† The residues shown in bold italics are the residues that have to be inserted into the model polypeptide.



### 6.3.2.2 Insertion at Residue Position Ser 186

Looking at the alignment of the serine protease family in the papers by Greer<sup>110,111</sup> it was noted that the insertion as it is positioned in the original alignment falls into a structurally conserved region. This structurally conserved region also contains the important residue at position 189 of the serine protease family. This residue is part of the substrate specificity pocket. It sits at the bottom of a pocket in the serine protease surface which is part of the binding site of the serine proteases. The residue which sits at the bottom of this pocket helps define the substrate selectivity of the particular serine protease. If this residue is charged then only oppositely charged residues are allowed into the pocket. Similarly if it is a bulky residue that is found at the bottom of the cleft then only small chain residues are able to dock into the pocket. The alignment in Greer's paper moved the insertion in the Bb sequence three residues before this important residue. It was decided to use this alignment instead. The reason being that although residue 189 is not highly conserved within the family of serine proteases it is an important residue to the function of the protease and so is unlikely to move far from this position. In my original alignment the insertion occurs just after residue 189 which would cause residue 189 to be significantly moved away from its original position. The sequence alignment was altered from:

							189							
C2	G	T	Q	E	-	-	-	D	E	S	P	C	K	G
FB	G	G	V	S	P	Y	A	D	P	N	T	C	R	G
TB	G	Y	L	E	G	G	K	D	-	-	S	C	Q	G

to:

							189							
C2	G	T	Q	-	-	Q	-	D	E	S	P	C	K	G
FB	G	G	V	S	P	Y	A	D	P	N	T	C	R	G
TB	G	Y	L	-	-	E	G	G	K	D	S	C	Q	G

conserved | | conserved region  
region

The primary sequence around insertion 186 is:

Gly Gly Val Ser Pro Tyr Ala Asp Pro Asn Thr  
Cys

The parameters used in the loop conformation generator were the default parameters as given except for:

- Short non bonded cut off = 5
- sphere radius = 10Å
- sphere violation = 10

The results showing the minimisation of the conformations generated by the loop conformation generator can be found in Table 6-7 on page 183.

Eventually conformation 27 was chosen to be added to the final model of Bb. The reason being conformation 27 was the conformation with the lowest energy after minimising the polypeptide fragment surrounding the insertion.

### 6.3.2.3 Insertion at Residue Position Gly 129

Due the length of the insertion at this position (7 residues) it became impractical to take two residues either side of the insertion to make a gap in the polypeptide chain for the insertion to fit into. Instead the polypeptide chain was rotated slightly. The idea using this method was to open up a gap in the polypeptide chain by choosing a  $\phi$  or  $\psi$  torsion angle which when rotated one or two degrees causes a large movement in the polypeptide chain. For insertion Gly 127 a torsion angle which causes a large movement in the position in the end of the gap is easy to find as the insertion is close to the end of a  $\beta$  strand. The primary sequence of the polypeptide chain either side of the insertion is:

Cys Leu Pro Cys Thr Glu Gly Thr Thr Arg Ala  
Leu Arg Leu Pro Pro Thr Thr Thr

The default settings in the loop conformation generator routine were used except for short non - bonded cut-off which was set at 10Å. The results showing the minimisation of the

conformations generated by the loop conformation generator can be found in Table 6-8 on page 188.

Two generated conformations have similar minimised energies well below any other conformation. These two conformation, conformations numbered 26 and 28, were taken for further studies.

While examining the protein it became clear that the insertion residue position 127 and the insertion at residue position 231 were very close to each other in the three dimensional structure of the model. Although the two insertions are distant in terms of the primary structure they are found at opposite sides of a cleft in the model's structure. As the two insertions are so close together there would be interactions between the residues of the two insertions. This meant that the insertions could not be modelled separately. As a result of the two insertions being so close to each other it was decided that one of the best conformations for the first modelled insertion, conformation 28 of insertion 127, should be introduced into the model while generating the conformations for insertion 231. Conformation 28 was chosen over conformation 26 in case some bias was added to the model in favour of conformation 26.

#### **6.3.2.4 Insertion at Residue Position His 231**

The insertion at residue position His 231 is eight residues in length. When the insertion is lengthened by a further two residues to start and open up some space in the polypeptide backbone, the number of residues to be modelled reaches close to the limit of the current version of the loop conformation generator. The primary sequence either side of the insertion is:

Val    Pro    Ala    His    *Ala*    *Arg*    *Asp*    *Phe*    *His*    *Ile*    *Asn*  
*Leu*    *Phe*    *Gln*    Val    Leu    Pro

The default settings in the loop conformation generator routine were used except for short non - bonded cut-off which was set at 10Å. The results showing the minimisation of the conformations generated by the loop conformation generator can be found in Table 6-9 on page 195.

Four generated conformations of insertion 231 have comparable low energies after they have been through the energy minimiser. These four conformations, conformations numbered 49, 60, 64 and 68 in Table 6-9, were tested with conformation 26 from insertion 127. The results of the energy minimisations of these four conformations of insertion 231 with conformation 26 from insertion 127 can be found in Table 6-10 on page 196.

Conformation numbered 49 of insertion 231 is clearly the lowest energy conformation for both conformation 26 and 28 of insertion 127. What has now to be decided is whether conformation number 26 or 28 is the lowest conformation for insertion 127. Using conformation 49 of insertion 231 two model fragments were built, one containing conformation 26 of insertion 127 and the other containing conformation 28 of insertion 127. The results of this minimisation can be seen in Table 6-11 on page 197.

Using conformation 28 for insertion 127 in the model to generate the conformations of insertion 231 produced one low energy conformation for insertion 231, conformation 49. Conformation 49 for insertion 231 was used to decide which of the two low energy conformations generated for insertion 127 is the better. When minimisation of the two model fragments were carried out conformation 26 for insertion 127 is clearly the lower energy conformation. Therefore conformation 49 for insertion 231 and conformation 26 for insertion 127 were chosen for the final model.

Choosing the higher energy conformation for insertion 127, that of conformation 28, while modelling insertion 231 appears to have cleared the ambiguity between which conformation to choose, either conformation 26 or 28. Both conformations are of comparable low energy at the start of the modelling. Choosing conformation 28, the conformation with the slightly higher energy, shows that even when biasing the generation of the conformations of insertion 231 in its favour, conformation 26 still eventually produces the conformation with the lower energy. Had conformation 26 for insertion 127 been chosen when modelling insertion 231 and the same result produced then it could easily be argued that choosing conformation 26 for insertion 127 biased the results in favour of insertion 26.

### 6.3.2.5 Insertion at Residue Position Glu 101

There are two important points about the positioning of insertion 101. The first is that it is immediately before residue Asp 102 which is part of the catalytic triad. Factor B is still an active serine protease and contains the three functional amino acids required for catalysing the breaking of the peptide bond (the other two residues are His 57 and Ser 195). These three residues have to be in a very particular spatial position for the catalysis to take place. Within the family of serine proteases, for members that still actively cleave peptide bonds and have known 3D structures these three residues are in almost identical position relative to each other. The spatial positioning of these three residues is held constant within the serine protease family by the three residues being within regions with highly conserved primary sequence. Having a large insertion next to Asp 102 in the alignment of Fb and TB is going to cause the Aspartic acid residue to move. It is therefore very likely that the sequences are improperly aligned at this region.

The second point also supports the fact that the sequences are misaligned in this region. On examining the model structure it is found that the insertion in question occurs in a strong secondary structure feature. The insertion occurs in a  $\beta$  strand. In the original alignment the insertion occurs five residues from the  $\beta$  turn. As insertions, particularly large insertions tend not to occur in secondary structure it is highly likely that the insertion should be aligned to the  $\beta$  turn rather than within the  $\beta$  strand structure itself.

The insertion to be modelled is nine residues in length. Two residues either side of the insertion are added to the residues to be modelled to create a gap in the polypeptide backbone to accommodate the insertion. This increases the number of residues in the insertion to more than the current version of the loop conformation generator can handle in a feasible time period. Therefore another method had to be employed to model this insertion.

The original alignment of Bb and trypsin just before Asp 102 is shown below. In the original alignment the residues involved in the  $\beta$  turn are underlined and Asp 102 is highlighted in bold:

```

FB   N Y N I N G K K E A G I P E F Y D Y D V A L
TB   S Y N S N T L N N - - - - - D I M L

```

#### 6.3.2.5.1 Attempt 1

The insertion was split in two and put into the model either side of the i+1 and i+2 residues of the  $\beta$  turn of the original model. This means the alignment between Bb and trypsin is:

```

FB   N Y N I N G K K E A G I P E F Y D Y D V A L
TB   S Y N S - - - - - N T - - - - - L N N D I M L

```

The loop conformation generator could not be used to model the insertions therefore each residue was inserted into the polypeptide chain one at a time using the 'Residue Insert' function within COMMET. This function places the residue to insert directly on top of the two residues either side insertion. It then, for each of the nine low energy  $\phi \psi$  torsion angle combinations from the Ramachandran map of the inserted residue, carries out a rapid pattern search minimisation with a high potential set forcing the  $\phi \psi$  torsion angles to the appropriate angles. Out of the nine different  $\phi \psi$  angle combinations tried the conformation which produces the lowest energy is chosen as the conformation for the inserted residue. After each residue is inserted using the "Residue Insert" function the polypeptide fragment around the insertion is put through the energy minimiser before the next residue is inserted into the polypeptide chain. The results can be found in Table 6-12 on page 199.

#### 6.3.2.5.2 Attempt 2

Again the insertion was split into two and put into the model either side of the i+1 and i+2 residues of the  $\beta$  turn of the original model. The new alignment is the same as for attempt 1 but the order in which the residues are inserted into the model is different. The results of the modelling can be found in Table 6-13 on page 201.

#### 6.3.2.5.3 Attempt 3

In this attempt the insertion was kept the original size and added one residue at a time between i+2 and i+3 residues of the  $\beta$  turn. The order the residues of the insertion were

inserted into the polypeptide chain was to keep adding the residues to the end of the  $\beta$  sheet. The alignment between Bb and trypsin for this attempt at modelling the insertion is:

```
FB    N Y N I N G K K E A G I P E F Y D Y D V A L
TB    S Y N S N - - - - - - - - T L N N D I M L
```

The results can be found in Table 6-14 on page 203.

#### 6.3.2.5.4 Attempt 4

The same alignment used in attempt 3 above was tried again. This time correcting the substitutions for the new alignment was carried out first. Then the residues were inserted into the polypeptide chain in the same position and same order as attempt 3. The results can be found in Table 6-15 on page 205.

Attempt 2 gives the most compact structure of all the attempts and has the lowest conformational energy. This conformation was put into the final model.

#### 6.3.2.6 Insertion at Residue Position Leu 143

The insertion at Leu 143 occurs in a loop region of the model. Again the insertion is too large to use in the loop conformation generator. Instead each of the nine residues in the insertion is put into the model and the polypeptide chain about the insertion is put through the energy minimiser. This is repeated until all nine residues of the insertion are inserted into the model. The results showing the minimisation of the conformations generated by the loop conformation generator can be found in Table 6-16 on page 206.

After Gln, the fourth residue of the insertion, was added to the model and minimised the following residues inserted caused the energy to increase sharply. The model was saved after Gln was inserted and minimised to be the starting point of further modelling. To try and keep the energy of the model low hydrogen atoms were added to the model and to the inserted residues to see if this would make a difference in the energy of the insertion. The following table shows the result of taking this strategy. Building the insertion with

hydrogen atoms already added to the model prevents the energy of the insertion from escalating. The results of this modelling strategy can be seen in Table 6-17 on page 207

This method produces a fairly compact structure with an acceptable energy. Further energy minimisation was carried out on this structure with the hydrogen atoms removed (see Table 6-18 on page 207).

Looking at the insertion more closely at this point showed that the aromatic ring of phenylalanine at residue position 144 had distorted from planar. The torsion and valence angles of the distorted aromatic ring were reset to their standard planar value and the energy minimisation restarted. The results of this modification can be seen in Table 6-19 on page 208. Again after several iteration of the energy minimisation steps the aromatic ring of the phenylalanine at position 144 was distorted. Once again the ring's torsion and valence angles were reset to the standard planar angles. The results can be followed in Table 6-20 on page 210.

Further energy minimisation and alterations to Phe 144 and Glu 142 reduces the energy of the fragment model to 221.0 kcalmol<sup>-1</sup>. In the model Phe 144 and Glu 142 point into the same cleft in the model surface. Once they are in this conformation, pointing into the cleft the energy of the fragment is low and the two residues stable.

#### **6.3.2.7 Insertion at Residue Position Arg 170**

In the original alignment of Fb with TB the insertion at residue 170 occurs within an  $\alpha$  helix structure and two residues away from a disulphide bridge found at Cys 168. Due to the disulphide bridge it is very unlikely that there will be much change in the structure of the proteins prior to the disulphide bridge. Also as this segment is in an  $\alpha$  helix it is in a stable conformation and is unlikely to change its conformation. Therefore it is reasonable to expect to find the  $\alpha$  helix up to at least residue 168 which is the Cys residue of the disulphide bridge. There are several possibilities that can occur after residue 168 in the model:



1. The Helix continues for one more residue, up to residue 169. After which there is a random coil due to the insertion at this point.
2. The  $\alpha$  helix is kept the same length in the Bb model as found in TB, extending up to residue 171. The insertion is moved to be just after the end of the helix.
3. The helix is extended for at least part of the insertion.

To examine the possibilities mentioned above a simple sequence analysis of this region of the Bb sequence was carried out manually. Chou and Fasman secondary structure prediction<sup>112</sup> was used as this algorithm can be carried out manually. Looking at the segment of the Bb primary sequence immediately before the insertion and the insertion itself the following probability figures for the residues being in the  $\alpha$  helix conformation are from the Chou Fasman algorithm:

Lys	Gly	Ser	Cys	Glu	Arg	Asp	Ala	Gln	Tyr	Ala	Pro
$I_{\alpha}$	$B_{\alpha}$	$i_{\alpha}$	$i_{\alpha}$	$H_{\alpha}$	$I_{\alpha}$	$I_{\alpha}$	$H_{\alpha}$	$h_{\alpha}$	$b_{\alpha}$	$H_{\alpha}$	$B_{\alpha}$
1.08	0.53	0.79	0.77	1.33	0.79	0.98	1.45	1.17	0.61	1.45	0.59

Gly	Tyr	Asp	Lys	Val	Lys	Asp	Ile	Ser	Glu	Val	Val
$B_{\alpha}$	$b_{\alpha}$	$I_{\alpha}$	$I_{\alpha}$	$I_{\alpha}$	$I_{\alpha}$	$I_{\alpha}$	$I_{\alpha}$	$i_{\alpha}$			
0.53	0.61	0.98	1.07	1.09	1.06	1.02	0.92	0.79			

Using the rules developed by Chou and Fasman the  $\alpha$  helix has a high probability of extending five residues into the insertion up to the Pro residue. As Pro is a strong  $\alpha$  helix breaker and tends not to occur at the C terminal of an  $\alpha$  helix it is very unlikely the  $\alpha$  helix would extend past the Pro residue.

As the insertion is too long to model using the loop conformation generator the number of residues to be modelled by the loop conformation generator was reduced by modelling the first five residues in the insertion in the  $\alpha$  helix conformation. This left seven residues of the insertion to be modelled. These residues along with three residues immediately after the insertion in the primary sequence were modelled using the loop conformation generator.

The three residues added to the loop to be modelled by the loop conformation generator were included to create a space in the model to add the insertion. The primary sequence around the insertion to be modelled looked like:

Lys	Gly	Ser	Cys	Glu	Arg	<u>Asp<sup>*</sup></u>	<u>Ala</u>	<u>Gln</u>	<u>Tyr</u>	<u>Ala</u>	<i>Pro</i> <sup>†</sup>
<i>Gly</i>	<i>Tyr</i>	<i>Asp</i>	<i>Lys</i>	<i>Val</i>	<i>Lys</i>	<b>Asp<sup>‡</sup></b>	<b>Ile</b>	<b>Ser</b>	Glu	Val	Val

The first part of modelling this insertion was building the  $\alpha$  helix. First a five residue helix was built using the COMMET's build function. This function allows the user to build a short chain of residues of a given primary sequence and then define the  $\phi$  and  $\psi$  torsion angles of the model. After the five residue  $\alpha$  helix was built it was manually added to the protein model at residue position 170. This was accomplished by joining the two models at the C<sub>carbonyl</sub> atom of residue 170 and the N<sub>amide</sub> terminal atom of the short  $\alpha$  helix fragment. This newly formed bond was shortened to the correct length and the torsion angle spanning the new bond set the correct value for being part of an  $\alpha$  helix.

Three residues before the insertion and the newly inserted residues in the  $\alpha$  helix conformation were allowed to move in the energy minimiser. The C<sub>carbonyl</sub> atom of the last residue in the  $\alpha$  helix was not bonded to any other residue during this procedure. To prevent the helix from unwinding this C<sub>carbonyl</sub> atom was held fixed while the model was put through the energy minimiser. The results of this minimisation can be seen in Table 6-21 on page 211.

Looking at the  $\phi$   $\psi$  torsion angles of the  $\alpha$  helix at this stage the angles do not drift far from the idealised torsion angles at the start of the energy minimisation and are still very much in the allowed region of the Ramachandran map for  $\alpha$  helix  $\phi$   $\psi$  torsion angles.

---

\* The residues in bold and underlined were part of the insertion but were modelled as part of the  $\alpha$  helix.

† The residues shown in bold italics were the part of the insertion modelled using the loop conformation generator.

‡ The residues in plain bold were modelled in the loop conformation generator along with the residues from the insertion.

To model the remaining residues of the insertion the loop conformation generator was used. The default settings in the routine were used except for short non-bonded cutoff which was set at 10Å. The results of minimising the conformations generated by the loop conformation generator can be found in Table 6-22 on page 220.

Eventually conformation 96 was chosen to be added to the final model of Bb. The reason being conformation 96 was the conformation with the lowest energy after minimising the energy of the polypeptide fragment surrounding the insertion.

## 6.4 Global Minimisation

The final model is built by putting all the modifications into the one model. This gives a model with the correct primary sequence for Bb but due to the many alterations carried out on the model has regions of high steric energy. Next the disulphide bridges were remade. Then new charges for the inserted residues were calculated. This was done using the 'Delre Charge' function within COMMET. This function calculates the charge on each atom in a residue. The charge calculation depends on the local environment of the residue which is most closely affected by the residue before and after it in the primary sequence. The residues in the model that were not altered or inserted kept the original charges from the bovine trypsin model. The model is now ready to be put through the energy minimiser.

At the start of the energy minimisation procedure the model is put through 10 iterations of the energy minimiser before being checked for serious steric clashes that the minimiser can not resolve. This problem is most easily noticed by watching the bond lengths for any dramatic deviations from the normal. After each set of ten iterations of the energy minimiser the bonds in the model was analysed for large deviations. The first 60 iterations of the energy minimiser are shown below:

All energy in kcalmol <sup>-1</sup>						
	10×	10×	10×	10×	10×	10×
	→	→	→	→	→	→
*****	*****	25010	17310	15950	1570	1500

At this point the backbone breaks between residues Ile 118 and Arg 119 due to residue Tyr 114. The aromatic group of the Tyr has bad steric contacts with the O<sub>carbonyl</sub> of Ile. Going back to the original model the COMMET routine 'SITAR' was used on residues Tyr 114 and Tyr 101. Tyr 101 was put through SITAR as this was also causing steric clashes in the region. This slightly altered model was now used as the starting position for the energy minimisation. The energy minimisations were started again. The first 70 iterations of the energy minimiser are shown below:

All energy in kcalsmol <sup>-1</sup>							
	10× →	10× →	10× →	10× →	10× →	10× →	10× →
*****	12610	9310	8007	7476	7172	7007	6866

At the end of this energy minimisation procedure the residue Ile 118 again had a very high energy bond. The residue was clashing with Lys 79. Using the 'SITAR' function on Lys 79 moves it into such a position that it clashes with many other residues. Using 'SITAR' on Ile 118 is a better option producing less clashes. The energy minimisation of the model continues after the modifications to Ile 118 is carried out. The next set of iterations of the energy minimiser give the energies for the model:

All energy in kcalsmol <sup>-1</sup>			
	10× →	10× →	10× →
1140	4906	4070	3864

Carrying out an analysis of the bond lengths in the model shows that the phenyl ring in residue Phe 49 becomes greatly distorted and the bond lengths grow large. Phe 49 was clashing with residue Lys 111. The aromatic ring of Phe 49 is set to the correct geometry with the correct bond lengths, after which the 'SITAR' routine is carried out on it. The

energy minimisation carried on after the above alterations were made to the model. The next 80 iterations of the energy minimiser are:

All energy in kcalmol <sup>-1</sup>								
	10×	10×	10×	10×	10×	10×	100×	100×
	→	→	→	→	→	→	→	→
2535	2320	2154	2067	1946	1853	1769	1216	992.3

At this stage the hydrogen atoms were added to the model. The addition of hydrogen atoms was carried out in two stages. First the polar hydrogens were added to the model and then the remaining hydrogens added to the model. The results of energy minimising the model after the addition of just the polar hydrogens is:

All energy in kcalmol <sup>-1</sup>						
	10×	10×	10×	10×	500×	500×
	→	→	→	→	→	→
7844	7457	7358	7232	7181	6149	5799

After the addition of the aliphatic hydrogens the model was again minimised:

All energy in kcalmol <sup>-1</sup>				
	10×	10×	500×	500×
	→	→	→	→
15940	8867	8016	6198	5992

Any improvements in the energy of the protein model now come from trying to mimic the environment the protein would find itself. Looking at the crystal structure of the original structure bovine trypsin there are some solvent molecules, in this case water, that are strongly bound to the protein structure. These water molecules help stabilise the structure of the protein. As the new model, Bb is homologous to bovine trypsin, a starting point for adding important waters to the new model is taking the strongly bound water molecules from the original crystal structure and putting them into the new model in the same relative

position. This is achieved by first superimposing the TB crystal structure onto the new Bb model. The residues used to superimpose the two models are chosen as those least likely to be in different positions in the two models. These residues are the six residues involved in the three disulphide bridges common to both proteins and the three residues involved in the catalytic triad of the serine protease family. Using these nine residues TB was superimposed onto Bb. The water molecules associated with TB were also moved as part of the TB model.

The TB residues are removed from superimposed molecules leaving behind Bb and the crystallographic waters. To continue with the energy minimisation of the model the Bb residues were held fixed and the water molecules allowed to move in the energy minimiser procedure. The results of this are:

All energy in kcalmol <sup>-1</sup>			Hydrogen atoms fixed			
	10× →	10× →		10× →	10× →	10× →
*****	454200	453800	*****	31950	31360	31220

At this point the entire model is allowed to move freely in the energy minimiser routine. The results of carrying out the energy minimisation routine on the model are:

All energy in kcalmol <sup>-1</sup>							
	10× →	10× →	500× →	500× →	1000× →	1000× →	1000× →
67280	10220	8561	5817	5691	5541	5433	5355

At this point in the model building process the model is examined closely for any possible errors. The most easily noticed one is that residue His 57, which is part of the catalytic triad, is in the wrong orientation for an active serine protease. As previously mentioned, the residues involved in the catalytic reaction in the serine protease family are highly conserved and found to be in the same spatial position for the members of the family that have had their crystal structure determined. As Bb is an active serine protease this very strongly

suggests that the three residues central to the catalytic reaction are in the same spatial position in Bb as other members of the family that already have their crystal structure determined. As His 57 was in the wrong orientation it was decided that this part of the model required to be remodelled.

Starting from the model produced by the last energy minimisation the water molecules and hydrogens were removed from the model.

The alignment between Bb and BT around this region is

TB	Val <sup>53</sup>	Ser	Ala	Ala	His <sup>57</sup>	Cys	Tyr	Lys	Ser	Gly	Ile <sup>63</sup>
Bb	Leu	Thr	Ala	Ala	His	Cys	Phe	Thr	Val	Asp	Asp

TB	Gln	Val	Arg	Leu	Gly <sup>69</sup>	Glu	Asp	Asn	Ile <sup>73</sup>	-	-
Bb	-	-	-	-	Lys	Glu	His	Ser	Ile	-	Lys

TB	-	Asn	Val	Val	Glu	Gly	Asn	Gln	Gln <sup>81</sup>	-	Phe
Bb	Val	Ser	Val	Gly	Gly	Glu	Lys	-	-	Arg	Asp

TB	Ile	Ser	Ala	Ser <sup>84</sup>
Bb	Leu	Glu	Ile	Glu

The above region, that is residues 53 Leu to 84 Ser was cut out of the crystal structure model of BT and inserted into the equivalent region of Bb. The residues previously in the model of Bb in this region were discarded. The appropriate residues were substituted so that the new section of polypeptide chain added to Bb was of the correct primary sequence. Next the two residue insertion 73 Ile was added to the model using COMMET's protein build function.

After the substitutions and the two residue insertion is remodelled in the Bb model the four residue deletion at Asp 63 is modelled. The modelling is carried out taking residues either side of the deletion and modelling this section of the peptide chain with the loop conformation generator. The primary sequence of Bb either side of the residues to be modelled with the loop conformation generator is:

Ala His Cys Phe Thr Val Asp Asp Lys Glu His Ser  
Ile Lys Val

The default setting in the loop conformation generator routine were used except for the short non-bonded cut-off which was set to 10Å. The results of energy minimisation on the conformations produced by the loop conformation generator can be found in Table 6-23 on page 222.

Conformation numbered 1 has the lowest conformation by a considerable margin and was inserted into the model.

With the new conformation for the region inserted into the model the global minimisation of the Bb model can be resumed. First the entire model is put through the energy minimiser

All energy in kcalmol <sup>-1</sup>				
	10×	10×	10×	100×
	→	→	→	→
6387000	5628	2078	1563	706.9

After the above energy minimisation was completed the hydrogen atoms were added to the model. This was carried out in two stages. First the polar hydrogens were added the resulting structure was put through the energy minimiser:

All energy in kcalmol <sup>-1</sup>				
	10×	10×	100×	500×
	→	→	→	→
5209	5013	4919	4875	4497



The second step is the addition of the remaining hydrophilic hydrogens to the structure. This drastically increases the number of atoms in the resulting model. Putting this model through the energy minimiser gives the following energy results:

All energy in kcalmol <sup>-1</sup>			
	10× →	500× →	500× →
7168	6304	4779	4589

Next the oxygen atoms from the crystallographic waters were added to the model. The oxygen atoms were added to this model in the same position as they were in the previous partially minimised model of Bb before the alterations to the polypeptide chain took place. These positions originally came from the waters closely bound to bovine trypsin as found in its crystal structure. The results of carrying out the minimisation of this model are:

All energy in kcalmol <sup>-1</sup>				
	10× →	10× →	10× →	100× →
*****	22650	9494	9271	9212

At this point the hydrogen atoms of the water molecules were added to the model. The entire Bb model was fixed and only the newly inserted hydrogen atoms of the water molecules were allowed to move while the entire model is run through the energy minimiser. The results are:

All energy in kcalmol <sup>-1</sup>			
	10× →	10× →	10× →
379500	11670	10830	1006

Then just the water molecules of the model were allowed to move freely in the energy minimiser routine while the polypeptide chain was kept in a fixed conformation:

All energy in kcalmol <sup>-1</sup>					
	10×	10×	10×	10×	100×
	→	→	→	→	→
10880	9641	9566	9527	9513	9477

Finally the entire model was allowed to freely move during the energy minimisation runs. The resultant energies are:

All energy in kcalmol <sup>-1</sup>							
	10×	100×	500×	500×	1000×	1000×	1000×
	→	→	→	→	→	→	→
2368	7396	4815	4442	4339	4202	4118	4115

## 6.5 Refinements To The Model

After global minimisation was carried out the model of Bb was evaluated by putting it through a structure analysis package. The structural analysis package used was the one found in the WhatIf package (see section 7.1.1 on page 223). This showed there were numerous serious errors in the model which had to be fixed. The errors in the structure which were tackled in the refinement process were:

1. The wrong chirality of certain amino acids
2. The wrong chirality of certain Ile C<sub>β</sub> atoms
3. The surface loop at Insertion 170 extends too far out of the surface

As the refinement modelling was carried out several years after the initial project the original software and hardware used in the rest of the project was not available. Instead the molecular modelling package InsightII<sup>®</sup> version 2.3.0 and the molecular simulations package Discover<sup>®</sup> from Biosym Technologies<sup>113,114</sup> was used.

There were a total of 35 amino acids which had the wrong chirality at the  $C_{\alpha}$  atom. It has long been known that the naturally occurring chiral amino acids only ever occur in the L isomer in nature. Therefore any amino acids in the model which have the D isomer are very obviously incorrect. It is an easy procedure to step through the residues in the model and change the D amino acids to L. The Biopolymer® module of InsightII allows the user to change the amino acids chirality. After changing the chirality of the amino acid the Biopolymer module steps through the library of possible side chain rotamers and selects the rotamer with the least steric clashes. The library of rotamers is developed from analysis of the Brookhaven database. It was noted that each side chain prefers to adopt certain conformations (rotamers). It is much quicker stepping through the list of possible rotamers for an amino acid than rotating through all the possible conformations that the side chain can adopt.

A suggested reason why there were so many D amino acids in the structure is because the hydrogen atoms weren't added to the model until quite late into the energy minimisation procedure of the model. It wasn't until the heavy atom model was nearly fully relaxed that the hydrogen atoms were added. During the energy minimisation of the heavy atom model there was no constraints on keeping the chiral  $C_{\alpha}$  atoms in the correct chirality. When the hydrogen atoms were added by the COMMET modelling package the criteria for where the hydrogen atoms should be placed was purely geometrical, no consideration of the chiral centre's preferred chirality was taken into consideration. In such a case with hind sight the chirality of the amino acids in the model should have been checked when the hydrogen atoms were added and not left until the refinement stages of the model.

The next step in the refinement was to correct the chiral  $C_{\beta}$  atoms of the Ile residues which were the wrong chirality. Throughout nature only the one isomer is used in nature. Ile is no exception even though it has two chiral centres ( $C_{\alpha}$  and  $C_{\beta}$  atoms) only the one enantiomer is used in nature. The same reason as above for the incorrect chiral  $C_{\alpha}$  atoms is the likely cause of the incorrect chirality in some Ile  $C_{\beta}$  atoms. To fix the chirality the Biopolymer function "invert" was used. This works by reflecting through  $180^{\circ}$  the chiral centre along the axis chosen. The axis is made up by the first atom selected, the chiral centre selected and the third atom selected. In this case the chiral centre was the  $C_{\beta}$  atom and the other two

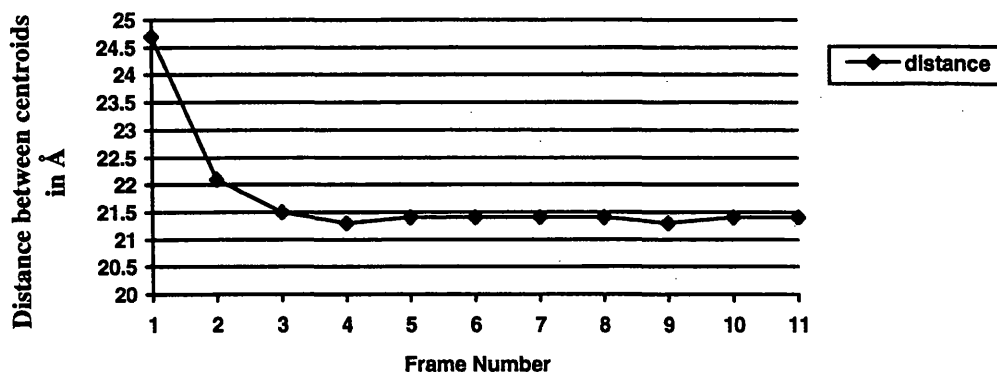
atoms defining the axis were  $H_{\beta}$  and  $C_{\delta 2}$  atoms. This effectively changed the position of the  $H_{\beta}$  and  $C_{\delta 2}$  atoms.

The next step in the refinement was to address the problem that the loop at insertion 170 extends too far from the protein surface. Although it is true that that surface loops are highly flexible and extend out into the solvent they do tend to rest on the protein surface. This is not the case for the second half section of Ins 170. To solve this problem the existing conformation was used as the starting point for further modelling studies. Molecular Dynamics was used in the attempt to find a conformation for the loop which does not extend out into the solvent to the same extent as the starting conformation. To encourage the loop to move back against the protein surface a constraining force was applied to the section of the loop allowed to move in the molecular dynamics. The constraining force that was set up such that it forced an atom towards the protein surface only if it was more than 16Å from the centroid of the model. The atoms constrained in this way were the backbone heavy atoms (N,  $C_{\alpha}$ , C, O). This was to encourage the side chains of the residues that were moving in the molecular dynamics to point into the solvent. The reasoning behind this is that it is well known that surface residues tend to point out into the solvent. If the side chains were also constrained to move towards the surface there would be a danger that the side chains may become buried and expose the backbone to the solvent.

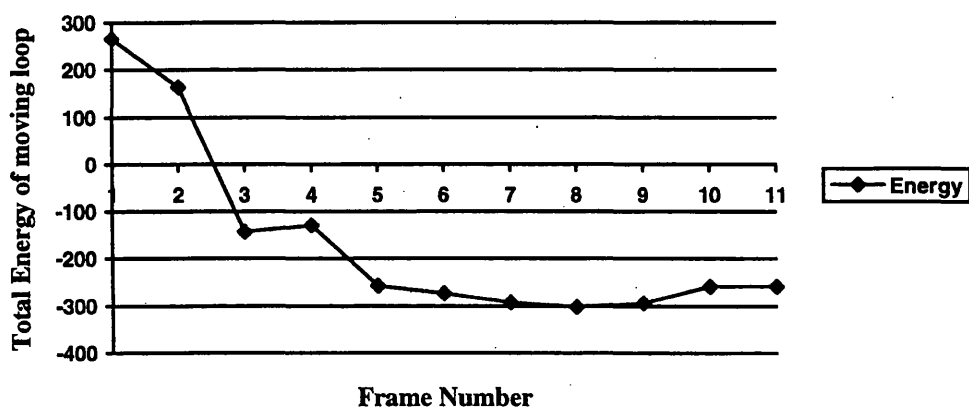
The residues allowed to move in the molecular dynamics run were from the fourth residue of Ins 170, which is the Tyr, to Val 176. This was a total of 15 residues. The dynamics run was carried over 25,000 iterations with the temperature at 500.0 K and using the constraints talked about above. The reason for the elevated temperature was to allow the backbone to move out of its present conformation and sample other conformations. The coordinates of the model was saved to a history file every 2,500 steps. This gave 10 conformations to examine.

The analysis of the loop conformation centred around the distance from the centroid of the protein model and the centroid of the moving residues in the loop. The smaller the distance between the two centroids the closer to the surface the loop is. Frame 8 gave the lowest

total energy for the loop but frame 4 gave the smallest distance between centroids (see Diagram 6-2 on page 170 and Diagram 6-3 on page 170). Frame 8 was chosen because the difference in the distance was negligible.



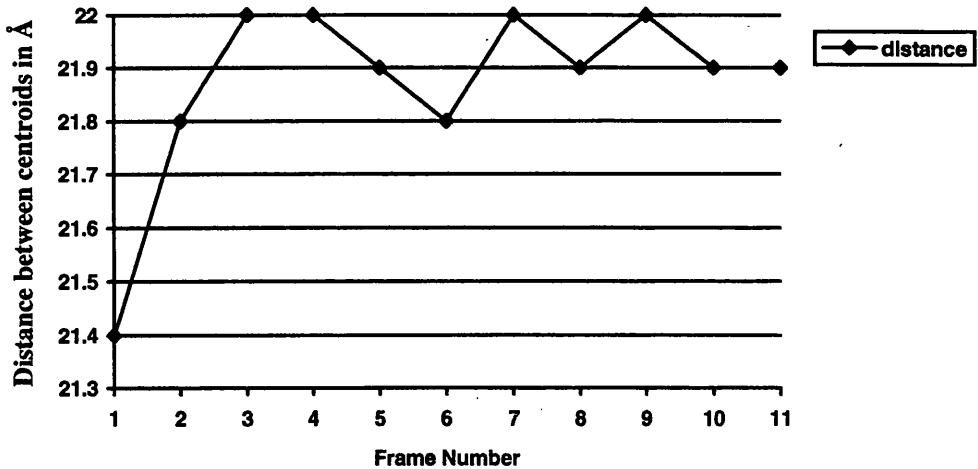
**Diagram 6-2:** The chart shows the distance between the centroid of the protein and the centroid of the moving residues for each frame in the molecular dynamics.



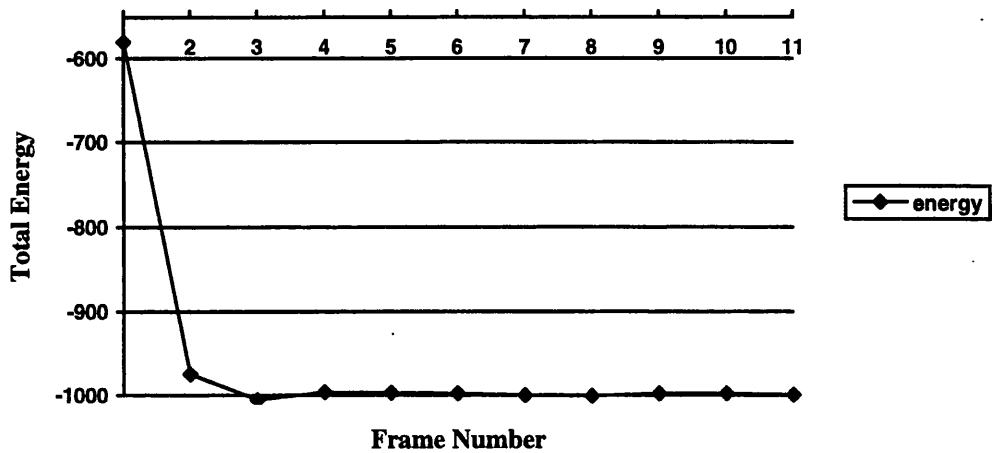
**Diagram 6-3:** This chart shows the total energy of the moving residues in each frame of the molecular dynamics.

Starting from the conformation of Frame 8 the same residues were allowed to move in a second molecular dynamics run. This time the temperature during the run was set to 300.0K and no constraints were used. This allowed the conformation of ins 170 to find the local minimum conformation. Again the molecular dynamics was carried out for 25,000 steps and the co-ordinates saved to a history file every 2,500 steps. The same analysis was

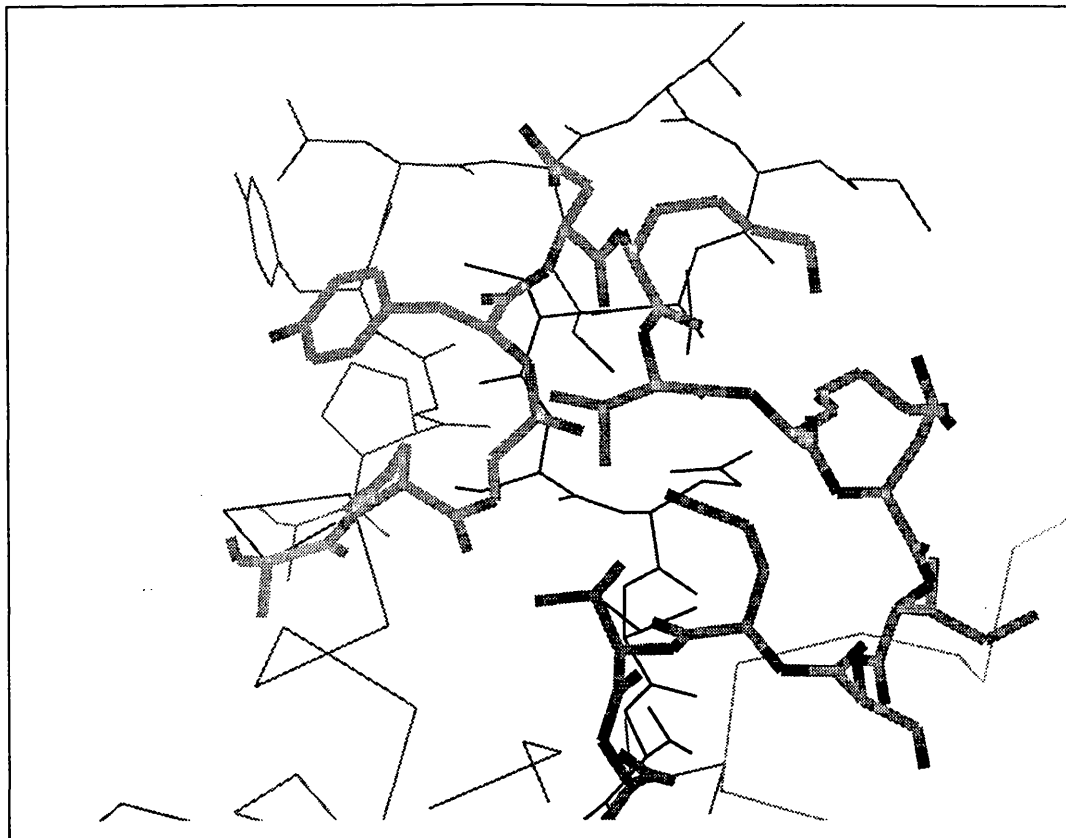
used for this dynamics run as in the first run (see Diagram 6-4 on page 171 and Diagram 6-5 on page 171). This time the conformation from frame 6 was chosen as it had the smallest centroid to centroid distance after the energy of the loop had settled. The new conformation of Ins 170 can be seen in Diagram 6-6 on page 172.



**Diagram 6-4:** The chart shows the distance between the model centroid and the centroid of the moving residues for each step in the molecular dynamics simulation.



**Diagram 6-5:** This chart shows the total energy of the moving residues for each step in the molecular dynamics simulation.



**Diagram 6-6: New and old conformation of Ins 170. The old conformation is in purple. This shows the new conformation of the insertion is lies more on the surface.**

Finally the global minimisation was carried out. As before the minimisation procedure was carried out in stages. First the entire backbone was held rigid as the side chain residues were allowed to be minimised.

Then just the  $C_{\alpha}$  atoms were fixed and the rest of the model allowed to move. At this point constraints were added to the model to try and fix some poor  $\phi/\psi$  angles in the model. To decide which  $\phi/\psi$  angles to constrain the BT crystal structure and the Bb model were put through WhatIf's MOTIF function. This superimposes two structures on the merit of conserved sections of their structure, disregarding the variable regions of the structures. At the end of this function an alignment is given which tells the user which sections of the two structures WhatIf considers conserved. The aligned conserved regions in the structure were

used to constrain the equivalent  $\phi, \psi$  angles in the Bb model to the corresponding  $\phi, \psi$  angles from the BT crystal structure. The constraints were set up such that there was a force applied to the torsion angle if it deviated by more than 5Å from the torsion angle found in the BT crystal structure. In total 272 torsion angles were constrained during the minimisation.

The model put through the WhatIf analysis is taken from the result of this minimisation. Resource and time constraints prevented any further refinement being carried out. The biggest problem was the lack of resources. As the initial modelling was carried out three years ago the original machine and software package were no longer available. Only limited access to molecular modelling packages was available at Daresbury Laboratory, which was compounded when Biosym's Discover license was not renewed. The final minimisation calculations were carried out thanks to the good will of colleagues at Proteus Molecular Design Ltd.



Ins 30 Short	All energy in kcalsmol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
1	1436000	2914	2947	3134						
2	54290000	86550	2122	2203	5047					
3	16720000	192300	1747	1262	1201					
4	26135	1021	850	651.4	544.9	504.1	471	369.3	333.7	
5	246100	3116	1740	1207						
6	*****	6079000	13620	5927						
7	5361	1681	1064	1010						
8	18320000	18810	1054	800.4	520.1	455.7	421.9	354.3	306.7	
9	*****	22820000	20300	15480						
10	7363000	21350	9380	9565						
11	7112000	31250	6521	5549						
12	36590	5607	4499	4603						
13	24920000	4298	1923	1736	1616					

Take conf 8 and add Hydrogens	10x	10x																				
	→	→																				
8	252.4	194.3	175.4																			

**Table 6-5: Result of energy minimisation on the conformations of insertion Gln 30 produced by the loop conformation generator.**

Ins 30 Long	All energy in kcalmol <sup>-1</sup>					
	10x →	10x →	10x →	10x →	10x →	10x →
1	4032000	21920	15690	14950		
2	*****	71570000	12820	8352		
3	8266000	20480	12110	11680		
4	*****	7167000	15680	12520		
5						
6	*****	833200	7617	5915	5264	
7	*****	283400	18130	17940	17630	
8	*****	10660000	35390	24810	23610	
9	1068000	19060	13580	13100	13050	
10	8391000	11090	3591	2889	2334	7232
11	9326000	32200	4607	2637	2599	1911
12	*****	1109000	22890	18660	17300	
13	*****	149600	15570	15310		
14	6308000	43000	8201	8131		
15	*****	*****				

Ins 30 Long	All energy in kcalsmol <sup>-1</sup>					
	10x →	10x →	10x →	10x →	10x →	10x →
1	4032000	21920	15690	14950		
16	*****	1663000	17950	15520	15470	
17	*****	20660	13830	13590	13400	
18	*****	687200	22690	15750	15630	
19	13090000	255500	251000	239500		
20	*****	*****	6933000			
21	*****	12930000	12830	12390	11850	
22	36590000	329200	15570	12390	11130	
23	*****	146300	16830	15370	15100	
24	17980000	269100	22060	21660		
25	*****	1493000	15010	9417	8453	7978
						7755

26	*****	6118000	5969000	592400				
27	*****	5792000	11130	11000	10920			
28	*****	89930000	89650000					
29	*****	*****	68230	16720	12380			
30	*****	*****	879300	12080	10630			

**Table 6-6: Results of energy minimisation on the conformations of 'long' insertion at Gln 30 produced by the loop conformation generator.**

Ins 186	All energy in kcalmol <sup>-1</sup>														
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
1	*****	*****	1395000	155000											
2	*****	*****	*****												
3	*****	*****	44930000												
4	*****	14360000	12840000												
5	*****	*****	347800	220900											
6	*****	*****	63760000												
7	*****	1078000	458500												
8	*****	*****	132500	52470	45810										
9	*****	89680000	48650000												
10	*****	312400	93780	89560	87320										
11	*****	41130	6286	5592	5379	5222	5002	4859	4535	4368					
12	*****	437100	18990	18560	18430										
13	*****	495000	38360	37980	37640										
14	*****	1318000	58340	57880	57090										
15	*****	21430000	227700	175400	172600										

Ins 186	All energy in kcal/mol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
16	*****	29400	19680	18310	17540					
17	*****	61870000	61650000							
18	*****	34250000	34010000							
19	*****	*****	*****							
20	*****	3055000	1989000							
21	*****	*****	329000							
22	*****	6485000	2149000							
23	*****	237200	73650	73850						
24	*****	32380000	12600000							
25	*****	*****	98450000	98290000						
26	*****	*****	*****							
27	11560000	4869	2721	2370	2185	2101	2056	1957	1883	1842
28	*****	*****	45390	45210						
29	*****	14350000	7868000							
30	*****	24870000	16920000							
31	752300	63450	39780	40060						

Ins 186	All energy in kcal $\text{mol}^{-1}$										
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
32	*****	664100	14660	9993	10030						
33	38870000	156200	99340								
34	*****	24770000	66020	29580							
35	*****	46940000	49500	38110							
36	4174000	6612	5172	4682	4470	4336	4246	4184	3862	3384	3132
37	*****	89790	10550	9792	9637	9248	9171	9068	9024	8993	
38	1583000	7202	3863	3566	3318	3257	3215	3187	3007	2870	2655
39	*****	244700	27380	24090							
40	*****	20650000	314500	308200							
41	*****	49500	45710	44380							
42	*****	10940000	64150	52770							
43	*****	1627000	23090	19750							
44	1992000	6153	4171	3877	4260	4166	4101	4068	4052	3971	3601
45	*****	965300	117200								
46	*****	2647000	1799000								
47	*****	74090	43510	40630							



Ins 186	All energy in kcalmol <sup>-1</sup>													
	10x →	762700	10x →	34980	10x →	32470	10x →	10x →	10x →	10x →	10x →			
48	*****													
49	*****	*****		1388000										
50	*****	60320		26180		23390								
51	*****	*****		120900		47370								
52	749000	8249		6047		5844		5759	5602	5093	3948	2853	2704	2623
53	2389000	44660		43570		41650								

Ins 186	All energy in kcalsmol <sup>-1</sup>										
Add Hs											
	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x
	→	→	→	→	→	→	→	→	→	→	→
27	3154	2595	2604	1711	1648	1624					
36	4515	4037	4042								
38	5725	5706	5688								
52	3327	3415	3354	1731	1685	1655					

Table 6-7: Results of energy minimisation on the conformations of insertion 186 produced by the loop conformation generator

Ins 127	All energy in kcal $\text{mol}^{-1}$									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
1	568.3	543.9	480.3							
2	480.3	388.0	351.1							
3	369.0	319.8	304.9							
4	382.8	365.5	358.7							
5	531.3	396.3	350.5							
6	402.2	373.7	337.4							
7	568.7	466.1	456.6							
8	372.4	346.1	301.9	302.0						
9	384.1	358.2	314.4	307.2						
10	507.1	350.2								
11	467.8	338.5	312.0							
12	435.3	358.1								
13	501.5	366.7	319.1							
14	435.7	335.1	237.6							
15	441.7	334.8	380.4	296.5						

Ins 127	All energy in kcalsmol <sup>-1</sup>			10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
	10x →	10x →	10x →										
16	425.1	484.9	306.5	302.9									
17	373.1	324.0	341.0	307.4									
18	1120	600.3	310.7	303.9	300.2								
19	339.9	317.2	269.9	265.6									
20	360.1	306.2	285.5	279.4									
21	361.0	323.8	308.1	263.3	259.4	275.3	252.0	247.4					
22	359.3	305.6	296.7										
23	357.3	299.1	263.9	257.4	254.7	255.6	254.1	251.8					
24	467.2	427.7	343.7										
25	619.8	443.2											
26	342.3	283.8	264.0	305.6	263.4	257.6	250.9	249.7					
27	353.2	299.5	287.1	277.7									
28	368.0	367.9	298.6	259.8	259.7	259.2	255.1	253.2					
29	353.1	298.8	277.3	271.1									
30	343.2	324.6	324.8										
31	420.6	285.8	277.4										

Ins 127	All energy in kcalsmol <sup>-1</sup>									
		10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
32	354.0	342.9	301.3	265.7						
33	312.8	285.7	258.2	261.3						
34	342.7	293.1	268.4	264.2						
35	347.6	389.8	371.3							
36	376.5	292.7	272.5							
37	369.0	290.3	277.8	274.116						
38	321.4	275.6	264.0	258.4	264.7	258.9				
39	362.9	280.4	272.1	258.3	252.6	251.7	249.0			
40	368.5	297.3	289.7	267.6						
41	440.8	309.9	274.2	271.4						
42	369.3	285.1	265.0	263.9						
43	3744	311.5	280.3	272.3						
44	485.2	383.7	340.7							
45	385.6	308.0	292.6							
46	402.4	537.1	346.8							
47	1172	404.2	279.5	260.8	257.6	252.8				

Ins 127	All energy in kcal/mol <sup>-1</sup>						10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
		10x →	10x →	10x →	10x →	10x →											
48	421.9	280.7	266.1	265.4													
49	330.6	337.8	271.6														
50	346.5	272.2	257.6	261.3													
51	390.0	308.2	289.1	288.8													
52	372.2	309.2	299.8	289.7													
53	349.2	286.3	275.5	270.9													
54	384.0	295.1	275.4	260.7	256.7	256.3											
55	425.0	313.3	299.5														
56	407.6	345.4	339.9														
57	337.6	280.2	272.8	279.4													
58	389.3	305.1	290.2														
59	436.0	323.5	299.1														
60	357.9	317.1	282.6	271.5													
61	375.3	299.3	291.1	283.0													
62	365.7	340.1	298.6	281.0													
63	381.8	285.4	284.7	267.2													

Ins 127	All energy in kcalsmol <sup>-1</sup>												
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →			
64	346.8	287.7	287.6	269.7									
65	391.4	299.6	278.2	275.6									
Add Hs	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →			
21	112.1	80.72	70.30	68.59	67.06	65.47	64.27	62.64	61.55	60.36	59.4	58.24	57.86
	57.86	57.05	56.13										
23	185.8	108.9	84.85	79.04	73.86	71.55	70.02	69.41	66.77	64.25	60.24	58.64	57.09
	57.09	55.99	54.19										
26	108.1	64.77	45.67	42.23	40.19	38.23	36.71	35.31	34.38	33.37	32.8	32.37	31.73
	31.73	31.08	30.99										
28	85.99	54.93	53.72	42.71	40.35	38.08	37.03	36.31	35.51	35.29	34.83	34.35	33.9
	33.9	33.26	32.66										
39	118.3	89.21	76.7	72.47	69.74	67.96	67.33	65.91	64.2	63.3	62.97	62.68	62.38
	62.38	61.82	61.43										

**Table 6-8: Results of energy minimisation on the conformations of insertion 127 produced by the loop conformation generator.**

Ins 231 using config 28 from Ins 127	All energy in kcal/mol <sup>-1</sup>													
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
1	868.4	701.5	591.7	541.8	421.4	410.3	366.5	361.2	336.2	329.2	326.3			
2	396.2	355.7	346.9	307.5	291.5	260.7	255.2	249.8	246.1					
3	425.4	284.6	231.9	217.8	175.6	164.4	133.2	124.6	118.7	103.3				
4	475400	1216	1059	608.2	498.2	451.2	349.9	311.8						
5	1793	910.9	894.5	793.3	653.1	630.06	623.8							
6	990.3	856.3	766.2	723.7	628.3	602.8	596.3							
7	1272	1127	1077	950.4	890.5	850	820.8	795.8						
8	1476	977.8	901.4	738.4	664.6	622.8	605.6	590.3						
9	1842	984.1	985.4	901.9	757.9	723	706.8							
10	942.5	808.8	752.8	726.9	677.8	652.5	640.1							
11	534.5	468.5	394	341	313.5	288.9	280.3	273.4						
12	48450	1468	1385	1359	1323	1291	1277							
13	3656	1825	1693	1643	1375	1335								
14	980.1	857.1	754.9	729.1	675	635	620							



Ins 231 using config 28 from Ins 127	All energy in kcal/mol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
15	31490	349.9	254.4	232.5	194	158	144.7	131.4	120.4	109.8
16	454.5	384.9	341.8	339.9	305.8	274.7	264.9	256.6	249	242.1
17	351	277	244.8	228.4	187.9	162.2	153.2	147	141.6	136.7
18	258.3	176.2	156.4	152.5	116.7	94.1	87.58	78.42	73.41	69.71
19	312.7	229.7	172.2	150.4	121.9	94	87.77	81.62	72.34	73.2
20	254.4	166.3	141.2	119.1	108.1	79.63	71.86	66.79	62.1	58.12
21	427	393.3	320.3	301.1	250.9	213.5	200	188.3	177.2	166.5
22	292.3	237	180.6	179.5	136.5	110.6	105.2	100.6	96.26	93.66
23	358.2	207.3	177.7	173.6	150.3	125.8	119	113.6	107.9	103.4
24	184.3	188.9	143.3	142.4	114.7	89.63	81.02	74.03	67.63	61.63
25	362.7	218.6	173.5	161	150.7	116.8	87.59	79.7	73.23	68.05
26	413.5	237	181.3	170.4	134.4	106.8	96.27	87.52	79.65	71.61
27	232.2	170.8	121.5	128.2	91.43	65.6	58.56	51.63	44.78	37
28	224.5	142.3	117.4	122.6	75.9	49.55	45.21	40.59	36.02	34.24
29	1788	551.3	474.2	442.6	397.1	287.9	250.7	229.7	217.4	

Ins 231 using config 28 from Ins 127	All energy in kcalmol <sup>-1</sup>											
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →		
30	300.1	267.3	206.7	198.6	160.2	136.5	130.6	125.7	10x →	10x →	10x →	10x →
31	460.3	405.3	309.9	278.2	220.3	205	193.3	180.9				
32	2094	156.7	135.2	130	86.79	35.77	48.64	44.06		39.89	36.28	
33	1966	1860	1814	1798	1760	1731						
34	248.7	199.5	206.5	179.3	140.1	109.9	103.1	96.57		90.99	86.62	
35	266.3	200.1	146.9	134.3	93.29	67.88	60.63	54.49		48.42	43.55	
36	254.7	170.5	126.4	115.4	83.11	56.71	49.72	41.32		35.28	30.21	22.89
37	184	172.8	133.1	130	99	85.02	56.52	50.84		45.96	42.28	
38	612.8	590.4	416.2	380.9	326.6	289.3	273.6					
39	362.2	278.5	217.9	202.4	167	136.4	127.3	119.6				
40	426.3	313.1	267.6	239.4	196.8	161.2	147.8	136.6				
41	1065	544.2	474.9	436.3	377.5	338.7	323	309.2				
42	177.4	166.4	137.4	147	114.5	90.66	84.43	79.84		76.39	73	
43	286.2	260.8	231.6	223.4	188	164	156.1	147.6				
44	193.2	161	101.2	101.5	73.11	47.07	42.44	38.47		35.11	32.29	28.06

Ins 231 using config 28 from Ins 127	All energy in kcal/mol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
45	200.2	177.5	134.5	133	98.88	69.85	63.77	58.93	54.95	51.74
46	343.1	239.5	207	199.5	159	127.8	116.8	106.3		
47	5556	228	174.2	172.7	136.4	102.6	92.93	83.02	74.24	68.2
48	395.2	210.2	190.8	194.4	144.8	111.6	101.6	94.41		
49	206.6	127.8	95.91	62.75	57.26	25.53	19.41	14.44	10.02	5.159
50	907.4	271.3	238.6	235.3	197.3	166.7	159.1	152.9		-4.173
51	359.7	264.3	230.3	226	188.9	158.6	148.7			
52	474.5	369.1	337.3	328.5	303.5	274.2	262.4			
53	551.4	343.2	292.1	274.4	230	201.9	189.9			
54	284.5	207	141.3	134.7	93.3	61.41	55.27	50.24	25.18	43.9
55	365.3	262.6	200	178.2	135.9	109.3	98.05	87.32	76.81	68.83
56	2589	704.8	577.2	535.9	481.1	430.8	403.2			
57	226	183.2	150.5	152.5	112.9	86.43	79.67	72.31	66.67	63.19
58	215.8	154.9	120.9	93.73	85.38	56.98	50.65	45.99	42.27	38.27
59	229.7	184.8	142.2	146	107	75.83	68.12	61.49	54.48	49.63

Ins 231 using config 28	All energy in kcal/mol <sup>-1</sup>											
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	
from Ins 127												
60	177	123	90.22	54.75	40.99	33.67	11.34	-3.091	-9.663	-12.9	-17.83	
61	344	285.8	246	226.8	178.7	149.1	137	126.5				
62	200	184.4	162.3	156.2	117.3	94.88	89.84	86.94	83.14	80.14		
63	219.9	136.1	116.4	101.9	73	45.73	37.08	28.28	20.1	14.26		
64	156.8	96.42	90.26	68.42	64.47	38.04	30.63	23.31	16.61	10.77	-0.1926	
65	377.5	266.7	191.8	184	136.2	100.5	89.8	80.64	73.68	67.75		
66	227.4	166	132.7	129.3	97.08	70.75	63.39	57.93	53.53	49.8		
67	401.5	273.6	226.3	120.3	179	148.2	136.9	74.74				
68	152	123.8	84.65	52.53	50.14	26.87	22.43	16.54	12.26	9.552	5.272	
69	2364	2226	2179	2172	2131	2104						
70	286.7	153.6	87.63	63.11	59.99	31.92	28.89	24.8	21.91	20.17	17.14	
71	236.9	199.4	162.7	152.7	126.8	99.89	92.9	86.18	81.1	77.29		
72	340.5	249.8	198.9	182.7	143	11.7	104.7	102.6				
73	438.9	405.8	336.4	312	256.2	220.4						
74	255.7	214.2	156.7	152.8	112.8	83.46	74.52	67.07				

Ins 231 using config 28 from Ins 127	All energy in kcal/mol <sup>-1</sup>																			
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →										
75	257.2	184	126.4	116.3	109	87.98	81.06	75.77												
76	426.4	282.8	204.5	175.3	133.4	101.5	93.13													
77	238.2	188.2	167.5	135.9	128.4	96.13	87.36	81.02												
78	305.4	204.2	174.2	169.8	135.6	109	101.1	94.61												
79	352.3	335.2	289.4	275.4	228	196.6	184.8													
80	241.6	186.6	120.7	117.7	84.08	58.37	50.78	44.14	39.24	35.11										
81	330.6	297.2	212.7	191.3	159.3	128.3	119.6													
82	2132	1544	1408	1340	1247															
83	824.3	637.9	521.9	472.4	409	370.2	352.5													
84	2374	2227	2111	2086	1994	1951														
85	9235	797.3	706.7	690.5	649.7	615.9	597.7													
86	368.1	298.7	242.9	211.4	159.7	124.8	112.3	100.4												
87	636.6	418.5	368.9	327.4	273	229.5														
88	450.3	343.5	319.9	269.1	258.3	224.3	214.5													
89	1532	960.9	598	484.6	408.7	349	320													

Ins 231 using config 28 from Ins 127	All energy in kcalmol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
90	817.5	567.7	463.8	425	361.5	320.1				
91	477.3	358.2	335.6	303.9	297	265.6	253.3			
92	538.4	362.3	301.6	285.7	240.4	203	187.8			
93	509.3	401.9	309.1	286	272.2	208.4	194.5			
94	715.1	489.4	442.1	398.3	378.4					
36	22.89	18.13								
49	-4.173	-13.15	-20.69	-24.72	-29.59	-32.93	-35.35	-37.8	-40.34	
60	-17.83	-21.21	-23.61	-25.31	-26.38	-27.28	-28.11			
64	-0.1926	-6.923	-10.62	-13.35	-15.92					
68	5.272	0.7455	-2.723							
70	17.14	12.91	9.615							

Table 6-9: Results of energy minimisation on the conformations of insertion 231 using conf 28 from insertion 127, produced by the loop conformation generator.

Ins 231 Using config 26 from Ins 127 [using confs from final energy from config 28		All energy in kcal/mol <sup>-1</sup>						
	10x →	10x →	10x →	10x →	10x →			
49	622.3	105.1	35.9	-1.488	-25.56	-27.94	-30.43	-33.66
60	973.2	41.24	10.29	-14.63	-15.1	-16.48		
64	129.6	33.18	9.228	-1.471	-4.91	-8.296		
68	362.1	60.64	28.03	13.9	8.333	3.564		

**Table 6-10: Results of energy minimisation on the four conformations of insertion 231 with the lowest energy using conformation 28 of insertion 127 now tested with conformation 26 of insertion 127.**

Ins 127 using config 49 of Ins 231	All energy in kcalmol <sup>-1</sup>			
	10x →	10x →	10x →	10x →
26	58.66	-12.63	-15.23	-16.04
28	45.83	0.4565	0.4437	3.888

**Table 6-11: Results of energy minimisation on the two conformations of insertion 127 (conformations 26 and 28) with conformation 49 of insertion 231.**



Ins 101 Moved to be at end of beta sheet. Add residues Asn → Glu after 96 and Phe → Ile after 98	All energy in kcal/mol <sup>-1</sup>					
	10x →	10x →	10x →	10x →	10x →	
Asn 96A	698.3	162.4	131.5	128.6	124.9	121.2
Phe 98A	*****	2451000	3829	3813	3810	
Gly 96B	8635	205.5	154.4	146.5	142.4	140.4
Glu 98A	3847	432	260.3	229.2	218.7	215.2
Lys 96C	352	288.7	256.6	252.3	247.6	239.7
Pro 98A	4368000	1474	1123	1068	1059	1057
Lys 96D	135900	1067	1044	1025	1015	1003
Ile 98A	863200	1491	1191	1118	1087	1065
Glu 96E	365800	1129	982	956.7	943.1	

Carry out substitution to correct sequence	All energy in kcalmol <sup>-1</sup>				
	10x →	10x →	10x →	10x →	10x →
216800	1133	1057	1009	985.2	981.7

**Table 6-12: Results of energy minimisation after each residue in the insertion is inserted into the model.**

Insert Asn → Glu after Ile 96, and Ile → Phe after Gly 89	All energy in kcal/mol <sup>-1</sup>							
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
Asn 96A	693.5	153.2	138.9	122.3	119.3	117.8		
Ile 98A	19670000	344.4	172.1	171.2	150.2	144.1		
Gly 96B	499.1	175.6	153.8	136.2	132.4	128.5		
Pro 98B	2021	242.6	184.9	171.9	163.4	154.7		
Lys 96C	520.8	169.6	135.7	128.3	123.3	119.2		
Glu 98C	484	236.7	154.2	135.1	129.6	125		
Lys 96D	406.1	159.2	121	116.8	114.7	113.2		
Phe 98D	*****	4741	2736	652.3	376.5	310.8		
Glu 96E	68532	508.6	320.9	262.9	222	210.2		

Substitute 97 Asn → Ala, 99 Lys → Tyr, 100 Lys → Asp and 101 Glu → Tyr	All energy in kcalsmol <sup>-1</sup>						
	10x →	10x →	10x →	10x →	10x →	10x →	10x →
	5310	354.9	283.8	236.7	209.9	205.3	200.2
Add hydrogens to the final structure and continue with energy minimisation	All energy in kcalsmol <sup>-1</sup>						
	10x →	10x →	10x →	10x →	10x →	10x →	10x →
	654500	766	151.4	40.31	29.79	19.7	18.18

**Table 6-13: Results of energy minimisation after each residue in the insertion is inserted into the model.**

Insert Gly, Glu after Asn 97, then Lys, Pro after Gly, then Lys, Ile after Lys, then Glu, Gly after Lys, and Ala after Glu	All energy in kcal/mol <sup>1</sup>								
	10x →	10x →	10x →	10x →	10x →	10x →			
Gly 97A	269300	7898	4813	3277	1779	796.4	539.1	10x →	10x →
Glu 97B	173200	5946	3580	3341	599	515.7	464.7		
Lys 97B	785.9	474.1	363.7	333.9	323.9				
Pro 97C	8293	467.4	385.7	354.2	333.1	331.2			
Lys 97C	3167000	563.7	376.7	352	346.5				
Ile 97D	41630000	539.6	413.4	365	340	330.9			
Glu 97D	7315000	727.9	483.8	415.5	378.9	367.6			
Gly 97E	128200	583.2	402.9	339.4	299.2	284.9			
Ala 97E	*****	3033	975	762	640.3	560.8			

Substitute 98 Gly → Phe, 99 Lys → Tyr, 100 Lys → Asp, 101 Glu → Tyr	All energy in kcalmol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
1973	753.1	585.1	501.3	490.7	451.9	441.8	435.5	433.9		
Add hydrogens to final structure and continue with energy minimisation	All energy in kcalmol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
2684	379.7	358	352.7	348.2	343.9	340.3				

**Table 6-14: Results of energy minimisation after each residue in the insertion is inserted into the model.**

Substitute Gly 98 → Phe, Lys 99 → Tyr, Lys 100 → Asp, Glu 101 → Tyr	All energy in kcalsmol <sup>-1</sup>							
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
54850000	2153000	7627	1941	1355	831.9	734.3	723.6	721.4
0								
Insert Gly, Glu after Asn 97, Lys and Pro after Gly, Lys and Ile after Lys, Glu and Gly after Lys and Ala after Glu	All energy in kcalsmol <sup>-1</sup>							
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
Gly 97A	3812	509.8	301.5	229.1	204.8	196.6		
Glu 97B	103930	372	259	221.5	196.6	193.1		
Lys 97B	1158	343	262.7	222.6	191.1	185.5		
Pro 97C	2182000	857.5	461.1	359	319.3	302.2		
Lys 97C	139100	680.3	337	275.7	247.6	237.2		

Ile 97D	6625	594.7	301.2	231.7	202.3	196.5				
Glu 97D	*****	2731	1306	1084	1030	971.1				
Gly 97E	1832000	10380	2626	1428	1019	731	603	419.5	352.1	269.1
Ala 97E	1028400	5312	3373	2741	484.1	465.1	392.8	354		
Add H's	All energy in kcal/mol <sup>-1</sup>									
		10x	10x	10x	10x	10x	10x	10x	10x	10x
		→	→	→	→	→	→	→	→	→
	380	341.7	291.5	280.5	259.6	242.7	238.9	236.9		

Table 6-15: Results of the energy minimisation after each residue in the insertion is added into the model.



Ins 143. Order in which the residues were inserted	All energy in kcalmol <sup>-1</sup>						
	10x →	10x →	10x →	10x →	10x →		
Leu	2406	723.8	491.0	442.1	397.0	366.0	352.4
Pro	1243	773.1	641.9	589.4	516.9	471.3	448.2
Ala	1485	904.0	695.5	584.5	506.5	461.4	446.9
*Gln	936.4	470.3	425.7	394.2	377.0	340.9	325.9
Asp	9582	9139	8634	8558	8536	8434	8405
Ile	10280	10130	9355	7729	7465	7223	7056
Lys	9013	6337	5930	5752	5618	5579	5562
Ala	27210	29150	28730	28360	28160	28100	28040
Leu	*****	126200	127700	127800	127500	127500	127500

**Table 6-16: Results of energy minimisation after each residue in the insertion is inserted into the model. The \* indicates where this model was saved and the rest of the results discarded due to the high energies.**

Ins 143. Starting from Gln in the insertion	All energy in kcalmol <sup>-1</sup>							
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
Asp	16720	2423	1959	1826	1752	1705	1693	
Ile	18290	4730	3957	3819	3798	3792		
Lys	18580	8641	8309	8129	8103	8067	8052	
Ala	27210	3735	3563	3517	3500	3454	3452	
Leu	*****	23000	2171	2118	2110	2073	1987	

**Table 6-17: Results of the minimisation after each residue of the insertion is inserted into the model. This is the second attempt where the hydrogen atoms are added to the model.**

All energy in kcalmol <sup>-1</sup>	
10x →	10x →
898.4	672.7
762.4	689.1
	672.7

**Table 6-18: Further energy minimisation of the insertion at residue 143 but with the hydrogen atoms removed from the model.**

All energy in kcalmol <sup>-1</sup>					
	10x →	10x →	10x →	10x →	10x →
912.1	751.9	707.5	693.6	670.8	664.0

**Table 6-19: Energy minimisation of insertion at residue position 143 after the phenyl ring of the phenylalanine at position 144.**

<b>All energy in kcalsmol<sup>-1</sup></b>														
	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	20x		
	→	→	→	→	→	→	→	→	→	→	→	→		
888.1	777.7	708.1	679.6	663.1	643.1	619.5	600.7							
<b>The aromatic ring on Phe 144 fixed again</b>														
	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	60x		
	→	→	→	→	→	→	→	→	→	→	→	→		
1050	908.8	762.9	696.9	649.9	620.2	598.8	540.1							
<b>The aromatic ring on Phe 144 fixed again</b>														
	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	10x	60x		
	→	→	→	→	→	→	→	→	→	→	→	→		
1050	908.8	762.9	696.9	649.9	620.2	598.8	540.1							

The aromatic ring on Phe 144 fixed again							
10x	10x	10x	10x	80x	80x	80x	100x
532.7	505.5	492.4	484.0	479.2	463.2	460.1	453.4

**Table 6-20: The phenyl ring of the phenylalanine residue at position 144 is fixed to the torsion and planar angles, and then the fragment is put through the energy minimiser.**

Model and $\alpha$ helix fragment	All energy in kcalmol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
11680	422.2	277.6	255.8	239.9	232.5	228.4	234.2	224.2		
Add hydrogens	All energy in kcalmol <sup>-1</sup>									
10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
74.33	52.44	43.89	25.47	13.61	-1.882	-28.45	-32.71	-35.53	-37.83	

Table 6-21: results of energy minimisation after the helix fragment is added to the model.

Ins 170	All energy in kcalmol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
1	*****	26610	7832	7326	6979					
2	32970000	5189	1461	825.4	716.2	686.1				
3	*****	83280	8477	8376						
4	*****	1091	644.1	571.5	534.7	522.8				
5	1465000	2412	1029	820.4	720.3	687				
6	*****	2017	81	880.8	811.1	774.1				
7	577300	1318	663.1	563.2	528.3	514.7				
8	11860000	1948	841.5	707.4	1994	1102	703.5	564.2		
9	25630000	2132	880.8	676.6	600	554.6				
10	8927	860.9	572.9	51501	424.7	457.9				
11	314900	1092	748.9	668.8	627.3	610.6				
12	*****	9515	6874	5732	5181					
13	*****	2951	2367	2052	1886					
14	7119000	2970	1637	1056	902.4	837.5				
15	*****	1767	879.4	726.7	661	616				

Ins 170	All energy in kcalmol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
16	*****	25690	10620	904.9	864.3	862.4				
17	7316	714.9	572.7	509.6	481.4	474.1				
18	40920	1032	603.5	528.7	491.4	477.8				
19	216400	798.9	586.6	525.4	495.5	488.4				
20	7904	703.5	569.7	513.6	482	473.4				
21	162600	1135	718.8	618.9	573.6	554.4				
22	*****	132800	13410	13280	13190					
23	649600	1854	816.2	698.4	646	618.6				
24	79810	554.3	450.5	402.3	378.5	374.9				
25	26340	1277	699	566.9	507.1	477.7				
26	199600	1286	698.4	586.5	534.1	511.1				
27	144900	1010	755.6	698	667.5	651.6				
28	50570	1273	689.9	612.3	569	548				
29	274100	1492	1001	900.7	851.1	820.8				
30	*****	1392000	20070	14420	14240					
31	52650	936.9	736.2	669.2	633	617.2				



Ins 170	All energy in kcalmol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
32	17350000	9476	3242	3098	3013					
33	27410	1169	614	520.6	473.8	454.1				
34	264600	1301	810.1	695.3	642.4	619.1				
35	*****	11070	6315	4613						
36	*****	2860	916	677.6	624.1	600.8				
37	4809000	4476	4076	3324						
38	*****	3165	1241	776.5	694.2	644.4				
39	*****	14140	6304	5763						
40	*****	7696	3140	1756	1460					
41	349800	955.9	678.8	585.7	538.5	515.1				
42	*****	397900	12640	10080	9929					
43	*****	8485	4745	2869						
44	283400	1180	671.1	584.7	541.7	529.3				
45	1538000	1188	713.3	641.5	611.7	598.2				
46	*****	216100	18660	14380						
47	*****	5908	2533	2365						

Ins 170	All energy in kcalsmol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
48	*****	5557	3505	2963						
49	3593000	1915	871.5	687.5	629	595.8				
50	8716000	2676	1372	957.9	823.2	749.6				
51	*****	5750	3259	2623						
52	998900	1680	1065	818.7	734	697.4				
53	*****	40500	1232	820	718.4	671.9				
54	59670000	4795	2509	1576	1377					
55	27490000	11680	5388	4869	4769					
56	441300	1848	909.7	758.1	702.3	670.5				
57	1108000	2447	1982	1351	1023	911.5				
58	*****	82450	3545	3149						
59	*****	3130	3338	2727						
60	*****	71640	14970	11080						
61	*****	99040	19250	18670						
62	234100	913.8	626.5	556.9	526.1	511.4				
63	6068000	1839	923.2	803.6	740.7	700.1				

Ins 170	All energy in kcalmol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
64	11890000	1451	645.8	572.6	540.3	520.9				
65	14170000	1676	674.5	590.5	538	527.4				
66	325700	8345	585.8	515.8	470.7	449				
67	41560	792.5	593.5	522.2	486.3	475.5				
68	1314000	1869	939.3	749.7	690	663.1				
69	38650000	1635	776.4	671.3	630.3	611.4				
70	1055	758.9	585.9	528.7	502.3	496.2				
71	594500	984.4	558.4	466.7	430.4	422.6				
72	4474000	1661	1143	732.2	643.5	582.9				
73	6812000	1705	822.3	686.4	634.8	613.7				
74	368300	833.6	616.5	533.8	490.4	474.9				
75	365400	833.6	616.5	533.8	490.4	474.9				
76	*****	49830	15070	12600						
77	4938000	1106	634.1	538.5	502.3	488.1				
78	*****	51780	7004	6471	6402					
79	*****	34090000	9138	8100						

Ins 170	All energy in kcalsmol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
80	41320	818.7	609	528	501.3	492				
81	30010	1008	685.4	599.8	554.4	537.3				
82	7194000	1015	716.1	624.9	580.3	555				
83	87540000	62290	1632	1087						
84	*****	14610	2184	1263						
85	*****	26990000	7482000							
86	*****	21950	8527	6422						
87	*****	116300	32730	27410						
88	370800	1808	864.2	647.2	538.5	523				
89	87520	704.9	507.1	447.3	415.5	405.5				
90	1671000	2401	6669.4	561.7	517	499.8				
91	183400	1671	872.6	603	546	520.2				
92	109800	1196	686.6	524.4	478.3	462.4				
93	1961000	1263	607.9	540.3	506.3	492.3				
94	32820000	3253	3262	1549						
95	2037000	1354	596.3	502.4	463.1	450.4				

Ins 170	All energy in kcalsmol <sup>-1</sup>										
		10x →	712.4	10x →	524.4	10x →	465.3	10x →	436	10x →	325.6
96	46130										
97	64830000		8419		5661		1492				
<b>Add Hs'</b>											
		10x →		10x →		10x →		10x →		10x →	10x →
4	94.82		62.3		49.23						
7	36.84		10.87		2.064						
8	108.7		74.52		54.05						
9	168		122.4								
10	21.31		0.7922		-6.896						
17	68.65		47.49		43.69						
18	43.99		18.97		10.02						
19	70.44		48.64		41.61						
20	28.88		6.896		0.2989						
21	153.7		113.3		85.9						
24	70.19		23.78		4.983						

Ins 170	All energy in kcalsmol <sup>-1</sup>										
		10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
25	1.088	-25.9	-36.52	-73.12	-97.48	-111.4	-136.4	-119.7			
26	120.9	94.66									
28	101.8	85.24	78.66								
33	-9.074	-27.88	-63.4	-88.15	-99.23	-126.3	-140.4				
41	121.9	78.77	60.91								
44	40.72	9.497	-6.409								
62	120.7	80.59	51.12								
64	126.3	83.18	74.16								
65	127.5	86.91	66.46								
66	20.74	-6.274	-17.7	-50.58	-61.63						
67	98.64	59.47	39.42								
70	73.77	47.66	41.12								
71	-11.52	-34.92	-42.22	-62.76	-75.62						
74	5.64	-22.13	-36.57	-76.32	-102.2	-121.8	-137	-138.6	-143.8		
75	189.8	160.3									
77	142.1	111	96.56								

Ins 170	All energy in kcalsmol <sup>-1</sup>									
		10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
80	99.71	73.41	66.44							
81	45.61	-1.92	-15.26	-36.33	-80.91	-94.73	-103.5			
88	71.21	36.15	20.99							
89	-62.72	-84.91	-93.89	.122.1	-137.5	-142.3	-145	-151.2	-154.9	
90	47.88	21.38	12.38							
91	122.4	87.52	70.11							
92	90.68	59.43	50.66							
93	128.4	107.7	97.39							
95	-14.68	-33.95	-40.58	-63.15	-74.36					
96	-42.02	-62.24	-70.12	-103	-119.6	-133.5	-143.9	-151.1	-153.5	

**Table 6-22: Results of energy minimisation on the conformations of insertion 170 produced by the loop conformation generator.**

Deletion 63	All energy in kcal/mol <sup>-1</sup>									
	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →	10x →
1	3824000	15860	14180	1350	1294	1245				
2	23750000	16620	6104	5178	4791	4668	4609			
3	21860000	8850	1919	1833	1807	1781				
4	*****	38720000	4068000	1229000						
5	*****	91560	21870	16470						
6	*****	21500000	20870000							
7	23920000	16870	7391	7176	7004	6859				
8	*****	*****	*****							
9	*****	33580000	118400	112700						
10	*****	*****	14930000	38150	17680	5775	3098	2994		
11	*****	198400	157800	1516						



	Further minimisations	50x →	50x →	50x →	50x →	100x →	100x →
1	1217	1187	1014	958.1	867.1	854.5	
3	1763	1626	1537	1396	1313	1186	

**Table 6-23: Results of minimisation on the conformations of deletion 63 Asp produced by the loop conformation generator.**

## 7 Discussion

### 7.1 Discussion of the Model

#### 7.1.1 Assessment Of The Model

After the final minimisation the protein model was evaluated by running it through an analysis program. The analysis routine was the Protein Analysis function "FULCHK" used in the What If<sup>115,116</sup> package. The entire report can be read in Appendix A (chapter 8). A discussion of the report follows:

1. The space group information was added by WhatIf automatically. It has no effect in the analysis of the protein model.
2. The model co-ordinates had already been rounded to the correct number of decimal places (three for Brookhaven files) when the co-ordinates were written to pdb format by InsightII.
3. This error is due to there being no crystal information in the pdb file and can be ignored
4. No problems with nomenclature. Previous problems with nomenclature appear to have been caused by having H atom co-ordinates in the pdb file. The pdb file format was designed for protein crystal structures which due to the nature of crystallography do not have H atoms assigned. Therefore there is sometimes problems with reading in pdb files which do contain H atom co-ordinates.
5. Everything is OK, see point 4 above.
6. Everything is OK, see point 4 above.
7. Everything is OK, see point 4 above.
8. Everything is OK, see point 4 above.
9. Everything is OK, see point 4 above.
10. Everything is OK, see point 4 above.

11. Everything is OK, see point 4 above.
12. Everything is OK, see point 4 above.
13. Everything is OK. WhatIf could properly assign all the atoms in the file to the correct amino acid structure.
14. No side chains had the wrong chirality. This is a vast improvement on the original structure where 35 amino acids had the wrong chirality, a major error as only the L structure is ever found naturally in nature. The Ile C<sub>β</sub> atoms all now have the correct chirality. The problem that the chirality checker is having is due to the constraints on the backbone atoms of the model during the minimisation. This is causing the peptide bond to deviate significantly from planarity. This would quickly be resolved by carrying out further minimisations reducing the constraints on the backbone model. Unfortunately resource and time constraints do not allow further energy minimisation runs to be carried out.
15. Again the problem of the improper dihedral angles is down the constraints on the backbone during the final energy minimisation carried out.
16. The problem here is the way the original modelling package used, COMMET, renumbered the insertions. The standard expected by WhatIf is to use the last residue number before the insertion and use this as the base to number the insertion. The first residue in the insertion should be iA, the next iB, the next iC and so on. Instead COMMET numbers each insertion from 1, the next is numbered 2, the next 3 and so on. This means that the model has several residues named 1, 2, 3, etc. This is cause of the error.
17. Weights checked OK. They were all assigned 0.00 in the pdb file as default.
18. All atoms that WhatIf was expecting from examining the primary sequence were found in the pdb file.
19. C terminal oxygen atom was missing. This was missed out by COMMET package as the Bb primary sequence extends past the BT sequence. As no homology existed for this section of the primary sequence it was not modelled.

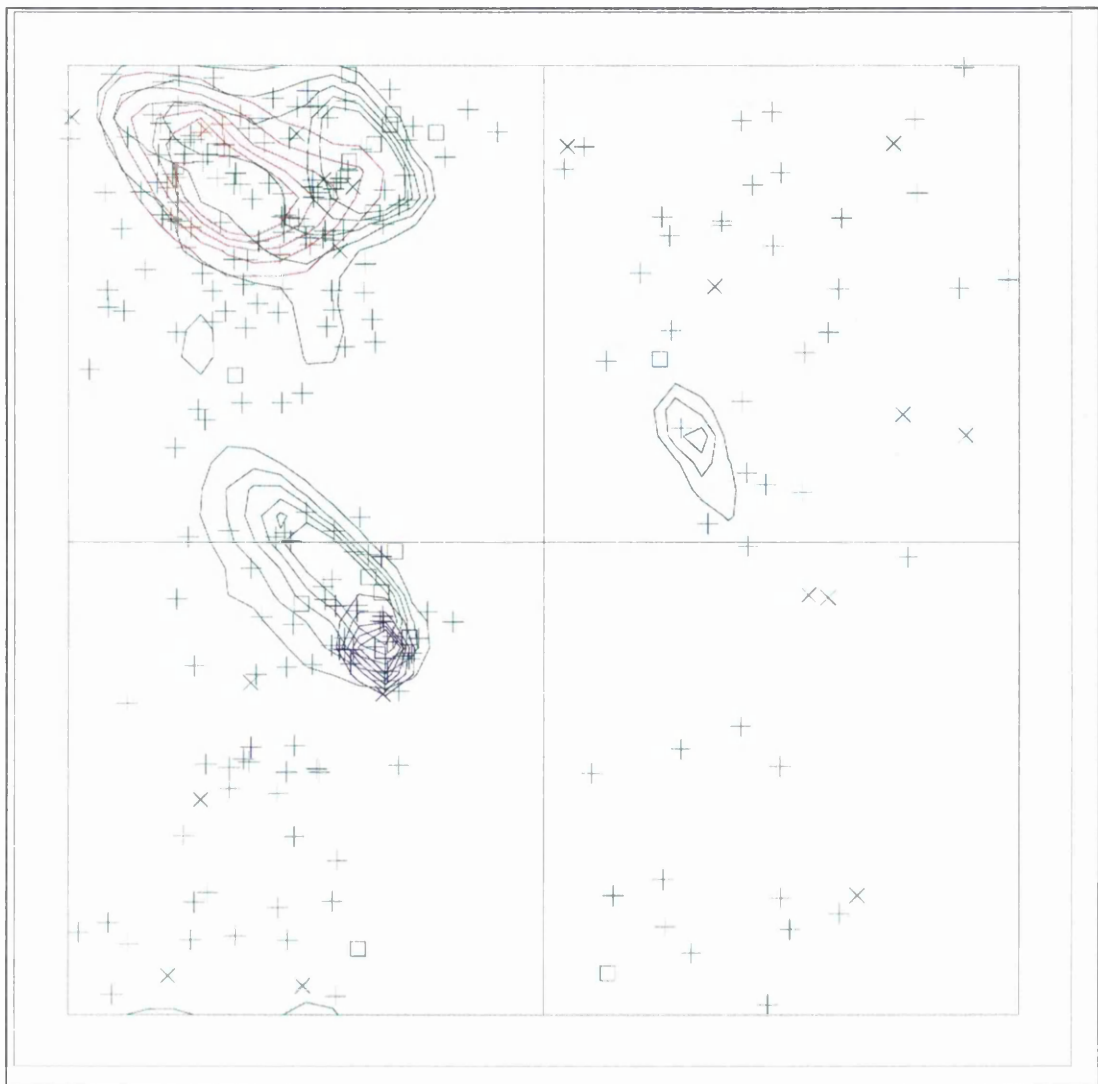
20. Same reason as in 19. It is normal to add a neutral group to the end of C terminal end of the model so that the model does not have an extra charge where none exists in the native protein.
21. In total 97 unusual bond length were reported by WhatIf. Two reasons are possible. Having this many in a minimised structure strongly suggests that a different bond length library was used in Biosym's CVFF force field. This is made more plausible considering that this problem was not seen in the original model minimised using the force field used by COMMET. The second reason is due the constraints on the backbone atoms used during the minimisation procedure. This second reason accounts for the many backbone bonds that deviate from the normal. Removal of the constraints while carrying out further minimisation would likely fix these bond angles.
22. The high bond length deviation is likely due to the constraints on the backbone dihedral angles. The z score is not that far above the normal that removing the constraints and carrying out further minimisations would solve this problem.
23. The possible cell scaling problem can be ignored as being a possible reason for the bond lengths being inaccurate. No cell units are applicable as this is a homology model and not a crystal structure.
24. 571 angles were reported as deviating from the default value by more than  $4\sigma$ . Again many of these angles are bond angles associated with the backbone of the protein model. The cause of these angles deviating from the normal is the constraints on the backbone torsion angles. The remaining bond angle deviations are probably due to a different bond angle being used as there were only 17 deviations reported in the original model.
25. The high bond angle deviation is likely due to the constraints on the backbone dihedral angles. The z score is not that far above the normal that removing the constraints and carrying out further minimisations would solve this problem.
26. The analysis found 15 side chains where a planar group deviates from planarity by more than 4.0 times the expected value. The cause of these deviations will be steric clashes with other side chain atoms. Examination of each residue is required to sort out

the probable cause of the clash. The problem can then be fixed by changing the torsion angle of the affected side chain or the side chain causing the problem. This causes only minor local changes in the structure. Of the side chains that were reported two of the residues are highly distorted from planarity.

27. A total of 17 aromatic side chains had a connecting atom severely out of the plane of the aromatic ring. Seven of these warnings were for the hydroxyl group of the Tyr residue. These can be easily fixed by examination of the area and the movement of only a few atoms. The more serious problem concerns the other 10 warnings where the  $C_{\beta}$  atom is reported to being out of the aromatic plane. Due to the bulk of the aromatic ring it will involve moving more atoms to try and accommodate the aromatic side chain in a different conformation. These might hopefully improve as the structure is further refined with the removal of the constraints on the backbone torsion angles.
28. 5 Pro residues had unusually high puckering amplitude, two of which are worrying high. The three Pro with the lower values have their backbone torsion angles constrained. Removing the constraint would allow the conformation to relax a bit more and remove the problem. The two Pro with the very puckering amplitude are more worrying. The likely cause here is steric clashes. Examination of the area around these two residues would hopefully point to the atom in the model causing the puckering.
29. A total of 6 Pro residues have unusual Pro puckering phases. All the reported Pro residues have their backbone torsion angles constrained during the energy minimisation. None of the phases are widely out which suggests removing the constraints would solve this problem.
30. Only 4 residues have a worrying score for their torsion angles. No comment in the report tells you if the backbone torsion angles are to blame or the side chain torsion angles. For the 4 worrying residues it is likely to be the backbone torsion angles that will be wrong. The reason being there are several residues with bad backbone torsion angles, which can be seen in the Ramachandran map later in this report.
31. There is still problems with poor  $\phi, \psi$  torsion angles but not as bad as the original model. Although there are more lines reporting errors many of the are due to the  $\Omega$  angle being poor. The reason for the poor  $\Omega$  angles is the constraints on the  $\phi, \psi$  angles

on many of the residues. This has improved the look of the Ramachandran map but has made the  $\Omega$  angles very badly constrained to planarity. Removing the backbone torsion angle constraints would help ease the  $\Omega$  torsion angle problem without making the  $\phi, \psi$  torsion angles worse.

32. The Ramachandran z score is still low but improved upon from the original model.
33. The warning about the  $\Omega$  torsion angle restraint not being strong enough is because the minimisation has been set to favour good  $\phi, \psi$  torsion angles by using constraints. This score can be greatly improved by the gradual easing of the constraints during following energy minimisations.
34. Although the analysis mentions that the  $\chi_1 / \chi_2$  correlation z-score is low this is a reasonable score considering the poor homology between the starting structure, bovine trypsin and Bb. With the sequence identity being so low this meant that there were many non conservative side chain substitutions made. These substitutions have their side chain atoms remodelled as they are structurally different from the original side chain. To add the problems was the problem that 35 residues ended up with the wrong chirality. This is something that should not happen if their were proper chirality restraints added during minimisation of the model before the H atoms were added.
35. The diagram below shows a Ramachandran plot of the  $\phi, \psi$  torsion angles in the model of Bb. It is an improved map from the original model, especially in the bottom right quadrant which is disallowed to all residues except Gly. The allowed regions for helical residues is coloured blue, for strand residues in red, all other regions green. For residues that are part of a helix are shown in blue, strand residues red. The x - signs represent Gly, square represent Pro, and small + signs all other residues.



**Diagram 7-7: Ramachandran plot for the refined model of Bb. Produced by the protein structure analysis routine "FULCHK" in WhatIf.**

36. The inside / outside distribution warning is expected. This model, Bb, is only part of the Factor B protein. It is homologous to the serine proteases so will fold like a serine protease, hence we can model the homologous section. But this is only one domain of the protein. The other section of the protein, Ba, will be in close contact with the serine protease domain and it is therefore expected that part of what is normally expected to be the surface of the serine protease fold will in fact be buried and in contact with the Ba domain of the protein. The same is also true when Factor B is cleaved into Ba and

Bb. Bb remains bound to the C3 / C5 convertase through protein - protein interactions. These interactions are nearly always through a large section of aliphatic residues on the surface of the two proteins. A normal score is 1.16, and Bb as modelled scored 1.236. This suggests that Bb is found on the surface of the C3 / C5 convertase which would also agree with its role in the complex which is cleave further C3 or C5 proteins. This role suggests that it should be easily accessible to the surrounding solution of the complex it is attached to.

37. Couldn't find output plot from WhatIf!!
38. The DSSP program is inaccessible to the WhatIf package. Therefore WhatIf inferior secondary structure prediction program was run. This showed that the secondary structure elements from the BT structure were conserved.
39. The analysis reported 24 abnormally short interatomic distances. None of the clashes involved both atoms being part of the polypeptide backbone. This means that the backbone does not clash with itself at any point. For the worst clashes it would be a simple procedure to examine the structure around the clash and manually alter the cause of the clash. This would cause minimal disruption to the model.
40. 24 residues were reported to having unusual packing environments. A slight improvement on the original model. Reasons for poor packing score are: poor packing, misthreading, sequence misaligned, crystal contacts (not applicable in this case), contacts with a co-factor (not applicable), or the residue is part of the active site. The main reasons for the poor packing score in this model will be poor packing and misalignment of the sequences. Comparative / Homology modelling are dependent on the fact that the sequence alignment between the known structure and the sequence of the model is correct. If the alignment is wrong it is guaranteed the model will be wrong for the misaligned section. As the homology between the two sequences was so low then there is likely to be sections in the alignment of the two sequences that are wrong. Some of the poor packing is also likely to be some residues having the wrong conformation after non conservative substitution.



41. The only section in the model where the packing environment was abnormal for three consecutive residues is in the middle of one of the loops. It is not unusual for loops to be have unusual conformations so the warning is not too severe.
42. The average quality control value for the structure is reported to be very low. The value given of -2.749 is acceptable for low homology models which normally have the range of -2.00 to -3.00.
43. Couldn't find quality value plot !!
44. A low packing z-score was reported for 16 residues. A score below -2.5 warrants having a look at the structure as they are "unusual". This was not possible due o the time constraints.
45. 3 sections of at least four residues had a second generation packing z-score below -1.75. Reasons are misthreading or part of a strange loop. The first stretch between Tyr 24 and Gln 27 is the only real problem section as the other two sections that were reported are part of modelled insertions and are therefore part of surface loops.
46. The abnormally low structural average packing z-score is worrying. It is acceptable for low homology models to fall in the range -2.0 to -5.0. The value reported for Bb is an overall -6.79. This is likely to be due to the very low homology between the BT and Bb.
47. Couldn't find the plot output from WhatIf !!
48. There is a warning in this section on backbone oxygen evaluation. The only report was for a Gly residue. As it is normal to get a few Gly residues in this section it is not a worry that the analysis routine pulled up one Gly residue from the model.
49. Only two residues were reported to having unusual rotamers. It is not necessarily an error if a few residues have a rotamer value below 0.3. No residues in the refined model had a value below .03.
50. There were 126 unusual backbone conformations reported. Two reasons for the unusual backbone conformations will be the constraints on the  $\phi, \psi$  torsion angle during the minimisation causing problems with the  $\Omega$  torsion angle. The other reason will be the

large number of insertions and deletions that had to be carried out while building the model.

51. The low reported z-score for the backbone conformation is due to the reasons given above in point 50
52. The average B-factor problems is a crystallographic check. This can be ignored for this model.
53. Again the B-factor plot is ignored because there is no crystallographic data in the model.
54. The report shows that 5 residues could have part of their terminal group on the side chain flipped to give a better H-bonding pattern. This is easy to carry out causing only local disruption. Unfortunately time restraints prevent further refinement from being carried out.
55. The rms Z-scores are much higher than 1 for the His residues reported. This suggests that the geometric assignment given here does not correspond to the type used in the refinement.
56. A total of 84 buried hydrogen bond donors are not involved in a hydrogen bond. This is a result of the many alterations to the starting structure to build the final model. All the alterations have severely disrupted the hydrogen bond network in the core of the protein.
57. There are only 8 buried hydrogen bond acceptors in the model that are not involved in a hydrogen bond. Again the reason for there being any at all is because of the numerous alterations to the starting structure which has disrupted the hydrogen bond network at the core of the model structure.
58. The reason for the poor results here is the fact that this is a very low homology model. The sequence homology between the starting structure, bovine trypsin, and the unknown structure Bb is very low. This means that there are likely to be misaligned sections in the sequence alignment, and the large number of non conservative substitutions, deletions and insertions introduce many more errors in to the model.

### 7.1.2 A Look At The Model

An examination of a three dimensional image of the backbone of the Bb model shows that the protein fold has remained stable during the energy minimisation procedure, see Diagram 7-8 on page 234. The secondary structure has remained stable even though the model was highly strained at the start of the energy minimisation procedure due to the large number of substitutions, deletions and insertions modelled.

The relative positions of the seven large insertions in the structure of Bb can be seen in Diagram 7-9 on page 234. All the insertions occur on the surface of the protein model and remain on the surface during the energy minimisation procedure. The insertion at position 170 where the first five residues of the insertion were modelled as  $\alpha$  helix can be seen in the top right hand corner of this diagram. It is obvious from the ribbon plot that these five residues have remained in the  $\alpha$  helix structure throughout the energy minimisations procedure.

As mentioned in Chapter 5, while discussing the results of the insertions at position 129 and 231, it was noted that the two insertions were very close to each other in the three dimensional structure of the model. This can be clearly seen in Diagram 7-10 on page 235. In the diagram it is clear that the ends of the insertions are interacting with each other.

Looking at the acidic and basic residues (Asp, Glu, Lys, Arg, His) all the internal acidic and basic residues form salt bridges with each other. The remaining acidic and basic residues are all found on the surface of the protein (see diagram Diagram 7-11 on page 236). The only exceptions are Glu 70, Lys 160, Arg 153 and a basic residue which is part of Ins 230. Examination of the model around Glu 70 does not show any obvious potential salt bridge. The only H-bond Glu 70 is able to form is with the  $N_{amide}$  of Gly 44, Diagram 7-12 on page 237). This will relieve some of the energy of having a buried acidic group in the model but the Glu 70 is still in a high energy position. For Lys 160 the situation is slightly better as it forms a favourable H-bond with Asn 189. Again this is not ideal as there will still be some energy strain due to having a buried basic group, Diagram 7-13 on page 238. Finally for Arg 153, it forms two hydrogen bonds. The first to the  $O_{carbonyl}$  of Cys 136, and a second to the O carbonyl of Pro from Ins 186. For each of these buried acidic

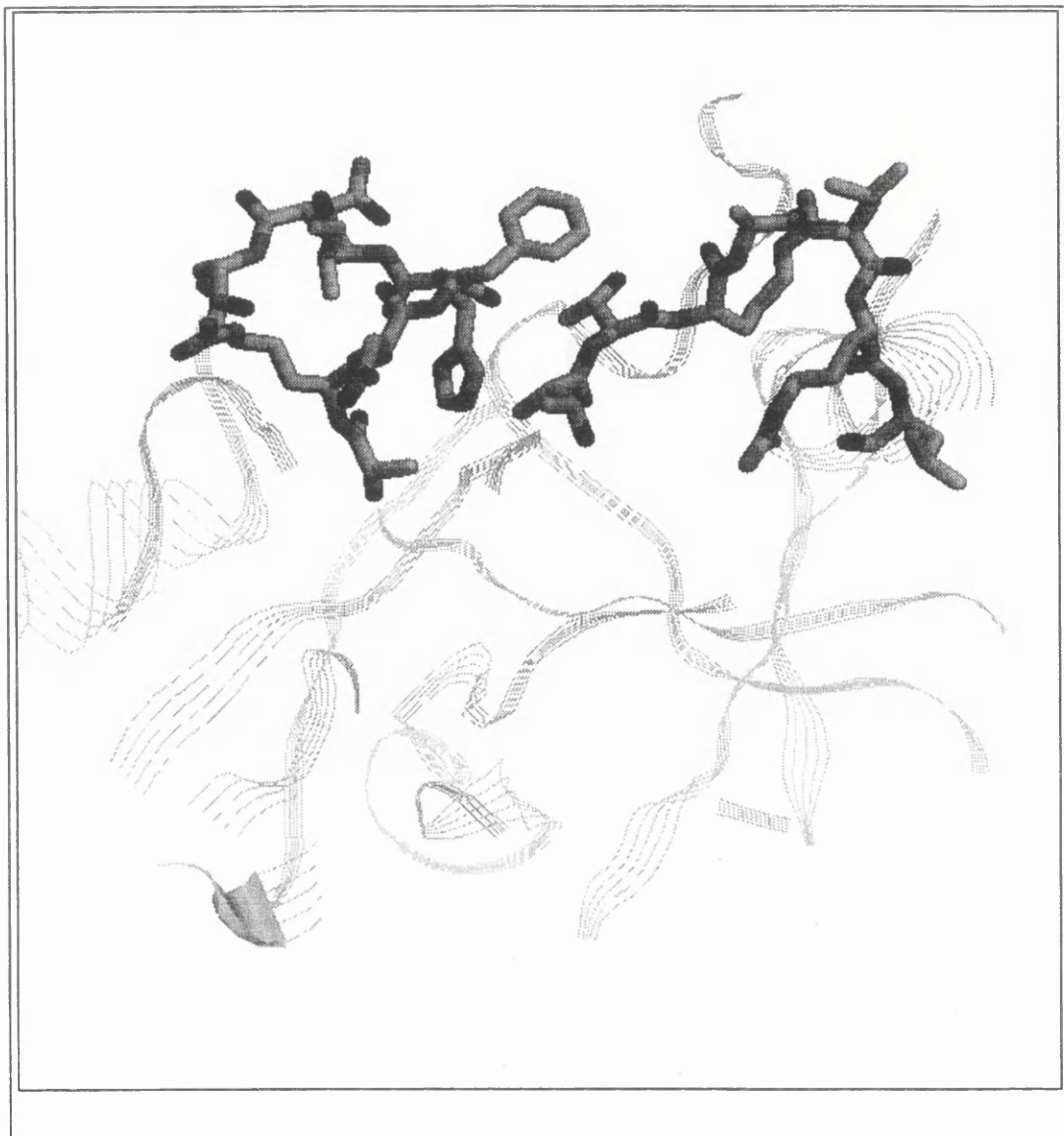
and basic groups there will be considerable energy for not being part of a salt bridge, but some energy is relieved by forming H-bonds with nearby potential sites, see Diagram 7-14 on page 239.



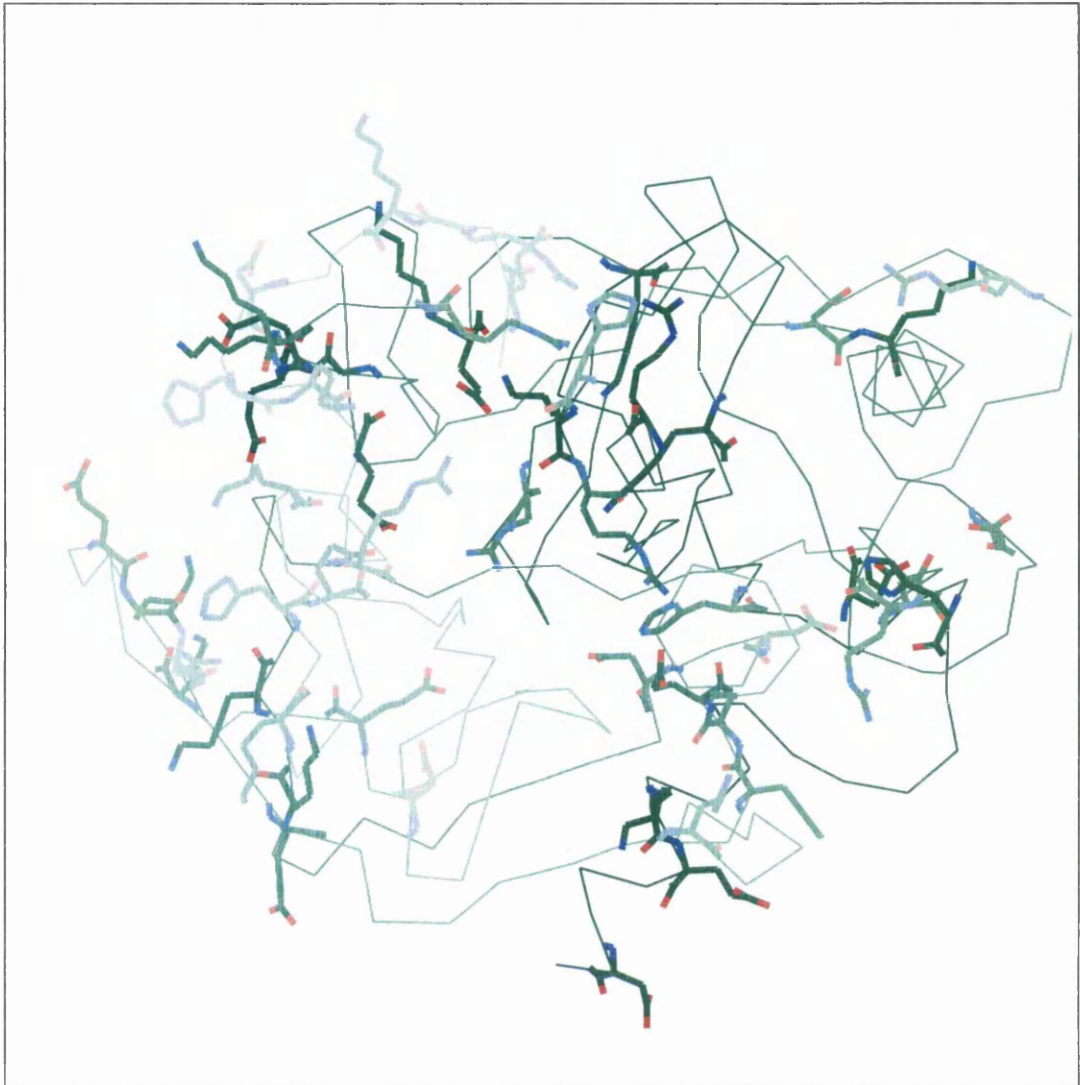
**Diagram 7-8: A ribbon drawing tracing the  $C_\alpha$  of each residue. It shows the  $\alpha$  helices and sections of  $\beta$  strands that have remained stable during the energy minimisation procedure.**



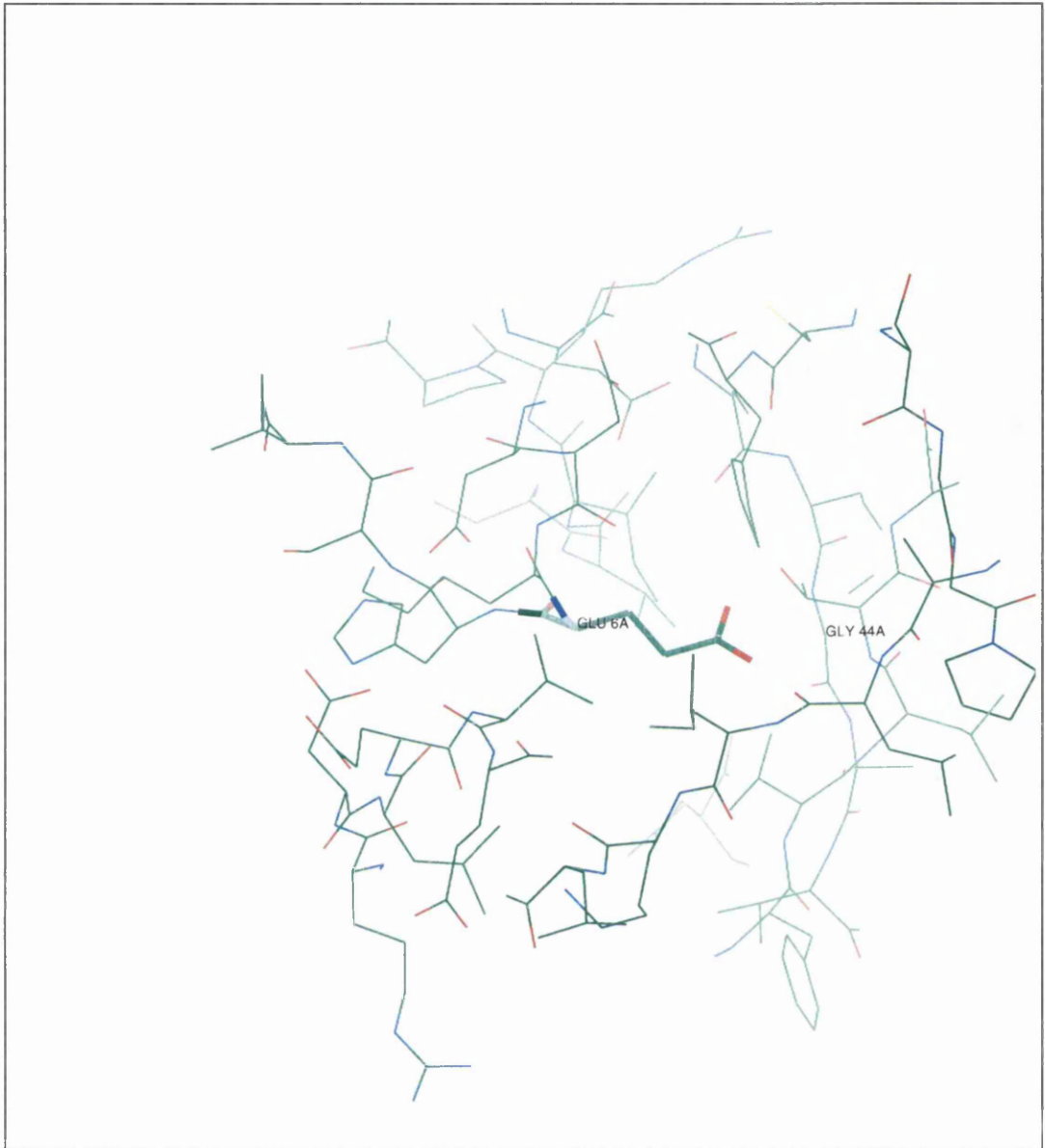
**Diagram 7-9: A ribbon representation of the backbone of the Bb model. The 7 large insertions are shown in black.**



**Diagram 7-10: The two insertions at 129 and 231 are shown in stick representation. The surrounding backbone is shown in a ribbon representation. The diagram shows how close the two insertions are in the three dimensional structure of the protein.**

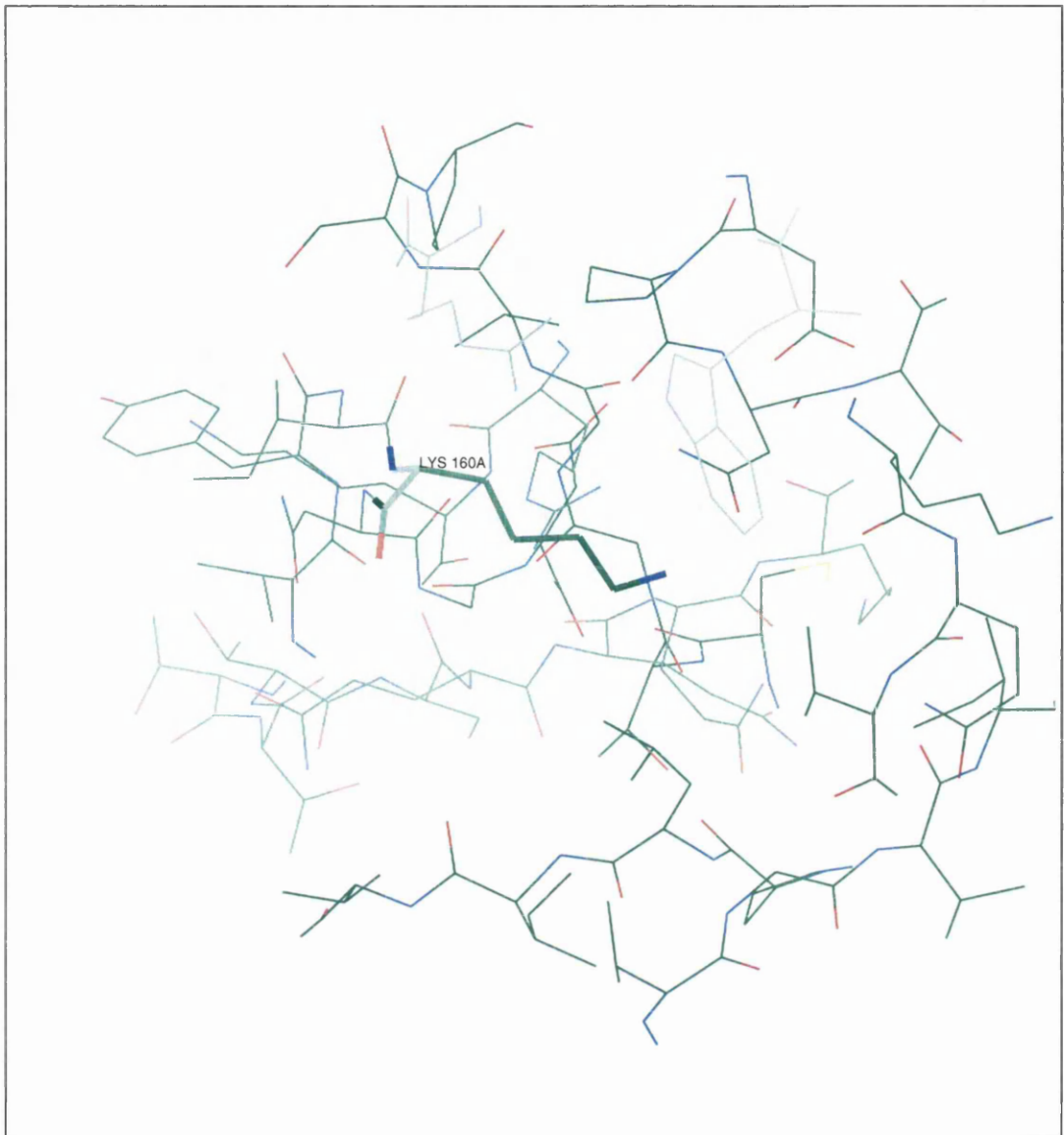


**Diagram 7-11: View of all acidic and basic (Asp, Glu, Lys, Arg, His) residues not involved in a salt bridge. There is also a Calpha trace.**

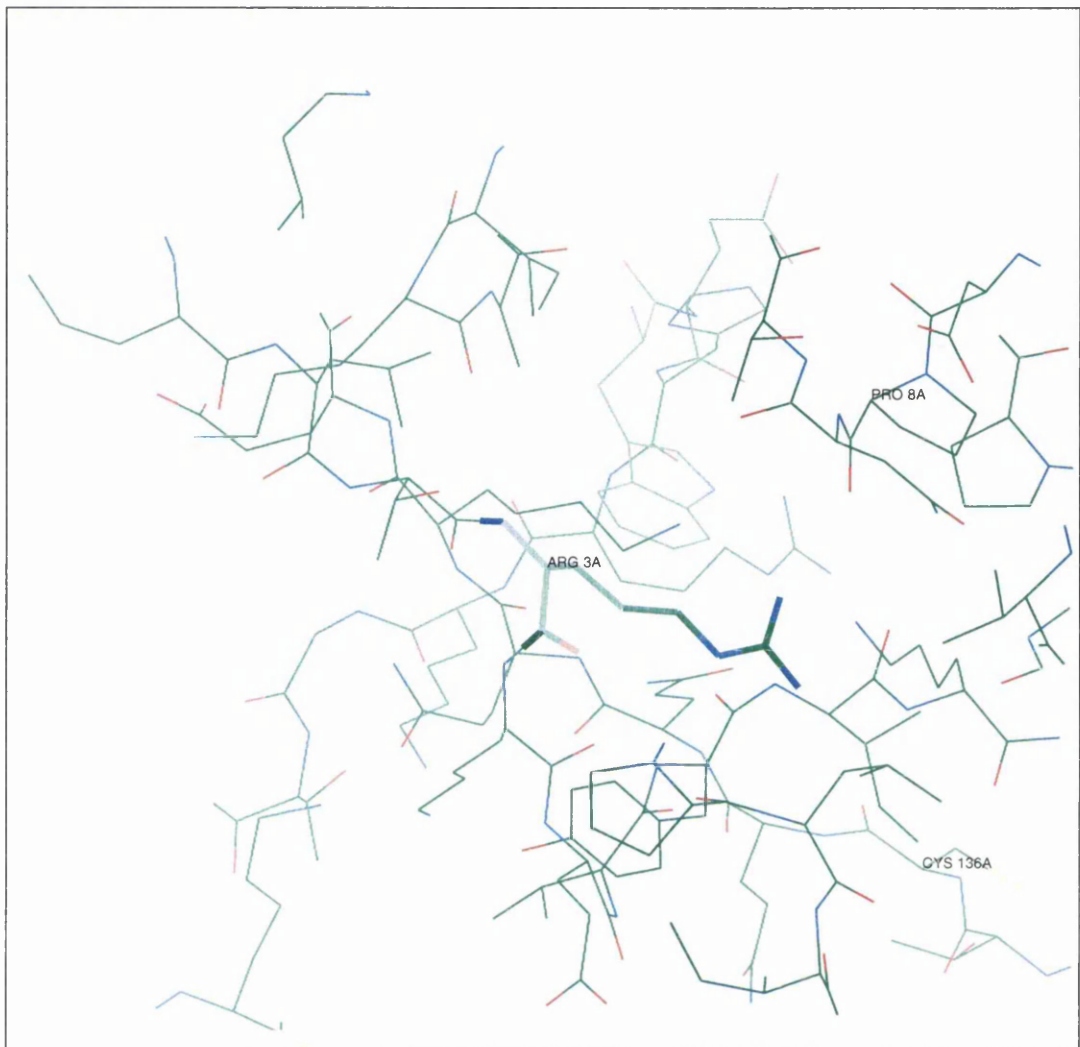


**Diagram 7-12: The residues around Glu 70 are shown. Gly 44 is labeled. Glu 70 forms hydrogen bond with Namide of Gly 44.**





**Diagram 7-13: This shows the residues around Lys 160. Asn 189 forms a strong H-bond with Lys 160.**



**Diagram 7-14: This shows the residues around Arg 153. Arg 153 forms H-bonds with Cys 136 and Pro 8 which are also labeled in the diagram.**

### 7.1.3 Discussion on the Active Site

Bb is an active serine protease so it was important that the three residues His 57, Asp 102 and Ser 195 are kept in the same spatial arrangement relative to each other. The primary sequence of Bb around each of these three residues are the most homologous sections of the alignment of Bb with bovine trypsin. The corresponding sections of alignment between Bb and bovine trypsin's are given below:

57

TB	W	V	V	S	A	A	H	C	Y
Bb	F	V	L	T	A	A	H	C	F
	*		*	*					*

102

TB	N	N	D	I	M	L	I	K	L	K
Bb	D	Y	D	V	A	L	I	K	L	K
				*						

195

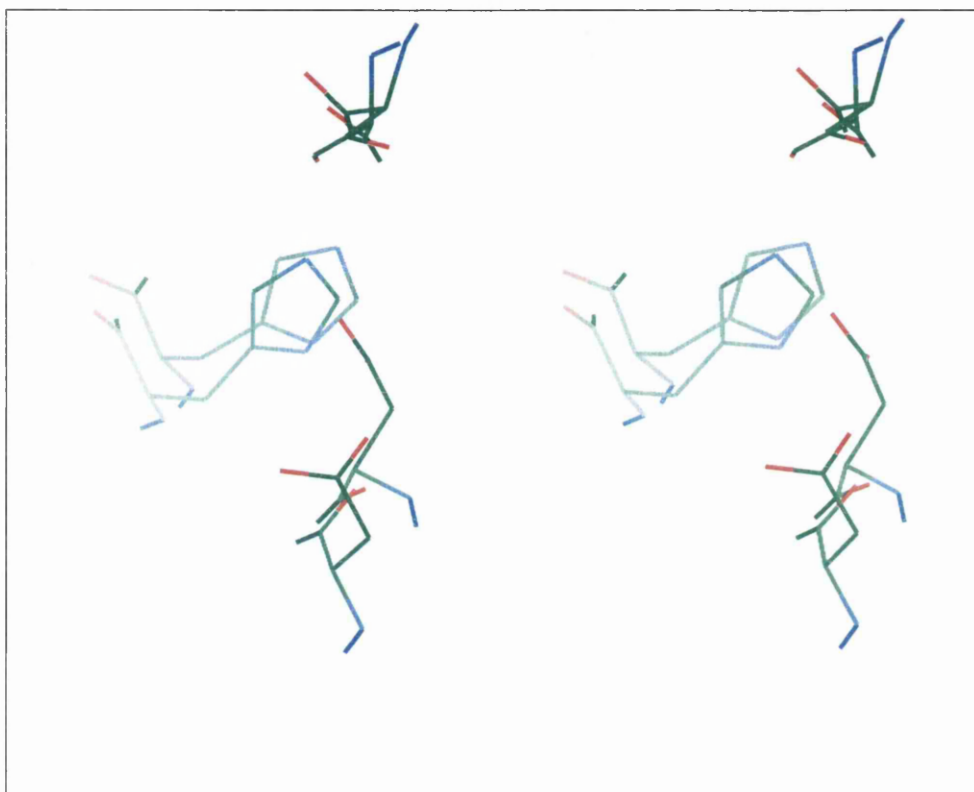
TB	S	C	Q	G	D	S	G	G	P	V	V
Bb	T	C	R	G	D	S	G	G	P	L	I
	*									*	*

The His and Ser residues of the of the three catalytic residues are near the middle of a conserved region in the alignment. The fact that His 57 is next to a preserved cysteine residue also gives an indication of how important it is that the three dimensional structure of this region does not change. With both the His and Ser residues being near the centre of the conserved region of primary structure it strongly backs up the case that these two residues will be the same spatial position in the model of Bb as is found in the structure of bovine trypsin.

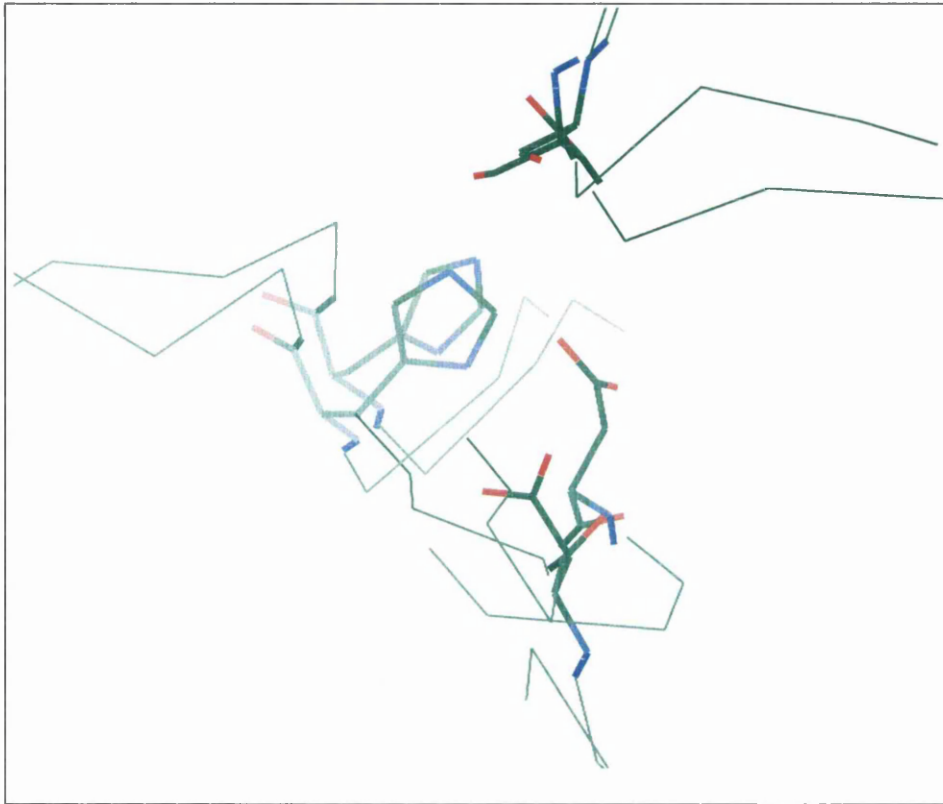
For the third residue in the catalytic triad, Asp 102, there is a large insertion just one position away in the original alignment between Bb and bovine trypsin. Looking at the structural environment of the Asp 102 in the bovine trypsin structure it is clear that the residue is in a  $\beta$  sheet. In fact the insertion at position 101 is positioned in this section of secondary structure in the original alignment. Two facts combine to bring one to the conclusion that the sequences are misaligned in this section. The first is the fact that Bb is an active serine protease and therefore Asp 102 must be found at the same three dimensional position as it is found in all other known structures of active serine proteases. The second fact is that large insertions do not occur in segments of secondary structure. The insertion is moved in the alignment to where the  $\beta$  turn in the bovine trypsin structure

is, allowing Asp 102 to be far enough away from the insertion so that it is unaffected by the loop in the conformation it adopts in the Bb model.

In the final minimised model the three catalytic residues can be superimposed onto the catalytic residues from the starting bovine trypsin structure with a high degree of overlap, see Diagram 7-15 on page 241 and Diagram 7-16 on page 242.



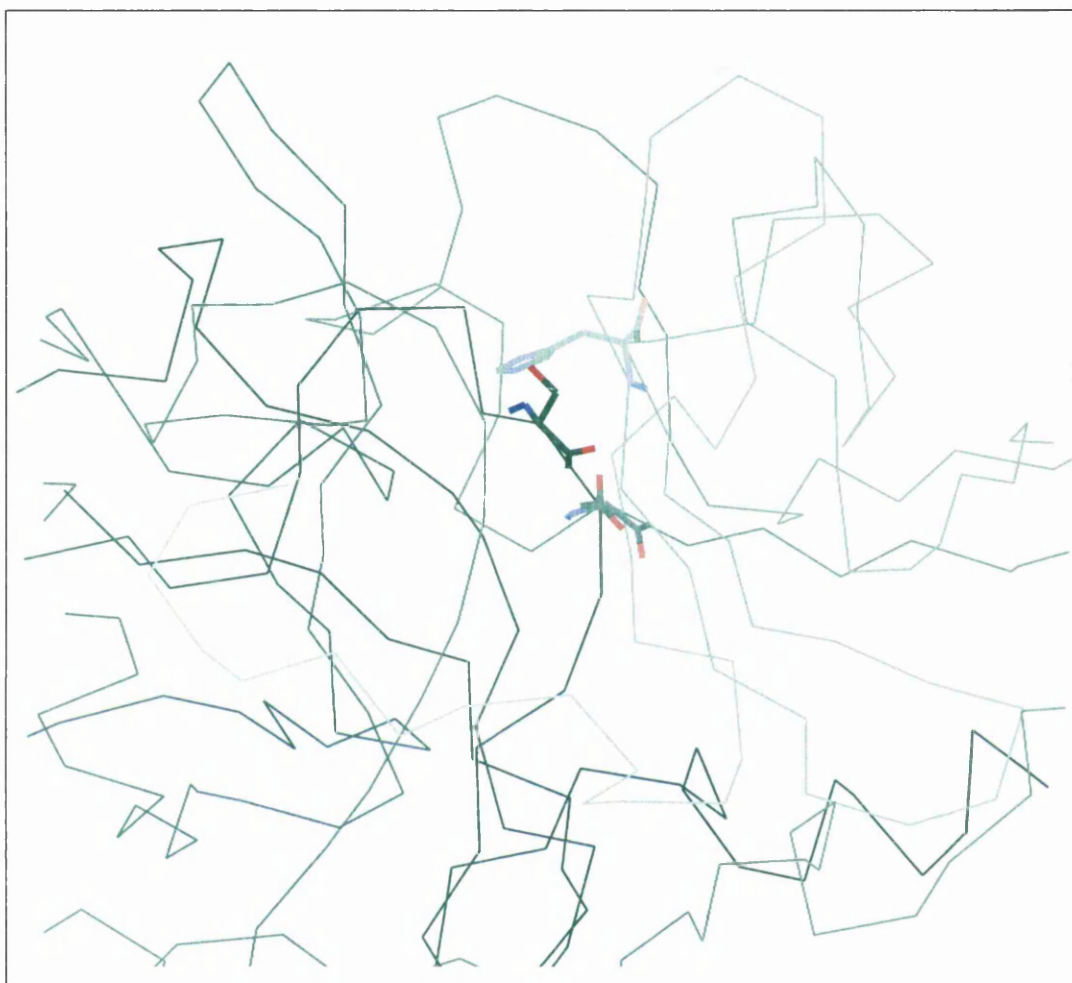
**Diagram 7-15: Stereo view of the active site of Bb superimposed onto the active site of the starting structure, bovine trypsin.**



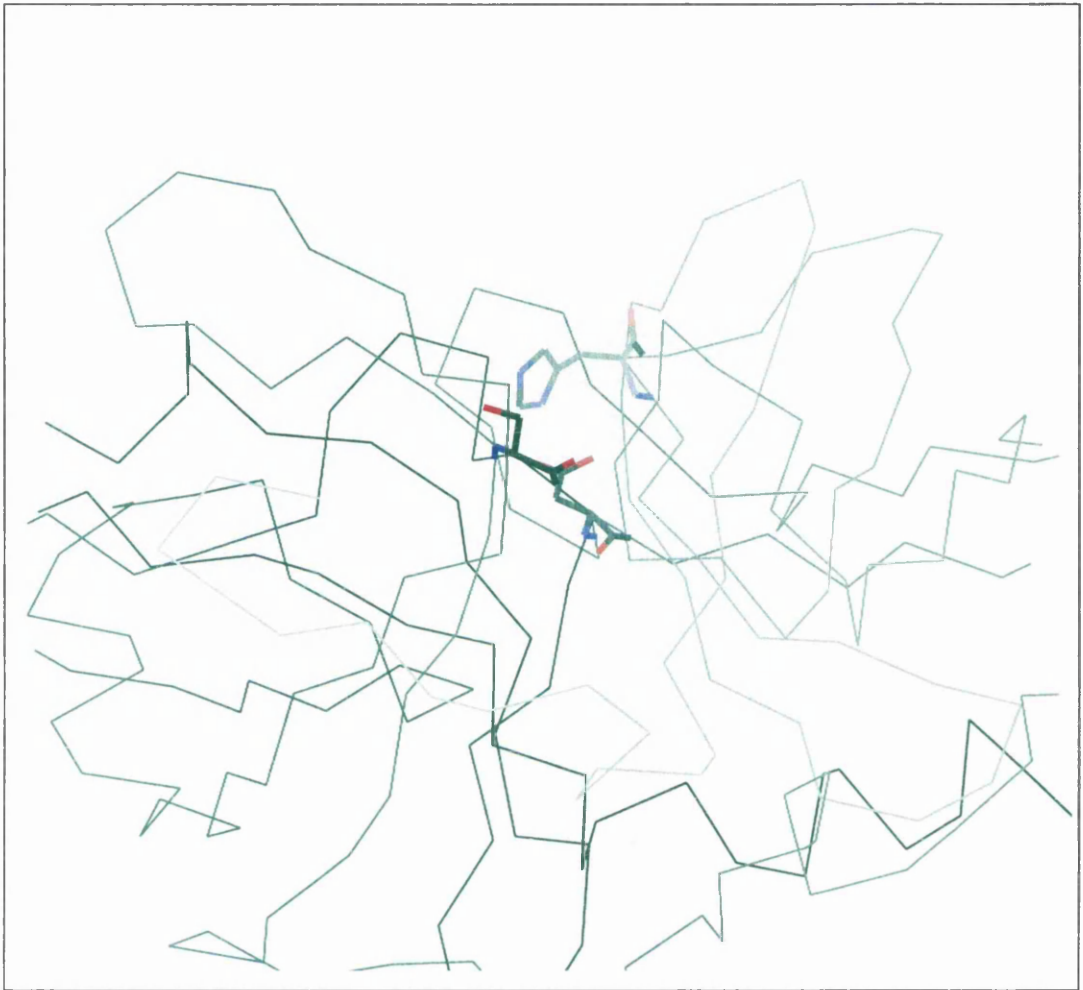
**Diagram 7-16: Active site of the Bb model superimposed onto the active site of bovine trypsin. The backbone of both structures is shown several residues either side of each residue of the active site.**

A final point about the active site is that in Bb some of the inserted loop appear to extend around the active site. This can be seen in Diagram 7-17 on page 243. This has the effect of burying the active site in a deeper valley. This would make the active site less accessible to substrates as they had to work their way down into the valley to reach the active site. Having the active site in the valley would also make Bb more specific as there would be plenty of potential enzyme - substrate interactions with the substrate protein and the walls of the valley. This seems to support the fact that Bb has only one substrate in nature, that is to be part of the C3 / C5 convertase in the Alternative Pathway of the Complement System and activate C3 or C5. The exact same view for TB can be seen in Diagram 7-18

on page 244. Here it is obvious that the active site is much more exposed than in Bb. This allows easy access to many different protein substrates. This agrees with TB role in the body as a general digestive enzyme.



**Diagram 7-17: A view along the catalytic triad of Bb with the Calpha trace shown in green. It shows the insertions extend a good distance past the active site causing the active site to be more buried than in TB.**



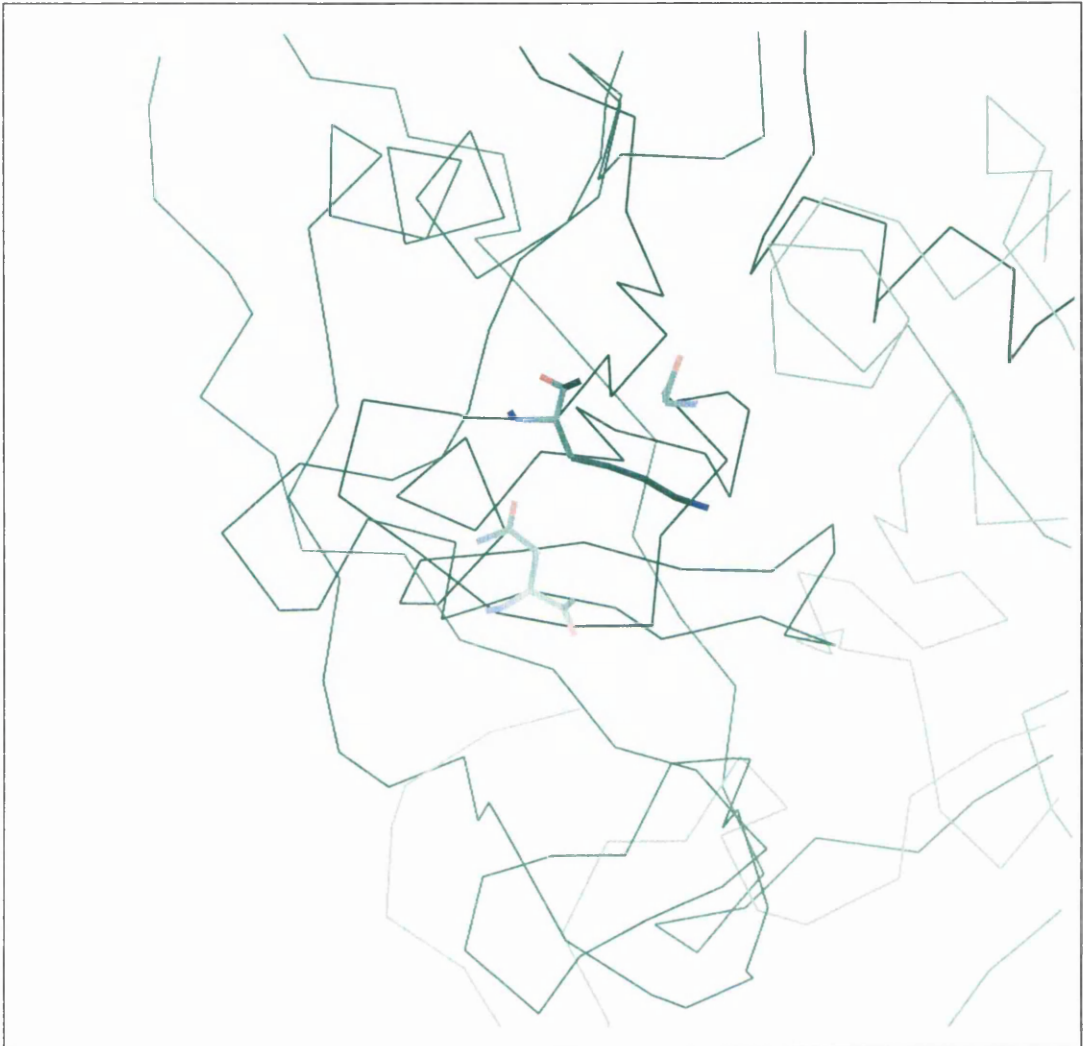
**Diagram 7-18: The same view for TB looking along the catalytic triad, with Calpha trace in green. Shows the active site much more exposed than for Bb.**

#### 7.1.4 The $S_1$ Binding Site

In serine proteases there is a well defined substrate binding site, called  $S_1$  which strongly influences the specificity of the substrate that the serine protease will cleave. The  $S_1$  site consists of a pocket in the surface of the serine protease enzymes. At the bottom of the pocket is position 189 whose side chain extends into the pocket and can influence which type of residue is favourably bound in the pocket. At the entrance to the pocket, at either side are residues 216 and 226 which determine how wide the entrance to the pocket is (see 3.2.3.2.1 for a fuller description). In BT the important residues are Aps 189, Gly 216 and

Gly 226. With the Gly residues at the mouth of the pocket the entrance is not blocked allowing the P<sub>1</sub> residue from the substrate access to the pocket. At the bottom of the pocket Asp 189 strongly binds to basic side chains. It is therefore found that TB is strongly specific to having Lys or Arg at P<sub>1</sub>. For Bb the residues that influence the S<sub>1</sub> site are Asn 189, Gly 216 and Lys 226. With the Lys residue at the entrance to the S<sub>1</sub> binding pocket it strongly suggests that the entrance to the pocket will be at best if not completely blocked. Diagram 7-19 on page 246 shows that Lys 226 lies across the entrance to the binding pocket and into the pocket itself.





**Diagram 7-19: Alpha trace of Bb with Asn 189, Gly 216 and Lys 226 shown. Lys 226 can be seen to block the entrance to the S<sub>1</sub> pocket.**

## **7.2 Where Protein Modelling Is At**

### **7.2.1 Why Protein Modelling Is Necessary**

Modelling of protein structures is necessary due to the large number of DNA - derived sequences that are available compared to the number of known three dimensional protein structures. Even with the tremendous increase in the rate at which protein structures are being determined there is still an enormous number of sequences with unknown structure.

This is all due to the molecular biology revolution of the early eighties. Modern molecular biology techniques now make it feasible to isolate and sequence the DNA which encodes for proteins. The entire process of extracting the DNA fragments and reading the code can be fully automated.

### 7.2.2 There Are Only Certain Folding Motifs Proteins Adopt

As more protein structures were solved it became apparent that certain fold patterns were used repeatedly. Proteins with similar function had a similar three dimensional structure. This enabled proteins to be grouped into families of related function and structure<sup>117, 118</sup>. Even some proteins with diverse functions were found to be structurally related. These folds appear more conserved than the primary sequence of the proteins<sup>119,120</sup>. Analysis of the protein folds (motifs) has shown that there could be a limit to the number of motifs that proteins adopt. The number of non homologous motifs could be as low as one thousand distinct motifs<sup>121</sup>. Already it has been determined that there are over 100 distinct non homologous motifs in the Brookhaven database<sup>122</sup>.

The reason for the distinct number of motifs available to proteins comes from the nature of the amino acid residues themselves. All the twenty naturally occurring amino acid residues, except for Gly, are chiral and all exist in nature in one enantiomer. This means that there is a difference in energy when there is a right handed join between two sections of secondary structure joined by a small number of residues to a left handed join e.g. the join between an  $\alpha$  helix and a  $\beta$  sheet. It also appears that the motifs found in protein structures are able to be stabilised by many different random sequences, whereas the other possible folds that a polypeptide sequence could adopt seem only to be stabilised by a few specific sequences<sup>123</sup>.

Modelling techniques take advantage of the fact that proteins only adopt certain folds. In comparative modelling a template from a known three dimensional structure of the same motif as the unknown sequence is known to adopt is used as a starting model. Threading methods take a sequence of unknown structure and try and find which motif the unknown best fits. It aligns the unknown's sequence to the three dimensional structures. The other main technique is useful if the sequence with unknown structure can not be successfully

fitted onto one of the known motifs. Genetic algorithms offer a hopeful *ab initio* method to fold protein sequences using the same optimisation procedures as natural genetic evolution.

## 7.2.3 Modelling Techniques

### 7.2.3.1 Comparative Modelling

Once a protein's sequence has been determined and it has been found to be a new member of a structurally characterised protein family, it is relatively straight forward to build a molecular model of the protein using a set of simple guidelines<sup>124, 125</sup>. The first step is to determine the structurally conserved regions and variable regions in the protein family the unknown belongs to. This can be carried out by looking at the known three dimensional structures of members of the family. The superposition of the known three dimensional structures of the family will clearly show the regions in the structure that are conserved and regions that have variable structures. Having as many three dimensional structures as possible in this process greatly improves in defining the structurally conserved regions and the variable regions. The next stage is to align the amino acid sequence of the family. The alignment is not carried out by the normal criteria for sequence alignment, that is sequence homology, it is strictly done on the basis of the structural information of the structurally conserved and the variable regions. The unknown is then aligned using the patterns found in the structurally conserved regions of the sequence alignments. It is essential for the alignment of the new sequence to be correct. If the alignment of the sequence is incorrect then the model will definitely be incorrect for the misaligned section of sequence. This is one of the reasons that the comparative modelling method currently depends so heavily on the retention of sequence homology.

For the structurally conserved regions the co-ordinates from one of the known structures can be taken. The amino acid residues are substituted and the side chains altered so that the section has the correct sequence to that of the unknown. This substitution procedure does not involve any alteration to the backbone atom co-ordinates. The variable regions are usually modelled by a database search, conformational search or some sort of molecular dynamics.

### **7.2.3.2 Threading**

A new development over the past two or three years in protein structure modelling has been the approach of identifying entire protein folds from the amino acid sequence<sup>126</sup>. This approach employs techniques for aligning sequences with three dimensional structures, known as threading, which are used to select the native folds of a given sequence from a database of alternatives<sup>127, 128</sup>. This method is very much dependent on the fact that proteins with very different sequence and functions can adopt very similar folds, and, moreover that known folds fall into a limited set of families<sup>129</sup>.

Structure is conserved to a greater extent than sequence and threading may extend the range of molecular modelling to sequences which show little or no sequence similarity to proteins with known structure.

#### **7.2.3.2.1 Motifs as linear Profiles**

It is possible to represent folding motifs as linear profiles of local environment properties, such as solvent accessibility<sup>130</sup> and secondary structure<sup>131</sup> at each residue site in a protein of known structure. These local environment properties are assumed to remain the same in any protein which folds into a given motif. A residue preference for a particular environment can be calculated by examining the family of proteins where at least one member of the family has a known three dimensional structure. Representing the motif in this manner has the attraction of simplicity. Optimal threads, or alignments, of sequence and folding motif may be found using the dynamic programming algorithms which have been well developed and long used in sequence alignment.

#### **7.2.3.2.2 Motifs as Contact Matrices**

Several other groups have chosen to represent folding motifs as a two dimensional contact matrix<sup>132,133</sup>. This representation defines the spatial relationship of residue sites e.g. that position 2 in the primary sequence is close in three dimensional space to position 110 in the primary sequence. When a sequence is threaded through a folding motif, the identities of each position are assigned. An evaluation score, or appropriate conformational energy is then calculated by summing up the energies of all the contact pairs. The lookup tables

defining these contact energies are the evaluation. These energies have been empirically derived by analysing the pattern of pairwise contacts in known protein structures. The intentions are that the contact energies will represent entropic and solvation effects that are crucial in protein folding, but which can be represented in conventional molecular mechanic calculations only by lengthy simulations.

The contact energies are specific enough to distinguish correct sequences and alignments among the many million alternatives. It oddly appears rather too easy to find the best alignments of sequences and core folding motifs, even though this represents a formidable combinatorial optimisation problem. The answer appears to be that there are always few “misthreads” that have favourable contact energies<sup>134</sup>.

Once the core motif region has been aligned and the co-ordinates of the backbone for the sequence set the problem becomes much like the comparative modelling technique to solve. The side chain atoms must be set and any loop regions (insertion) must have their conformation calculated. The same methods of loop conformation generation are used: Molecular dynamic simulations, database searches, and conformation searches. These methods are all employed to calculate the loop conformation.

### **7.2.3.3 Genetic Algorithm**

Genetic algorithm methods are so called because they utilise the same optimisation procedures as natural genetic evolution: mutation, crossover and replication operations on strings<sup>135</sup>. This is a fairly recent development to the protein folding problem and is still in its infancy, but it is quickly gaining in popularity. In genetic algorithms a population of current solutions is maintained. The solutions evolve by mutations and crossovers. The latter process, that is crossovers, is the heart of the method. Technically the operation consists of exchanging parts of strings between pairs of the solution, so as to yield new solutions. This has a large impact on the effectiveness of the search, since it allows exploration of regions of the search space not accessible to either of the two “parent” solutions. Through such interactions, good features from one solution can be transferred to the other solutions and explored. The population size is maintained by pruning, using certain criteria of fitness for each solution in such a way that better solutions have a higher

chance of reproducing. The diversity of the population is maintained to allow for a large sampling of individual solutions so that many combined features may emerge. Experience with other “co-operative” problem solving methods has shown this feature of exchange of information between solutions is often a powerful way of extending the effectiveness of a search.

The common method of introducing mutations into the structure sequence in the genetic algorithm is an extension of the more familiar Monte Carlo methods<sup>136</sup>. A population of evolving conformations is maintained. Each conformation changes independently for some time by the Metropolis Monte Carlo<sup>137</sup> procedure in the usual manner, in a process equivalent to the accumulation of point mutations. Then selected polypeptide chains are cut and rejoined to another chain cut at the same point. This action is the crossover mutation in the genetic algorithm. Metropolis style criteria are used to see if each newly generated conformation should be accepted. Those that are accepted are used as the next generation of the population and enter the Monte Carlo phase again and the process is reiterated. The simple genetic algorithms that have been developed are dramatically more effective in searching than Monte Carlo methods alone.

As mentioned earlier this genetic algorithm method is still a fairly recent application to the problem of protein folding<sup>138</sup>. Using only simple criteria the results look fairly promising with the algorithm able to identify the main chain topology the protein will adopt<sup>139</sup>. The algorithm is not dependent on the size of the known structure database like the other two modelling techniques discussed, where the fold the unknown protein structure is a fold pattern that must already exist in the structural database. The genetic algorithm is very much a *de novo* protein folding algorithm<sup>140</sup> which is producing some success.

#### 7.2.4 Comparison Of The Modelling Techniques

Currently, comparative modelling / homology modelling is the most successful at producing reasonable structures of proteins. The reason being that it relies the most on information from the structural database to build the model. But because it relies so much on information from the structural database it is the modelling technique which can be used to build the three dimensional models of the most limited set of known protein structures.

Comparative modelling depends on the fact that the unknown sequence is a member of a family of proteins where at least one member of that family has a known three dimensional structure. The sequence alignment of the unknown sequence and the sequence of the protein with known three dimensional structure is crucial in getting the comparative model correct. Any misalignment of the two sequences guarantees the structure of the model will be wrong for the misaligned residues<sup>141</sup>. This dependence on getting the sequence alignment correct is not a problem where the sequence of unknown three dimensional structure is highly homologous to the known structure's primary sequence. With a high sequence identity between the two primary sequences (>70% sequence identity) then there will be no or few problems with getting the alignment of the sequences correct, hence a reasonable model can be expected to be built. If on the other hand the identity (and homology) of the sequences is much less (<30% sequence identity) then the alignment algorithm normally used (usually some implementation of the Needleman and Wunch algorithm) is likely to produce an alignment with sections of the sequences misaligned. Even after manual adjustments to the alignment, or carrying out multiple sequence alignment, there will still be regions of the alignment that are uncertain.

Threading algorithms can be used if the unknown structure's primary sequence is not significantly homologous to any known three dimensional structure's sequence. Threading algorithms use the fact that structure folds and motifs are more conserved than the primary sequence. They are also highly dependent on the fact that there is a limited number of distinct, non-homologous motifs found in natural proteins and that as more classes of motifs are crystallised the number of unknown classes decreases. Threading algorithms use a variant of the alignment routines to align the sequence to different motifs. The threading algorithms so far can match the correct motif to a sequence but then the same methods used in comparative modelling (substitutions, deletions and insertions) are required to complete the model. At the moment threading algorithms offer a more sensitive alignment for a sequence that has a very low similarity to other sequences of known structure but other techniques are required to build the model from the alignment.

The genetic algorithm method is a highly promising *ab initio* modelling technique. It has shown some promising results using simplistic criteria. It is not dependent on finding some

homologous sequence of similar fold in the three dimensional structure database so can be used to model novel classes of proteins that have not yet had a member of that family of proteins been crystallised. At the moment it is still being developed and can not yet offer high resolution models. To date the genetic algorithm method still only produces models of low resolution.

Although comparative modelling / homology modelling is the most successful and widely used modelling technique it does have severe shortcomings that still have to be improved. The hardest of these is the modelling techniques used to determine the conformation of a loop / many residue insertion. There are a few methods that can be used for loop modelling each with their own pitfalls. The modelling of loops can be considered as a mini protein folding problem. At an insertion the unknown structure deviates from the starting model, so the starting model is of no use unless the loop sequence of the loop in the unknown structure's sequence is homologous to the sequence of the loop in the known structure, but this is rarely the case. If the model sequence is a member of a class with numerous proteins in that class with known structure, then these structures' sequences can be aligned and if a loop segment in the unknown is homologous to one of the loops in an known structure this known loop structure can be used in the model. This method can only be used in a few cases where there are many known structures in the protein family i.e. the serine proteases. This still leaves the majority of model loops where a structure of the loop has to be determined. Even for protein families with many known structures only a very few loop will be modelled directly from existing loops found in one of the known structures of the protein family.

## 7.2.5 Methods Used To Model Loop Regions

### 7.2.5.1 Database Searches

Database search methods search the database of known structures (Brookhaven Database<sup>142</sup>) for a segment of a structure that will fit into the space for the insertion. For this to be accomplished at least three of the residues from the conserved region of the model have to be included in the loop so that the new segment can be properly aligned with the rest of the structure. The new segment's sequence must be homologous to the loop



region sequence being modelled. Once a suitable loop structure been found from the structural database it is attached to the model and its residues substituted so that its sequence matches the sequence of the model being built.

#### **7.2.5.2 High Temperature Molecular Dynamics Simulated Annealing**

The idea with high temperature molecular dynamics simulated annealing is that the high temperature molecular dynamics run will produce many different conformations covering a wide range of conformational space for the loop being modelled. This is achieved by carrying out the molecular dynamics at a very high temperature. After so many iterations the current conformation is used as the starting point for simulated annealing of the loop. The loop will be in a strained high energy conformation due to the high temperature that the original molecular dynamics was carried out at. Having the molecular dynamics run at such a high temperature gives the loop sufficient thermal energy to jump from conformation to conformation without getting trapped into a local minima. Running the molecular dynamics for long enough and at a high enough temperature it is hoped the loop will be able to reach every conformation theoretically feasible for it to adopt. Taking conformations at time intervals throughout the molecular dynamics run there should be a spread of conformations for the loop. These act as the starting positions for the simulated annealing part of the method. During the simulated annealing the temperature is slowly cooled down to 0 K while the molecular dynamic simulation is continued.

#### **7.2.5.3 Monte Carlo Simulated Annealing**

This is a similar method to the molecular dynamics simulated annealing except the changes in the conformation are determined by the Metropolis Monte Carlo algorithm. The original conformation is randomly generated. Then a random segment of the chain is altered. If the new energy of the loop is of lower energy than the starting conformation it is automatically accepted. If the energy is higher than the starting conformation then the conformation is accepted on a probability proportional to the difference in energy between the old and new conformations and some “temperature” factor. The temperature factor is slowly cooled (reduced) which means that over time of the simulation run there is less chance of accepting the new conformation if it is of higher energy than the previous conformation.

After cooling a new conformation is randomly generated and the cooling method is started again. After so many attempts there should be a consensus from the different runs what is the lowest energy conformation for the loop.

#### **7.2.5.4 Conformation Searching**

This has previously been described in detail earlier in Chapter 4. Conformational searching used by other researchers have a similar methodology to that described.

#### **7.2.6 Comparison Of The Loop Modelling Techniques**

Each of the methods described above have serious short comings. The simulated annealing methods are computationally intensive but worse they do not guarantee to examine every conformation. The annealing process can be guaranteed to find the global minimum but only over a very long time scale and with a very slow and controlled rate of cooling. They did not produce many promising results and appear to have fallen from popularity over the last few years. This may change if a genetic algorithm method is developed for the smaller case of insertions rather than for the entire protein folding problem which was discussed earlier.

Database methods assume that the set of known structures is large enough to contain adequately accurate conformations of all short segments of chains that occur in proteins. This is true for short segments of chain up to seven residues in length, when an overall low root mean square deviation (r.m.s.d.) is the criterion for similarity. However as longer lengths are considered the percentage of unobserved structures rises rapidly. For the database of known three dimensional structures of proteins segments longer than seven residues are not well represented. Increased database size will reduce the fraction of missing conformations, but even for eight residue segments a much larger database will be required than there is at present. There is no prospect of obtaining an adequate database for lengths longer than that<sup>143</sup>.

From the viewpoint of comparative modelling / homology modelling, there are serious problems in using the database to select conformations for even short lengths of chain. There are two reasons for this:

1. In order to align the new segment with the rest of the structure, at least three additional residues must be included. For seven residue segments it is estimated only ~4% of structures are absent from the database. But for eight residues this rises to at least 30% and is much greater for nine residues. Therefore the information required is not present in the database in a large fraction of cases when building five residue fragments.
2. Even when the segment, including the root, is well represented in the database, aligning with just the root residues produces unacceptable results in a large percentage of cases. This problem is not an artefact of a poor alignment procedure, but rather reflects the fact that obtaining a low overall r.m.s.d. by aligning the ends of a segment with a reference structure is a more stringent requirement than aligning the whole segment and requires a larger database for adequate sampling.

The systematic search methods are guaranteed to sample all feasible conformational space. There is no requirement to sample in an even distribution as much of the  $\phi$   $\psi$  conformation space is disallowed due to steric clashes in the backbone atoms. Unfortunately this still leaves eight or nine  $\phi$   $\psi$  angle pairs in COMMET and 11 pairs in CONGEN<sup>144</sup> for each residue except for Pro and Gly. This means that the number of discrete conformations that have to be generated increases exponentially with the length of the segment being modelled. This does have the advantage that conformation space is adequately sampled for short and long segments, unlike database searches, but the time taken to generate all conformations becomes prohibitive. For CONGEN this is seven residues and for COMMET about ten residues.

To make the time taken to generate all the conformations for long residues acceptable two steps can be taken. Increase of computational power available to run the calculation and secondly decrease the density of sampling. Increasing the computational power is not a particularly elegant solution because the length of time it takes to generate all the conformations grows exponentially. Therefore to increase the number of residues in the segment by one you have to exponentially increase the computing power available to take the same amount of time as the previous calculation. Decreasing the density of sampling of the conformation space required is a better method for speeding up the calculation by

drastically reducing the number of conformations generated by the algorithm. Unfortunately the density of sampling required is dictated by the ability of the available discriminatory functions to identify a correct conformation against the background of incorrect conformations. In future, discriminatory functions less sensitive to errors than the current energy based ones may be developed and these may be better able to identify more approximate solutions.

Although time consuming, systematic conformational searching is more robust at finding the global minimum structure than database searches and simulated annealing techniques. There is a limitation on the length of the segment though before less thorough and less reliable procedures have to be used.

### **7.3 How Computer Generated Models Of Proteins Are Used In Research By The Pharmaceutical Industry**

To pharmaceutical companies involved in research computer generated models of proteins are a great benefit in the drug discovery process. The use of models of the target protein can drastically cut down the time required to develop a new drug, greatly reducing the number of potential drugs candidates that have to be synthesised and tested in the laboratory, and potentially reduce the number and severity of any side effects.

Many drugs are effective by binding to one specific protein in the body and acting as an inhibitor but occasionally agonist or superagonist. The ideal drug candidate is a molecule which binds only to the desired protein. The best drug candidates bind strongly to only one specific site on the target protein. The targeting of the drug to a specific protein reduces the number of side effect that the drug may cause by not having the drug interact with other proteins in the body.

## **7.4 What Use Is The Model Of Bb To The Pharmaceutical Industry**

The key use for the model of Bb will be in rational drug design studies, the reason being it is a key protein in the Alternative Pathway of the Complement System (see section 2.2 on page 35). If the enzyme function of Bb were inhibited then:

1. The Alternative Pathway would not get any further than the spontaneous nucleophilic attack of C3 to C3i.
2. The positive amplification loop which is a central feature to the Alternative pathway would not get started.
3. The Alternative Pathway C3 convertase and C5 convertase would be unable to attack C3 and C5 respectively

The effect of blocking the catalytic function of Bb would be to effectively block the Alternative Pathway.

The beauty of blocking Bb is that the Classical Pathway is unaffected since Bb plays no part in the Classical Pathway. The Classical Pathway continues to function normally ensuring that the body still has a functioning immune system that can respond to antigens detected by the antibodies. The bodies effectiveness to fight off bacterial, viral and fungal infections remains.

The reason behind wanting to block the Alternative Pathway of the complement system is the fact that it is believed that the Alternative Pathway is linked to auto immune diseases. It is believed that the positive amplification loop gets out of control and the immune response can become excessive. Effectively blocking the Alternative Pathway prevents this amplification loop making the immune response milder but hopefully as successful since the Classical Pathway is still active.

## 7.5 Rational Drug Design

### 7.5.1 QSAR

Traditionally drugs were discovered by screening as many different molecules as possible for the desired biological effect. It involved trying to cover as wide a range of compounds as possible and was very much a process of luck. Even when a possible drug candidate was found it was a process of trial and error modifying the original molecule to get an improvement on the biological activity. This whole process was becoming ever more expensive and the time taken to develop a drug to market ever increasing.

A major early development in rational drug design was the introduction of Quantitative Structure - Activity Relationship method (QSAR). This method involves large databases of compounds describing each compound by its chemical structure and activity. The first step in the analysis is to replace the chemical structure by some general description in terms of free energy, and to assume that steric, electronic and solvent dependent properties make linear contributions to that free energy.

In practise it is some pharmacological property  $Q$  (i.e. the inhibitory effect of a compound to the target protein) which varies in a series of compounds which are to various extents analogous compounds. These  $Q$  values can be related to physiochemical or other properties (i.e.  $p, q, r, \dots$ ) of each analogue through a linear equation:

$$Q = a + bp + cq + dr + \dots$$

Where  $a, b, c, d, \dots$  are coefficients. Considering that any one analogue  $i$  has pharmacological property  $Q_i$  and other properties  $p_i, q_i, r_i, \dots$ , then we have a set of equations, one for each analogue, but each with the same coefficients  $a, b, c, \dots$ . These coefficients can be determined as those which give the best fit overall to all the  $Q$  values and the corresponding values of  $p, q, r, \dots$ . This best fit is usually found by a least squares

method. In fact, Q need not be related to p, q, r... through a linear equation. The only demand is that Q be a continuous function of p, q, r... .

Failure to take into account any conformational behaviour can lead to discontinuous relationships between Q and the properties p, q, r... and no curve can be fitted by a least squares or similar method. The promising QSAR method is an approach well known to be the most likely to fail in the case of conformationally flexible compounds.

The QSAR method infers the structural conformation of the binding site by the shape and chemical properties of the analogues used in the study. Using QSAR any interesting or unusual features of the binding pocket can be easily missed unless one of the analogue structures in some way uses the feature. For example, for the QSAR method to detect a polar group in the binding site of the target protein one of two cases must occur in the data set of analogue compounds. A hydrogen bond must form between the specific polar group on the target protein and a polar group of one of the analogue compounds. This will have the effect of increasing the binding of the compound to the target protein and so information about the existence of the polar group will be incorporated into the QSAR model. The other possibility of detecting the polar group in the binding site is if a hydrophobic segment of one of the analogue compounds interacts with the specific polar region. This has the effect of decreasing the binding affinity of that particular compound to the target protein, again this gives information about the existence of the polar group in the binding site of the target protein. If on the other hand none of the analogue compounds in the data set interact with the particular polar group in question then no information about the existence of this polar group will exist in the QSAR model. The effect will be that the polar group in the binding site of the target protein will not be used in designing compounds with improved binding to the target protein. Any new leads using this polar group in the binding site will be missed in the design study using the QSAR methods. Some other screening of compounds against the target protein is required to bring this polar group to the attention of the design study.

## 7.5.2 Protein Modelling

When the three dimensional structure of proteins became available through x-ray crystallography it opened up a whole new field in rational drug design. It was possible to manipulate the view of a protein on the computer screen and visually examine the site of interest. This site of interest could be the catalytic centre of an enzyme, the site of protein - protein interaction, or the docking site of a small organic molecule. This ability to visualise the three dimensional structure of the area of interest on the target protein allows for greater understanding in how to design a candidate drug compound with better binding affinities.

Protein modelling techniques can be used in rational drug design when two key points are satisfied. Firstly the underlying molecular biology of the disease must be reasonably well understood so that a single key protein can be targeted. This key protein should be either central to developing the disease, or be the cause of the symptoms. For example, the reverse transcriptase of the AIDS virus carries out the reverse transcription of the virus's RNA genetic information into DNA. It is the DNA that is then inserted into the cell's genome and is replicated by the host's DNA transcription machinery. The initial reverse transcriptase is part of the protein coat of the AIDS virus. Inhibiting the reverse transcriptase prevents it from transcribing the viral RNA to DNA, an essential step as the cell itself has no mechanism to do this. The viral RNA can not be used by the cell to produce the proteins coded by it hence the AIDS virus is not replicated if no functional reverse transcriptase is present. The reverse transcriptase is an ideal target as it is essential for the AIDS disease to develop. It plays an essential role in the replication of the virus and has no counterpart in the mammalian cell. The other approach, of targeting a protein that is the key cause of the symptoms is something that the pharmaceutical companies also adopt. Here inhibiting the key protein will not cure the disease but only alleviate the symptoms. This then means the inhibitory compound must continually be taken or else the disease will flare up again. An example of this approach is in the development of treatments for rheumatoid arthritis. Here the target protein is usually some key protein in the inflammatory cascade. When this protein is inhibited the painful inflammation around the arthritic joints is reduced. This treatment does not tackle the underlying cause of the



rheumatoid arthritis which is an auto immune response to proteins found on the membrane surface of the joints.

The second key point is that the three dimensional structure of the target protein must be known. The best, most reliable three dimensional structures are determined using x-ray crystallography. The problem with x-ray crystallography is in crystallising the protein. The first problem to overcome is often that the protein you want to crystallise is produced in only minute quantities. It is therefore necessary to use molecular biology techniques to transfer the gene that codes for the protein to a biological system where it is expressed more strongly. This will give you quantities of the protein to attempt to crystallise it. This is the second problem as most proteins are very difficult to crystallise and it can take a lot of work and skill to get the conditions right for crystallisation to occur and crystals of the required size to be formed.

An alternative approach is to find a protein which already has its three dimensional structure solved and is homologous to the target protein you want to study. After building the three dimensional model of the target protein the quality of this model, which is then used in the protein - drug binding studies, directly affects the success of the drug design process. The greater the homology between the target and starting protein primary sequences' the closer the conformation of the protein model is to the biological protein and the more likely it is to design a drug that will bind strongly and selectively to the protein in vivo.

The biggest problem faced by protein modellers is that compared to the number of protein sequences known relatively few three dimensional structures of proteins have been solved. This is a direct result of the molecular biology revolution. An excess of 50 times the number of protein sequences as three dimensional protein structures is known. It is then highly likely that the primary sequence of a target protein is known but its three dimensional structure unknown. This is when protein modelling can be used to build a model of the target protein from the structure of an homologous protein with known structure. Starting from the known structure the model can be altered in steps as described earlier to produce an accurate a model as possible of the required target protein.

### 7.5.2.1 Confidence In The Protein Model

Different parts of the homologous protein model will have different confidence factors about how well the conformation of the model is likely to fit the protein *in vivo*. Areas of secondary structure and high homology have a higher confidence level than the modelled loop or surface regions. This can affect how good the model is at predicting which possible drug candidate will bind to the protein the most effectively. This can be seen as both good and bad news for drug design.

Where a drug is designed to block the catalytic function of an enzyme then the computer model should have a high confidence factor around this region as the catalytic site is highly conserved through a family of proteins. This gives the modeller confidence in knowing that the conformation and structure of the catalytic site is well known and a high probability of being correct in the target protein. The down side of this is that enzymes are normally highly specific. They behave in a manner like a lock and key with their specific substrate. The shape of the substrate is like a key fitting into the lock of the enzyme's catalytic site. Therefore certain regions of the substrate binding site around the catalytic site are highly diverse within a family of proteins. It is more usual for a drug design to take some of the diverse regions into account so that the drug candidate is selective to only one member of a protein family. The confidence factor of these non-homologous regions is much lower as the structure and exact conformation of this region of the model is specific to the protein and likely to be quite different from the starting protein model. This has the effect of reducing the confidence in any compounds designed around this region of the model to bind as well as or better than another compound.

### 7.5.3 Building the Compound Around The Protein Model

The newer approach to rational drug design, that of protein modelling, gives a direct view of the structural conformation and chemical environment of the drug binding site. This allows compounds to be designed so that they can 'fit' into the binding site with no steric clashes between the designed compound and the protein atoms. The chemical groups on the designed compound can be chosen to complement the chemical environment of the binding

site on the protein model. Both of these will enable the designed compound to bind more strongly to the target protein.

Building the compound to fit into the binding site of the target protein model starts by examining the surface of the binding site and building a three dimensional map of its features. The site is split into hydrogen acceptors, hydrogen donors, general hydrophobic regions.

The hydrogen acceptors found on the protein are the carboxyl groups of the polypeptide backbone, those found on the acidic residues (Asp and Glu) and the residues with amide groups on their sidechain (Asn and Gln). These carboxyl groups can act as acceptors in a hydrogen bond formed between the target protein and the bound compound. The hydrogen donors found on the protein surface are the  $N_{\text{amide}}\text{—H}$  group of the peptide bond, the  $N_{\text{amide}}\text{—H}$  group on the basic residues (Lys, Arg and His), and the OH groups on the polar residues (Ser, Thr and Tyr). These chemical groups can act as hydrogen donors in a hydrogen bond formed between the target protein and the bound compound. These acceptor and donor sites have directional and positional orientation. The best hydrogen bonds are formed when the complementary polar group is a given distance and direction from the site on the protein surface. The most favourable hydrogen bonds are formed when the donor—H—acceptor valence angle is  $180^\circ$ . This gives a cap of sites around each hydrogen bond acceptor and donor site.

The hydrophobic sites are general areas on the map. The hydrophobic regions of the binding site will attract the hydrophobic regions in the compound. These areas follow the contours of the surface of the binding site but are not directional. The driving force in the hydrophobic interactions is the loss of entropy by the water molecules of the solvent in contact with hydrophobic surfaces. The hydrogen bond network of the solvent water molecules is disrupted around the hydrophobic region. So it is the force of the water trying to minimise this disruptive effect and driving the hydrophobic regions together that causes the two hydrophobic surfaces into close contact and to remain in close contact.

Once a complete map of the binding site of the target protein has been built potential drug candidates can be designed. Fast but simple scoring functions such as number of steric clashes and number of favourable interactions are used as the point is to build as many compounds as possible. The compounds are made by starting with a seed structure and building from that. The seed structure can be a single functional group that the drug must contain and the rest of the compound is grown from this starting position. Alternatively a few functional groups are positioned into the binding site at strategic positions to have specific favourable interactions with the target protein. The algorithm then joins up these fragments with bridging elements, adding other functional groups where possible. To speed up the building process libraries of pre-built molecule fragments are used which can be scanned quickly. This allows functional groups and bridging groups to be added to the growing molecule rapidly rather than trying to build the compound atom by atom. The building functions know how to add these fragments together keeping the geometry of the growing compound correct. The functional groups are the heteroatom containing sections such as alcohols, esters, amides and groups such as benzyl derivatives. These groups are used to bind to hydrogen bond acceptor and donors on the protein's surface, or in the case of the hydrophobic fragments to interact with large hydrophobic pockets or regions on the protein surface. The bridging groups are mostly rigid extended structures used to span the gap between the functional groups. Using mainly rigid structures with as few rotational bonds as possible is an attempt to reduce the complication of having too flexible a compound.

One of the first tests that the compounds produced by the molecule building routine is put through is an inspection to see how feasible it is to synthesise the molecule. Until recently there were no rules in the compound building algorithm on synthesis routes and it took an experienced organic chemist to tell whether a compound could be synthesised and how easy it could be done. Now knowledge based systems are becoming available. These systems have knowledge about chemical reactions built into the program. They are also linked to databases of known synthesis routes which are searched to give a possible route to synthesise the compound at the bench and can give a list of possible starting material. These possible routes give an indication of the possible difficulty or simplicity in

synthesising a particular molecule. The list of starting material is retrieved from databases of available off the shelf compounds that can be readily purchased. It is the job of the program to amalgamate all this information together into a synthesis route from easily available starting material to the required compound via known chemical reactions. This idea of only producing compounds that can be synthesised as possible drug candidates can be taken into the modelling stages where the compound is built. By using libraries of fragments with known synthesis routes and with rules built in to the algorithm about different synthesis routes, the design compounds now produced as possible drug candidates are much more likely to be able to be synthesised.

#### 7.5.4 Advantage Of Using Protein Model Over QSAR

Having the 3D structure of the entire binding site allows the design of the drug candidate to take into account any feature of the binding site. Compounds can be designed to fit exactly the actual contours of the binding site, making use of all the features present. As discussed above one of the drawbacks with the QSAR method is that it has an incomplete picture of the binding site. The QSAR method is very dependent on the database set of analogous compounds to pick out as many features of the binding site as possible. The advantage of having a three dimensional model of the target protein is that all the features of the binding site are already known. The specific features of the binding site can be used to the best advantage with the compound designed to interact with as many of these features as possible. Finding new leads of compounds that can dock into the binding site well is much simpler with the entire three dimensional structure of the binding site known. It is possible to experiment with different classes of compounds fitting them into the binding site of the model examining how well they potentially fit without having to carry out laboratory experiments. With the surface of the binding site well known it is possible to design compounds that follow the conformation of the surface making plenty of favourable interactions with the features found on the surface of the binding site.

The problem of the flexibility of the compounds designed and of the binding site can be more readily tackled with protein modelling. At worst the conformational flexibility can severely hinder the drug design process but an attempt can be made to solve the problem

with the use of molecular mechanical techniques. At the very best the problem of conformational flexibility in the model greatly increases the time spent on the drug design process. The main problem with a flexible compound is knowing the conformation it will take when bound to the target protein. A favourable interaction of one polar site in the binding site with part of the compound can compensate for an increase in potential energy of the compound if it is in a slightly strained conformation. The outcome of this means that a compound may not bind to the binding site in its lowest energy conformation. Carrying out molecular dynamic simulations of the compound protein model can indicate how stable the compound is within the binding site. If the drug interacts strongly and favourably with the target protein in the binding site then little movement of the compound within the binding site will be observed over the time scale the simulation is allowed to run. The ideal situation is if the compound remains where it was designed to bind on the protein and the compound and target protein keep the same interactions over the time scale of the simulation. If the compound binds poorly to the binding site then during the molecular mechanics simulation the compound can be observed to 'wander' around the surface of the target protein or move off the surface completely. During the time scale of the simulation the compound will form arbitrary interactions with different regions of the protein surface as it moves around.

The molecular dynamic simulations can be very time consuming. The length of time for a calculation is dependent on the number of atoms in the model that is allowed to move and the number of time steps in the simulation. A reasonable time scale to take for each iteration in the calculation is 0.001 nanoseconds. Therefore for even 1 nanosecond of time in the simulation involves 1,000 iterations of the calculation. The number of atoms allowed to move in a calculation affects the time taken to carry out each iteration. The time taken for each iteration grows factorially as the number of atoms allowed to move in the simulation. This fast growth rate in time can be reduced by having a cut off distance, that is if two atoms are further than a given distance apart then the force between them is said to be negligible and so can be ignored. This greatly speeds up the time taken for each iteration. Due to the constraint in the size of each time step the calculation can take it still requires 10,000's of iterations for any significant time elapse in the simulation. Over short

time periods only valence bond vibrations will be seen in the simulation, any conformational changes occur in the nanosecond time scale.

### 7.5.5 Scoring Functions

Some measurement of how well a compound binds to the target protein model is required so that the best compounds can be selected for synthesis and screening with the protein in an assay. The QSAR method has a scoring function built into the developed model since the base of the model are physiochemical or other properties which can be related to some pharmacological property  $Q$ , in this case the binding affinity of a compound to a target protein. Therefore in the QSAR model that is developed, an increase in the value  $Q$  by a compound means that it is predicted to bind better to the target protein than a compound which has a lower  $Q$  value. This makes the decision of choosing which compounds are to be synthesised and screened against the protein easier. There is still the problem with QSAR in that it is not good at predicting new lead compounds to be screened so some care must be taken in screening a reasonably varied array of compounds in the chance one might bind well with the protein and give a new lead structure. The QSAR model can be refined with each iteration of drug design, compound synthesis and screening. The results of how well the compounds fared in the binding assay can be used to improve and refine the QSAR model. As more information is added to the model the confidence of its prediction abilities will increase.

Protein modelling on the other hand has no such easily defined scoring function. The scoring methods that have been used in rational drug design studies using protein modelling techniques are much more arbitrarily defined. The scoring functions used in these studies are more subjective than the scoring functions developed in the QSAR models.

One popular method is to use molecular mechanic calculation to carry out energy minimisations on the compound bound to the binding site of the target protein model. The best compounds are deemed to be those that have a high favourable interaction energy but where the compound is in a conformation close in energy to its global minimum energy. The high interaction energies come from hydrogen bonds forming between two polar

groups, one on a residue of the protein model and another on the compound model. These hydrogen bonds have the same energies and geometries as the intramolecular hydrogen bonds found in proteins. Another important type of interaction that gives rise to a favourable intermolecular energy is when two hydrophobic surfaces are in contact. Although the energies involved can be much weaker than the electrostatic energies involved in the hydrogen bonding it still plays a significant role in protein - drug interactions. As mentioned the energy of the conformation the compound adopts on binding to the target protein should ideally be close to the compound's global minimum. Finding the global minimum of small organic molecules up to 10's of atoms can be carried out on the desktop workstation using semi - empirical techniques. An accurate minimum energy conformation can be calculated using these methods. Comparing this global minimum energy to the calculated intramolecular energy of the compound after the model has been minimised gives an indication of the strain in the conformation that the compound must contend with while bound to the protein. If the strain in the conformation is too great the compound will not bind to the target protein.

Carrying out an energy minimisation calculation for an entire protein and compound complex is very computationally expensive. Other strategies must be used to decrease the time taken to analyse the interactions of one compound in the binding site of the target protein model. To speed up each individual minimisation the number of atoms that are free to move during the calculation is drastically cut. In the most severe cases only the compound is free to move during the energy minimisation. Before the start of the rational drug design study the target protein model will have been through an energy minimisation calculation and be at or very near its local energy minimum. This forces the compound to adopt the complementary shape of the binding site and is very unrealistic to what actually occurs.

For most proteins what actually happens when a compound binds to a protein is that residues in the binding site move and will move to produce more favourable binding with the docked compound. The conformation change of the protein to a more strained conformation is offset with the more favourable interactions formed with the compound in



this slightly strained conformation. Only those residues of the protein model neighbouring the docked compound change their conformation on binding to the compound. The change in the protein conformation on a compound docking to the protein model is localised to the area around the docked compound. The other residues of the protein model do not change conformation from their initial position. This allows the entire protein to be put through the energy minimiser once to give the local minimum energy conformation and when compounds are docked to the protein model only those residues in the binding site of the target protein need be allowed to move during the energy calculation. This approach greatly reduces the number of atoms that are free to move during the energy minimisation calculations when binding a compound to the target protein making the calculation significantly faster. This approach gives a more realistic model of what happens when the compound is bound to the target protein without increasing the length of the calculation too significantly.

Molecular dynamics can also be used to give an indication of how well a compound binds to the target protein model. Carrying out the molecular dynamics simulation at body temperature 310K gives a better indication of how the compound will bind in the screening process. If the compound binds tightly to the target protein model then little movement of the compound will be seen during the simulation. Ideally the interactions the compound was designed to make with the protein model will remain, specifically the same interatomic hydrogen bonds between compound and protein should stay or continually appear during the simulation time frame. If on the other hand the compound binds poorly to the protein model it will be seen drifting away from the starting binding site. Arbitrary hydrogen bonds will be formed during each time frame of the simulation as the compound moves across the surface. If the original binding of the compound to the protein model was particularly strained then the compound may leave the protein model's surface completely. Molecular dynamic simulations are also good for looking at flexible compound models as previously discussed. Again if the entire protein model is free to move during the molecular dynamics simulation this makes the calculation too slow to be feasible. The simulation calculation can be made to run quicker if fewer atoms in the model are allowed to move during the calculation. The residues in the target protein model that are free to move in the simulation

are those that point into the binding site. During a simulation at body temperature a compound that binds strongly to the target protein will remain within the binding site. This binding site is normally a reasonably well defined feature on the surface of the target protein like a cleft so there is a clearly defined subset of atoms in the protein model that are allowed to move during the molecular dynamics simulation.

Both molecular mechanic methods described above are time consuming. They do not allow for the quick analysis of many hundreds or thousands of compound designs. The molecular mechanic methods described above are used for more in depth studies of a few carefully selected compounds. For screening large numbers of compounds bound to the protein model a much quicker but more subjective method is required in selecting the compounds that best fit the binding site. These methods revolve around a simple scoring function that gives some indication of how good the compound is bound to the target protein model. They take into consideration only intermolecular effects and assume that the protein and compound are already in a minimum energy or low energy conformation. The methods add up all the favourable and unfavourable interactions between the compound and protein and the total gives some indication on how well the compound binds to the protein model. Each of the interactions taken into account by the function has a different weighting. These functions might not be as robust as the molecular mechanics methods but they allow for a faster throughput of possible drug candidates. This is important as the bottleneck in rational drug design is in the analysis of the potential drug candidate.

## **7.6 Conclusions**

The model of Bb as it stands is acceptable as a homology model but this does not mean that it is not impossible to improve the quality of the model. The analysis carried out on the structure clearly shows that there is plenty of scope for improvement. Many of the alterations that can be done are relatively quick and easy to carry out. A prime example of this is making sure all the residues in the model are of the L isomer. Any alterations carried out on the side chain atoms can be done relatively easily. After these fairly minor alterations are carried out the model can continue to be run through the energy minimiser. This will hopefully help improve the problems with the planarity of the peptide bonds in

the structure. Protein crystallographers have had to carry out such analysis checks on the crystal structures of proteins that they produce, therefore it is only reasonable to expect protein modellers to carry out the same checks on their structures.

Comparative / Homology modelling is still the most effective and reliable method of producing a three dimensional model of a protein sequence which has yet to have its structure solved by x-ray crystallography. This does not seem likely to change in the near future. The depressing point is the lack of progress in recent years of improving the quality of the model produced by comparative modelling. The main stumbling block of the comparative modelling technique is calculating the conformation of large insertions. The algorithms which generate all possible conformations an insertion can adopt are severely limited by the number of residues that they can handle. They also appear to have problems in deciding which conformation an insertion will adopt. Until these problems are solved the use of protein modelling will be restricted by the range of structures comparative / homology modelling can cope with.

## **8 REPORT OF PROTEIN ANALYSIS by the WHAT IF program. Date : 1997-22-12**

### **8.1 INTRODUCTION**

This document contains a report of findings by the WHAT IF program during the analysis of one or more proteins. It contains a separate section for each of the proteins that have been analysed. Each reported fact has an assigned severity, one of:

**error:** severe errors encountered during the analyses. Items marked as errors are considered severe problems requiring immediate attention.

**warning:** Either less severe problems or uncommon structural features. These still need special attention.

**note:** Statistical values, plots, or other verbose results of tests and analyses that have been performed.

If alternate conformations are present, only the first is evaluated.

hydrogen atoms are ignored.

### **8.2 Legend**

Some notations need a little explanation:

**RESIDUE:** Residues in tables are normally given in 4 parts:

1. - A number. This is the internal sequence number of the residue used by WHAT IF.
2. - The residue name. Normally this is a three letter amino acid name.
3. - The sequence number, between brackets. This is the residue number as it was given in the input file.
4. - The chain identifier. A single character. If no chain identifier was given in the input file, a '-' is given.

**Z-VALUE:** To indicate the normality of a score, the score may be expressed as a Z-value or Z-score. This is just the number of standard deviations that the score deviates from the expected value. A property of Z-values is that the root-mean-square of a group of Z-values is expected to be 1.0. Z-values above 4.0 and below -4.0 are very uncommon. If a Z-score is used in WHAT IF, the accompanying text will explain how the expected value and standard deviation

=====  
Compound code fb\_44a.pdb  
=====

# 1 # Error: Missing unit cell information

No SCALE matrix is given in the PDB file.

# 2 # Error: Missing symmetry information

Problem: No CRYST1 card is given in the PDB file.

# 3 # Error: Threonine nomenclature problem

The threonine residues listed in the table below have their

O-gamma-1 and C-gamma-2 swapped.

7 THR ( 22 ) A

123 THR ( 4 ) A

160 THR ( 2 ) A

196 THR ( 177 ) A

=====  
Compound code fb\_44a\_noH.pdb  
=====

# 1 # Warning: Class of space group could be incorrect

The space group symbol indicates a different class than the unit cell given on the CRYST1 card of the PDB file.

Possible cause: The unit cell may have pseudo-symmetry, or one of the cell dimensions or the space group might be given incorrectly.

Crystal class of the cell: CUBIC

Crystal class of the space group: TRICLINIC

Space group name: P 1

# 2 # Note: No rounded coordinates detected

No significant rounding of atom coordinates has been detected.

# 3 # Error: Matthews Coefficient ( $V_m$ ) very high

The Matthews coefficient [REF] is defined as the density of the protein structure in cubic Angstroms per Dalton. Normal values are

between 1.5 (tightly packed, little room for solvent) and 4.0 (loosely packed, much space for solvent). Some very loosely packed structures can get values a bit higher than that.

Numbers this high are almost always caused by giving the wrong value for Z on the CRYST1 card.

Molecular weight of all polymer chains: 31572.467

Volume of the Unit Cell  $V = 0.1000E+10$

Cell multiplicity: 0

Matthews coefficient for observed atoms  $V_m = 31673.160$

# 4 # Note: Valine nomenclature OK

No errors were detected in valine nomenclature.

# 5 # Note: Threonine nomenclature OK

No errors were detected in threonine nomenclature.

# 6 # Note: Isoleucine nomenclature OK

No errors were detected in isoleucine nomenclature.

# 7 # Note: Leucine nomenclature OK

No errors were detected in leucine nomenclature.

# 8 # Note: Arginine nomenclature OK

No errors were detected in arginine nomenclature.

# 9 # Note: Tyrosine torsion conventions OK

No errors were detected in tyrosine torsion angle conventions.

# 10 # Note: Phenylalanine torsion conventions OK

No errors were detected in phenylalanine torsion angle conventions.

# 11 # Note: Aspartic acid torsion conventions OK

No errors were detected in aspartic acid torsion angle conventions.

# 12 # Note: Glutamic acid torsion conventions OK

No errors were detected in glutamic acid torsion angle conventions.

# 13 # Note: Heavy atom naming OK

No errors were detected in the atom names for non-hydrogen atoms.

# 14 # Warning: Chirality deviations detected

The atoms listed in the table below have an improper dihedral value that is deviating from expected values.

Improper dihedrals are a measure of the chirality/planarity of the structure at a specific atom. Values around -35 or +35 are expected



for chiral atoms, and values around 0 for planar atoms. Planar side chains are left out of the calculations, these are better handled by the planarity checks.

Three numbers are given for each atom in the table. The first is the Z-score for the improper dihedral. The second number is the measured improper dihedral. The third number is the expected value for this atom type. A final column contains an extra warning if the chirality for an atom is opposite to the expected value.

1 TRP ( 16 ) A C	-9.0	-15.7	0.1
3 HIS ( 18 ) A CA	-6.2	20.0	34.4
3 HIS ( 18 ) A C	-8.2	-15.9	-0.1
7 THR ( 22 ) A CA	4.5	43.9	34.1
7 THR ( 22 ) A C	-6.9	-13.1	-0.1
8 ASP ( 23 ) A CA	-6.3	19.5	34.0
12 GLN ( 27 ) A CA	-6.5	19.6	34.2
12 GLN ( 27 ) A C	-4.1	-7.4	-0.1
18 ILE ( 5 ) A CA	-6.9	19.7	33.8
20 VAL ( 7 ) A CA	-6.7	19.3	33.7
21 ILE ( 33 ) A CB	-8.9	15.6	33.3
26 GLY ( 38 ) A C	11.8	19.2	0.1
27 HIS ( 39 ) A C	9.3	17.9	-0.1
28 GLU ( 40 ) A CA	10.8	57.2	34.4

28	GLU ( 40 )	A C	-10.2	-18.8	0.0
29	SER ( 41 )	A CA	-8.1	14.9	34.3
30	CYS ( 42 )	A C	-7.2	-13.0	-0.1
31	MET ( 43 )	A C	7.7	13.8	0.0
35	VAL ( 47 )	A C	4.6	8.6	-0.3
36	SER ( 48 )	A CA	-6.0	20.0	34.3
40	VAL ( 52 )	A C	6.4	12.1	-0.3
41	LEU ( 53 )	A C	5.2	9.5	-0.1
42	THR ( 54 )	A C	-6.5	-12.4	-0.1
46	CYS ( 58 )	A C	14.0	25.1	-0.1
47	PHE ( 59 )	A CA	6.6	49.6	34.2

And so on for a total of 147 lines

#### # 15 # Warning: High improper dihedral angle deviations

The RMS Z-score for the improper dihedrals in the structure is high.

For well refined structures this number is expected to be around 1.0.

The fact that it is higher than 1.5 in this structure could be an indication of overrefinement.

Improper dihedral RMS Z-score : 4.153

#### # 16 # Error: Decreasing residue numbers

At least one residue in each of the chains mentioned below has a residue number that is lower than the previous residue in that

chain ('-' represents a chain without chain identifier).

Chain identifier(s): A

# 17 # Note: Weights checked OK

All atomic occupancy factors ('weights') fall in the 0.0--1.0 range.

# 18 # Note: No missing atoms detected

All expected atoms are present.

# 19 # Warning: C-terminal oxygen atoms missing

The C-atoms listed in the table below belong to a C-terminal residue in a protein chain, but the C-terminal oxygen ("O2" or "OXT") that it should be bound to was not found.

278 GLU ( 245 ) A C

# 20 # Note: No extra C-terminal groups found

No C-terminal groups are present for non C-terminal residues

# 21 # Warning: Unusual bond lengths

The bond lengths listed in the table below were found to deviate more than 4 sigma from standard bond lengths (both standard values and sigma for amino acid residues have been taken from Engh and

Huber [REF], for DNA they were taken from Parkinson et al [REF]). In the table below for each unusual bond the bond length and the number of standard deviations it differs from the normal value is given.

Atom names starting with "<" belong to the previous residue in the chain. If the second atom name is "--SS", the disulphide bridge has a deviating length.

4 ARG ( 19 ) A	CZ NH1	1.244	-4.5
7 THR ( 22 ) A	CA CB	1.638	5.4
7 THR ( 22 ) A	CB OG1	1.550	7.3
18 ILE ( 5 ) A	CA CB	1.619	4.5
19 SER ( 6 ) A	CA CB	1.623	4.7
21 ILE ( 33 ) A	CA CB	1.621	4.5
22 ARG ( 34 ) A	N CA	1.562	5.5
22 ARG ( 34 ) A	CD NE	1.545	4.7
27 HIS ( 39 ) A	CG CD2	1.410	4.9
28 GLU ( 40 ) A	CA CB	1.634	5.2
29 SER ( 41 ) A	N CA	1.537	4.2
45 HIS ( 57 ) A	CG CD2	1.403	4.3
45 HIS ( 57 ) A	ND1 CE1	1.379	4.6
47 PHE ( 59 ) A	N CA	1.550	4.9
48 THR ( 60 ) A	N CA	1.534	4.0

49 VAL ( 61 ) A N CA 1.555 5.1  
49 VAL ( 61 ) A CA CB 1.654 6.2  
50 ASP ( 62 ) A CG OD1 1.360 5.8  
51 ASP ( 62A ) A N <C 1.191 -6.9  
52 LYS ( 69 ) A N CA 1.545 4.6  
52 LYS ( 69 ) A CA CB 1.628 4.9  
53 GLU ( 70 ) A CD OE2 1.359 5.8  
54 HIS ( 71 ) A CG CD2 1.408 4.8  
54 HIS ( 71 ) A ND1 CE1 1.375 4.3  
57 LYS ( 73B ) A N CA 1.552 5.0

And so on for a total of 97 lines

**# 22 # Warning: High bond length deviations**

Bond lengths were found to deviate more than normal from the mean standard bond lengths (standard values for protein residues were taken from Engh and Huber [REF], for DNA/RNA these values were taken from Parkinson et al [REF]). The RMS Z-score given below is expected to be around 1.0 for a normally restrained data set. The fact that it is higher than 1.5 in this structure might indicate that the constraints used in the refinement were not strong enough. This will also occur if a different bond length dictionary is used.

RMS Z-score for bond lengths: 1.806

RMS-deviation in bond distances: 0.037

# 23 # Warning: Possible cell scaling problem

Comparison of bond distances with Engh and Huber [REF] standard values for protein residues and Parkinson et al [REF] values for DNA/RNA shows a significant systematic deviation. It could be that the unit cell used in refinement was not accurate enough. The deformation matrix given below gives the deviations found: the three numbers on the diagonal represent the relative corrections needed along the A, B and C cell axis. These values are 1.000 in a normal case, but have significant deviations here (significant at the 99.99% confidence level)

There are a number of different possible causes for the discrepancy. First the cell used in refinement can be different from the best cell calculated. Second, the value of lambda used for a synchrotron data set can be miscalibrated. Finally, the discrepancy can be caused by a dataset that has not been corrected for significant anisotropic thermal motion.

Please note that the proposed scale matrix has NOT been constrained to obey the space group symmetry. This is done on purpose. The distortions can give you an indication of the accuracy of the determination.

Unit Cell deformation matrix

0.986365 -0.000932 0.000135  
-0.000932 0.987224 -0.000052  
0.000135 -0.000052 0.986245

Proposed new scale matrix

0.001014 0.000001 0.000000  
0.000001 0.001013 0.000000  
0.000000 0.000000 0.001014

With corresponding cell

A = 986.365 B = 987.224 C = 986.245  
Alpha= 90.006 Beta= 89.984 Gamma= 90.108

The CRYST1 cell dimensions

A =1000.000 B =1000.000 C =1000.000  
Alpha= 90.000 Beta= 90.000 Gamma= 90.000

Variance: 1570.906

(Under-)estimated Z-score: 29.211

# 24 # Warning: Unusual bond angles

The bond angles listed in the table below were found to deviate more than 4 sigma from standard bond angles (both standard values and sigma for protein residues have been taken from Engh and Huber [REF], for DNA/RNA from Parkinson et al [REF]). In the table below

for each strange angle the bond angle and the number of standard deviations it differs from the standard values is given. Please note that disulphide bridges are neglected. Atoms starting with "<" belong to the previous residue in the sequence.

1 TRP ( 16) A CB CG CD1 114.044 -8.6  
1 TRP ( 16) A CB CG CD2 136.320 6.8  
1 TRP ( 16) A NE1 CE2 CZ2 120.701 -6.3  
1 TRP ( 16) A CE3 CD2 CG 127.036 -6.9  
2 GLU ( 17) A <O <C N 116.017 -4.4  
2 GLU ( 17) A N CA CB 92.401 -10.6  
3 HIS ( 18) A N CA C 141.822 10.9  
3 HIS ( 18) A N CA CB 118.044 4.4  
3 HIS ( 18) A C CA CB 96.001 -7.4  
3 HIS ( 18) A CA CB CG 127.488 13.7  
3 HIS ( 18) A CG ND1 CE1 98.063 -7.5  
3 HIS ( 18) A ND1 CE1 NE2 120.816 7.0  
3 HIS ( 18) A CE1 NE2 CD2 97.813 -7.0  
3 HIS ( 18) A CD2 CG ND1 111.148 5.0  
3 HIS ( 18) A CB CG CD2 122.381 -5.2  
4 ARG ( 19) A <O <C N 113.671 -5.8  
4 ARG ( 19) A N CA CB 95.233 -9.0  
4 ARG ( 19) A C CA CB 126.058 8.4  
5 LYS ( 20) A C CA CB 122.541 6.5



7 THR ( 22 ) A <C N CA 129.258 4.2  
7 THR ( 22 ) A CA C O 109.062 -6.9  
7 THR ( 22 ) A C CA CB 102.015 -4.3  
7 THR ( 22 ) A CA CB CG2 118.899 4.9  
7 THR ( 22 ) A CA CB OG1 119.684 6.7  
7 THR ( 22 ) A CG2 CB OG1 97.976 -5.7

And so on for a total of 571 lines

**# 25 # Warning: High bond angle deviations**

Bond angles were found to deviate more than normal from the mean standard bond angles (normal values for protein residues were taken from Engh and Huber [REF], for DNA/RNA from Parkinson et al [REF]). The RMS Z-score given below is expected to be around 1.0 for a normally restrained data set, and this is indeed observed for very high resolution X-ray structures. More common values are around 1.55. The fact that it is higher than 2.0 in this structure might indicate that the constraints used in the refinement were not strong enough. This will also occur if a different bond angle dictionary is used.

RMS Z-score for bond angles: 2.987

RMS-deviation in bond angles: 5.608

**# 26 # Error: Side chain planarity problems**

The side chains of the residues listed in the table below contain a planar group that was found to deviate from planarity by more than 4.0 times the expected value. For an amino acid residue that has a side chain with a planar group, the RMS deviation of the atoms to a least squares plane was determined. The number in the table is the number of standard deviations this RMS value deviates from the expected value (0.0).

1 TRP ( 16 ) A	19.406
3 HIS ( 18 ) A	11.659
28 GLU ( 40 ) A	8.355
93 ASP ( 100 ) A	7.383
95 ASP ( 102 ) A	6.808
15 GLN ( 2 ) A	5.710
238 TRP ( 215 ) A	5.456
9 TYR ( 24 ) A	5.263
258 ARG ( 3 ) A	4.942
120 GLU ( 1 ) A	4.852
230 PHE ( 4 ) A	4.845
22 ARG ( 34 ) A	4.741
210 ASP ( 7 ) A	4.360
63 GLU ( 78 ) A	4.115
156 GLU ( 149 ) A	4.022

# 27 # Error: Connections to aromatic rings out of plane

The atoms listed in the table below are connected to a planar aromatic group in the sidechain of a protein residue but were found to deviate from the least squares plane.

For all atoms that are connected to an aromatic side chain in a protein residue the distance of the atom to the least squares plane through the aromatic system was determined. This value was divided by the standard deviation from a distribution of similar values from a database of small molecule structures.

75	HIS	( 91 )	A	CB	12.666
92	TYR	( 99 )	A	OH	10.691
256	HIS	( 1 )	A	CB	10.113
54	HIS	( 71 )	A	CB	9.278
165	TYR	( 7 )	A	OH	8.272
181	TYR	( 4 )	A	OH	7.571
185	TYR	( 8 )	A	CB	6.946
208	TYR	( 5 )	A	OH	5.991
185	TYR	( 8 )	A	OH	5.884
151	PHE	( 144 )	A	CB	5.408
94	TYR	( 101 )	A	OH	5.329
208	TYR	( 5 )	A	CB	5.254
107	TYR	( 114 )	A	CB	4.805

181 TYR ( 4 ) A CB 4.686  
38 TYR ( 50 ) A CB 4.500  
199 PHE ( 180 ) A CB 4.398  
78 TYR ( 94 ) A OH 4.112

**# 28 # Warning: Unusual PRO puckering amplitudes**

The proline residues listed in the table below have a puckering amplitude that is outside of normal ranges. Puckering parameters were calculated by the method of Cremer and Pople [REF]. Normal PRO rings have a puckering amplitude Q between 0.20 and 0.45 Angstrom. If Q is lower than 0.20 Angstrom for a PRO residue, this could indicate disorder between the two different normal ring forms (with C-gamma below and above the ring, respectively). If Q is higher than 0.45 Angstrom something could have gone wrong during the refinement.

89 PRO ( 98B ) A 0.72 HIGH  
143 PRO ( 2 ) A 0.73 HIGH  
211 PRO ( 8 ) A 0.55 HIGH  
221 PRO ( 198 ) A 0.45 HIGH  
254 PRO ( 229 ) A 0.55 HIGH

**# 29 # Warning: Unusual PRO puckering phases**

The proline residues listed in the table below have a puckering phase

that is not expected to occur in protein structures. Puckering parameters were calculated by the method of Cremer and Pople [REF]. Normal PRO rings approximately show a so-called envelope conformation with the C-gamma atom above the plane of the ring ( $\phi=+72$  degrees), or a half-chair conformation with C-gamma below and C-beta above the plane of the ring ( $\phi=-90$  degrees). If  $\phi$  deviates strongly from these values, this is indicative of a very strange conformation for a PRO residue, and definitely requires a manual check of the data.

76 PRO ( 92 ) A 100.4 envelop C-beta (108 degrees)  
113 PRO ( 120 ) A 48.9 half-chair C-delta/C-gamma (54 degrees)  
117 PRO ( 124 ) A -118.5 half-chair C-delta/C-gamma (-126 degrees)  
207 PRO ( 4 ) A 51.1 half-chair C-delta/C-gamma (54 degrees)  
211 PRO ( 8 ) A -53.4 half-chair C-beta/C-alpha (-54 degrees)  
254 PRO ( 229 ) A 138.7 envelop C-alpha (144 degrees)

# 30 # Warning: Torsion angle evaluation shows unusual residues  
The residues listed in the table below contain bad or abnormal torsion angles.

These scores give an impression of how "normal" the torsion angles in protein residues are. All torsion angles except omega are

used for calculating a 'normality' score. Average values and standard deviations were obtained from the residues in the WHAT IF database. These are used to calculate Z-scores. A residue with a Z-score of below -2.0 is poor, and a score of less than -3.0 is worrying. For such residues more than one torsion angle is in a highly unlikely position.

42 THR ( 54 ) A -3.1854

143 PRO ( 2 ) A -3.1510

89 PRO ( 98B ) A -3.1193

211 PRO ( 8 ) A -3.0911

185 TYR ( 8 ) A -2.9986

260 PHE ( 5 ) A -2.9651

230 PHE ( 4 ) A -2.9440

147 ILE ( 6 ) A -2.8843

261 HIS ( 6 ) A -2.8521

265 PHE ( 10 ) A -2.7788

91 PHE ( 98D ) A -2.7692

164 VAL ( 6 ) A -2.7508

208 TYR ( 5 ) A -2.7354

22 ARG ( 34 ) A -2.7111

244 CYS ( 5 ) A -2.7062

20 VAL ( 7 ) A -2.6770

150 LEU ( 9 ) A -2.6748

264 LEU ( 9 ) A -2.5946

54 HIS ( 71 ) A -2.5269

92 TYR ( 99 ) A -2.5235

240 VAL ( 1 ) A -2.5050

259 ASP ( 4 ) A -2.4937

21 ILE ( 33 ) A -2.4517

222 LEU ( 199 ) A -2.3591

10 HIS ( 25 ) A -2.3432

And so on for a total of 61 lines

# 31 # Warning: Backbone torsion angle evaluation shows unusual conformations

The residues listed in the table below have abnormal backbone torsion angles.

Residues with "forbidden" phi-psi combinations are listed, as well as residues with unusual omega angles (deviating by more than 3 sigma from the normal value). Please note that it is normal if about 5 percent of the residues is listed here as having unusual phi-psi combinations.

2 GLU ( 17 ) A omega poor

3 HIS ( 18 ) A Poor phi/psi, omega poor

4 ARG ( 19 ) A Poor phi/psi

7 THR ( 22 ) A omega poor

8 ASP ( 23 ) A Poor phi/psi  
9 TYR ( 24 ) A omega poor  
10 HIS ( 25 ) A Poor phi/psi  
11 LYS ( 26 ) A omega poor  
12 GLN ( 27 ) A PRO omega poor  
15 GLN ( 2 ) A Poor phi/psi  
16 ALA ( 3 ) A omega poor  
17 LYS ( 4 ) A Poor phi/psi  
18 ILE ( 5 ) A omega poor  
19 SER ( 6 ) A Poor phi/psi  
20 VAL ( 7 ) A Poor phi/psi, omega poor  
21 ILE ( 33 ) A Poor phi/psi, omega poor  
22 ARG ( 34 ) A Poor phi/psi, PRO omega poor  
23 PRO ( 34A ) A Poor PRO-phi  
24 SER ( 34B ) A Poor phi/psi  
25 LYS ( 37 ) A Poor phi/psi  
26 GLY ( 38 ) A Poor phi/psi, omega poor  
27 HIS ( 39 ) A omega poor  
28 GLU ( 40 ) A omega poor  
34 VAL ( 46 ) A omega poor  
36 SER ( 48 ) A omega poor

And so on for a total of 151 lines

# 32 # Error: Ramachandran Z-score very low



The score expressing how well the backbone conformations of all residues are corresponding to the known allowed areas in the Ramachandran plot is very low.

Ramachandran Z-score : -5.194

# 33 # Warning: Omega angle restraints not strong enough

The omega angles for trans-peptide bonds in a structure is expected to give a gaussian distribution with the average around +178 degrees, and a standard deviation around 5.5. In the current structure the standard deviation of this distribution is above 7.0, which indicates that the omega values have been under-constrained.

Standard deviation of omega values : 18.623

# 34 # Error: chi-1/chi-2 angle correlation Z-score very low

The score expressing how well the chi-1/chi-2 angles of all residues are corresponding to the populated areas in the database is very low.

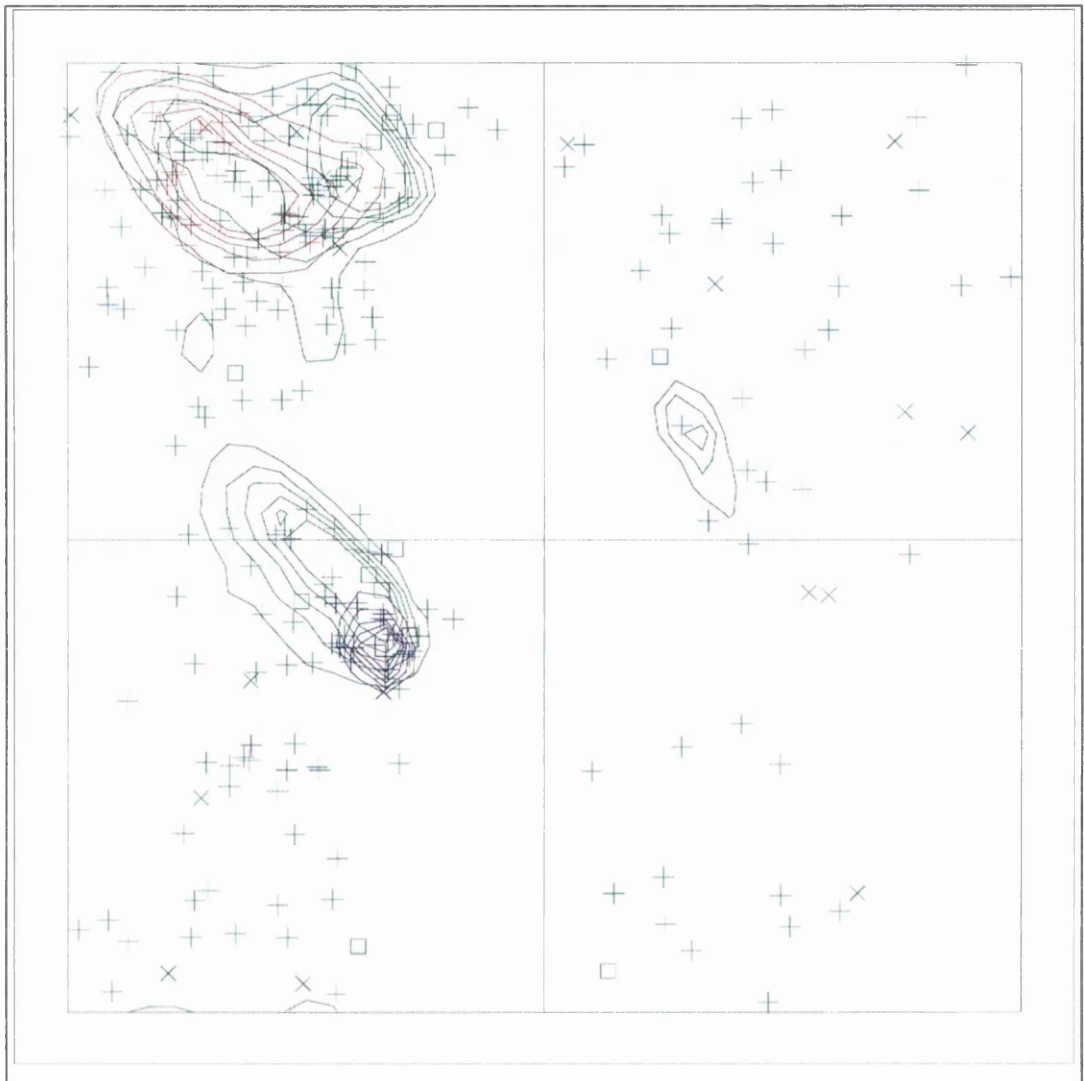
chi-1/chi-2 correlation Z-score : -4.444

# 35 # Note: Ramachandran plot

In this Ramachandran plot X-signs represent glycines, squares represent

prolines and small plus-signs represent the other residues. If too many plus-signs fall outside the contoured areas then the molecule is poorly refined (or worse).

In a colour picture, the residues that are part of a helix are shown in blue, strand residues in red. "Allowed" regions for helical residues are drawn in blue, for strand residues in red, and for all other residues in green.



Chain identifier: A

# 36 # Warning: Inside/Outside residue distribution unusual

The distribution of residue types over the inside and the outside of the protein is unusual. Normal values for the RMS Z-score below are between 0.84 and 1.16. The fact that it is higher in this structure could be

caused by transmembrane helices, by the fact that it is part of a multimeric active unit, or by mistraced segments in the density.

inside/outside RMS Z-score : 1.243

# 37 # Note: Inside/Outside RMS Z-score plot

The Inside/Outside distribution normality RMS Z-score over a 15 residue window is plotted as function of the residue number. High areas in the plot (above 1.5) indicate unusual inside/outside patterns.

In the TeX file, a plot has been inserted here

Chain identifier: A

# 38 # Note: Secondary structure

This is the secondary structure according to DSSP. Only helix (H), strand (S), turn (T) and coil (blank) are shown. [REF]

Secondary structure assignment

The DSSP executable was not found

/p6/exe1/whatif/dssp/DSSP.EXE

WARNING. You don't have the DSSP program installed. Therefore the emulator will be used. This emulator gives rather poor results,

but it prevents WHAT IF from crashing. See the writeup about this.

60 10 20 30 40 50

| | | | |

1 - 60  
WEHRKGTDYHKQPWQAKISVIRPSKGHESCMGAVVSFYFVLTAAHCFVDDKEHSIKVSV

1 - 60 T T T S TTT S 3 TSS T 3 T S  
TT 3

120 70 80 90 100 110

| | | | |

61 - 120  
GGEKRDLEIEVVLFPNYNINGKKEAGIPEFYDYDVALIKLKNLKYGQTIRPICLPCTE

61 - 120 T SSS T T T TTT SSS  
130 140 150 160 170

180 | | | | |

121 - 180  
GTTRALRLPPTTTCQQQKEELLPAQDIKALFVSEEEKLTRKEVYIKNGDKKGGSCERDAQ

121 - 180 T SS T T T S  
THHHHHHH H

240 190 200 210 220 230

| | | | |

181 - 240  
YAPGYDKVKDISEVVTPRFLCTGGVSPYADPNTCRGDSGGPLIVHKRSRFIQVGVISWGV

181 - 240 T T T T T T T T SS H  
S S

250 260 270

241 - 278		VDVCKNQKRQKQVPAHARDFHINLQVLPWLKEKLQDE					
241 - 278		T	T	S		T	HHHHHHHHHHHH

**# 39 # Error: Abnormally short interatomic distances**

The pairs of atoms listed in the table below have an unusually short distance.

The contact distances of all atom pairs have been checked. Two atoms are said to 'bump' if they are closer than the sum of their Van der Waals radii minus 0.40 Angstrom. For hydrogen bonded pairs a tolerance of 0.55 Angstrom is used. The first number in the table tells you how much shorter that specific contact is than the acceptable limit. The second distance is the distance between the centers of the two atoms.

The last text-item on each line represents the status of the atom pair. The text 'INTRA' means that the bump is between atoms that are explicitly listed in the PDB file. 'INTER' means it is an inter-symmetry bump. If the final column contains the text 'HB', the bump criterium was relaxed because there could be a hydrogen bond. Similarly relaxed criteria are used for 1--3 and 1--4

interactions (listed as 'B2' and 'B3', respectively). If the last column is 'BF', the sum of the B-factors of the atoms is higher than 80, which makes the appearance of the bump somewhat less severe because the atoms probably aren't there anyway.

Bumps between atoms for which the sum of their occupancies is lower than one are not reported. In any case, each bump is listed in only one direction.

```
142 LEU ( 1 ) A  CG -- 157 LYS ( 150 ) A  CA  0.485  2.715 INTRA
142 LEU ( 1 ) A  CD1 -- 157 LYS ( 150 ) A  CA  0.450  2.750 INTRA
142 LEU ( 1 ) A  CB -- 157 LYS ( 150 ) A  CA  0.408  2.792 INTRA
  4 ARG ( 19 ) A  CB -- 161 ARG (  3 ) A  CA  0.320  2.880 INTRA
142 LEU ( 1 ) A  CD1 -- 157 LYS ( 150 ) A  C   0.277  2.923 INTRA
138 LYS ( 140 ) A  C -- 148 LYS (  7 ) A  NZ  0.270  2.830 INTRA
142 LEU ( 1 ) A  CD1 -- 157 LYS ( 150 ) A  N   0.257  2.843 INTRA
149 ALA (  8 ) A  O -- 150 LEU (  9 ) A  CD1  0.257  2.543 INTRA
140 GLU ( 142 ) A  O -- 142 LEU (  1 ) A  CD1  0.254  2.546 INTRA
138 LYS ( 140 ) A  C -- 148 LYS (  7 ) A  CE  0.237  2.963 INTRA
150 LEU (  9 ) A  CD1 -- 214 CYS ( 191 ) A  C   0.198  3.002 INTRA
142 LEU (  1 ) A  CG -- 157 LYS ( 150 ) A  C   0.184  3.016 INTRA
138 LYS ( 140 ) A  NZ -- 236 ILE ( 213 ) A  CD1  0.165  2.935 INTRA
142 LEU (  1 ) A  CD2 -- 157 LYS ( 150 ) A  C   0.135  3.065 INTRA
209 ALA (  6 ) A  CB -- 248 LYS ( 223 ) A  N   0.115  2.985 INTRA
```

138 LYS ( 140 ) A NZ -- 236 ILE ( 213 ) A CG1 0.101 2.999 INTRA  
 140 GLU ( 142 ) A C -- 142 LEU ( 1 ) A CD1 0.101 3.099 INTRA  
 161 ARG ( 3 ) A O -- 162 LYS ( 4 ) A CB 0.099 2.701 INTRA  
 48 THR ( 60 ) A C -- 67 LEU ( 83 ) A N 0.085 3.015 INTRA  
 148 LYS ( 7 ) A N -- 159 LEU ( 1 ) A CD1 0.070 3.030 INTRA  
 150 LEU ( 9 ) A O -- 214 CYS ( 191 ) A SG 0.058 2.942 INTRA  
 138 LYS ( 140 ) A N -- 148 LYS ( 7 ) A NZ 0.019 2.981 INTRA  
 138 LYS ( 140 ) A NZ -- 236 ILE ( 213 ) A CG2 0.004 3.096 INTRA  
 67 LEU ( 83 ) A O -- 68 GLU ( 84 ) A C 0.003 2.797 INTRA

# 40 # Warning: Abnormal packing environment for some residues

The residues listed in the table below have an unusual packing environment.

The packing environment of the residues is compared with the average packing environment for all residues of the same type in good PDB files. A low packing score can indicate one of several things: Poor packing, misthreading of the sequence through the density, crystal contacts, contacts with a co-factor, or the residue is part of the active site. It is not uncommon to see a few of these, but in any case this requires further inspection of the residue.

124 ARG ( 5 ) A -7.50



185 TYR ( 8 ) A -7.20  
9 TYR ( 24 ) A -6.90  
128 LEU ( 9 ) A -6.60  
127 ARG ( 8 ) A -6.46  
25 LYS ( 37 ) A -6.43  
145 GLN ( 4 ) A -6.17  
126 LEU ( 7 ) A -6.11  
215 ARG ( 192 ) A -6.07  
260 PHE ( 5 ) A -6.04  
27 HIS ( 39 ) A -6.03  
208 TYR ( 5 ) A -5.96  
83 LYS ( 96C ) A -5.91  
109 GLN ( 116 ) A -5.91  
57 LYS ( 73B ) A -5.56  
137 GLN ( 139 ) A -5.54  
189 LYS ( 12 ) A -5.53  
11 LYS ( 26 ) A -5.43  
112 ARG ( 119 ) A -5.36  
154 GLU ( 147 ) A -5.28  
263 ASN ( 8 ) A -5.25  
265 PHE ( 10 ) A -5.09  
65 ARG ( 5 ) A -5.07  
181 TYR ( 4 ) A -5.00

# 41 # Warning: Abnormal packing environment for sequential residues

A stretch of at least three sequential residues with a questionable packing environment was found. This could indicate that these residues are part of a strange loop, but might also be an indication of misthreading.

The table below lists the first and last residue in each stretch found, as well as the average residue score of the series.

126 LEU ( 7 ) A --- 128 LEU ( 9 ) A -6.39

# 42 # Error: Abnormal average packing environment

The average quality control value for the structure is very low.

A molecule is certain to be incorrect if the average quality score is below -3.0. Poorly refined molecules, very well energy minimized misthreaded molecules and low homology models give values between -2.0 and -3.0. The average quality of 200 highly refined X-ray structures was  $-0.5 \pm 0.4$  [REF].

Average for range 1 - 278 : -2.749

# 43 # Note: Quality value plot

The quality value smoothed over a 10 residue window is plotted as

function of the residue number. Low areas in the plot (below -2.0) indicate "unusual" packing.

In the TeX file, a plot has been inserted here

Chain identifier: A

# 44 # Warning: Low packing Z-score for some residues

The residues listed in the table below have an unusual packing environment according to the 2nd generation quality check. The score listed in the table is a packing normality Z-score: positive means better than average, negative means worse than average. Only residues scoring less than -2.50 are listed here. These are the "unusual" residues in the structure, so it will be interesting to take a special look at them.

137 GLN ( 139 ) A -3.04

251 LYS ( 226 ) A -3.01

17 LYS ( 4 ) A -2.95

198 ARG ( 179 ) A -2.93

27 HIS ( 39 ) A -2.91

191 ILE ( 14 ) A -2.88

24 SER ( 34B ) A -2.83

52 LYS ( 69 ) A -2.75

20 VAL ( 7 ) A -2.73

254 PRO ( 229 ) A -2.71

10 HIS ( 25 ) A -2.65

138 LYS ( 140 ) A -2.63

250 GLN ( 225 ) A -2.63

160 THR ( 2 ) A -2.62

11 LYS ( 26 ) A -2.61

16 ALA ( 3 ) A -2.53

**# 45 # Warning: Abnormal packing Z-score for sequential residues**

A stretch of at least four sequential residues with a 2nd generation packing Z-score below -1.75 was found. This could indicate that these residues are part of a strange loop or that the residues in this range are incomplete, but it might also be an indication of misthreading.

The table below lists the first and last residue in each stretch found, as well as the average residue Z-score of the series.

9 TYR ( 24 ) A --- 12 GLN ( 27 ) A -2.50

15 GLN ( 2 ) A --- 18 ILE ( 5 ) A -2.41

160 THR ( 2 ) A --- 164 VAL ( 6 ) A -2.14

**# 46 # Error: Abnormal structural average packing Z-score**

The quality control Z-score for the structure is very low.

A molecule is certain to be incorrect if the Z-score is below -5.0.

Poorly refined molecules, very well energy minimized misthreaded molecules and low homology models give values between -2.0 and -5.0. The average quality of properly refined X-ray structures is 0.0+/-1.0.

All contacts : Average = -1.050 Z-score = -6.79

BB-BB contacts : Average = -0.378 Z-score = -2.69

BB-SC contacts : Average = -1.106 Z-score = -6.00

SC-BB contacts : Average = -0.248 Z-score = -1.34

SC-SC contacts : Average = -0.929 Z-score = -4.80

# 47 # Note: Second generation quality Z-score plot

The second generation quality Z-score smoothed over a 10 residue window is plotted as function of the residue number. Low areas in the plot (below -1.3) indicate "unusual" packing.

In the TeX file, a plot has been inserted here

Chain identifier: A

# 48 # Warning: Backbone oxygen evaluation

The residues listed in the table below have an unusual backbone oxygen position.

For each of the residues in the structure, a search was performed to find 5-residue stretches in the WHAT IF database with superposable C-alpha coordinates, and some constraints on the neighboring backbone oxygens.

In the following table the RMS distance between the backbone oxygen positions of these matching structures in the database and the position of the backbone oxygen atom in the current residue is given. If this number is larger than 1.5 a significant number of structures in the database show an alternative position for the backbone oxygen. If the number is larger than 2.0 most matching backbone fragments in the database have the peptide plane flipped. A manual check needs to be performed to assess whether the experimental data can support that alternative as well. The number in the last column is the number of database hits (maximum 80) used in the calculation. It is "normal" that some glycine residues show up in this list, but they are still worth checking!

82 GLY ( 96B) A 2.82 15

# 49 # Warning: Unusual rotamers

The residues listed in the table below have a rotamer that is not seen very often in the database of solved protein structures. This option determines for every residue the position specific chi-1 rotamer distribution. Thereafter it verified whether the actual residue in the molecule has the most preferred rotamer or not. If the actual rotamer is the preferred one, the score is 1.0. If the actual rotamer is unique, the score is 0.0. If there are two preferred rotamers, with a population distribution of 3:2 and your rotamer sits in the lesser populated rotamer, the score will be 0.66. No value will be given if insufficient hits are found in the database.

It is not necessarily an error if a few residues have rotamer values below 0.3, but careful inspection of all residues with these low values could be worth it.

98 LEU ( 105 ) A 0.38

123 THR ( 4 ) A 0.39

# 50 # Warning: Unusual backbone conformations

For the residues listed in the table below, the backbone formed by itself and two neighboring residues on either side is in a conformation that is not seen very often in the database of solved protein structures. The number given in the table is the number of

similar backbone conformations in the database with the same amino acid in the center.

For this check, backbone conformations are compared with database structures using C-alpha superpositions with some restraints on the backbone oxygen positions.

A residue mentioned in the table can be part of a strange loop, or there might be something wrong with it or its directly surrounding residues. There are a few of these in every protein, but in any case it is worth looking at!

3 HIS ( 18 ) A 0  
8 ASP ( 23 ) A 0  
9 TYR ( 24 ) A 0  
12 GLN ( 27 ) A 0  
14 TRP ( 1 ) A 0  
15 GLN ( 2 ) A 0  
18 ILE ( 5 ) A 0  
19 SER ( 6 ) A 0  
20 VAL ( 7 ) A 0  
21 ILE ( 33 ) A 0  
24 SER ( 34B ) A 0  
25 LYS ( 37 ) A 0



27 HIS ( 39 ) A 0  
28 GLU ( 40 ) A 0  
46 CYS ( 58 ) A 0  
51 ASP ( 62A ) A 0  
54 HIS ( 71 ) A 0  
55 SER ( 72 ) A 0  
56 ILE ( 73 ) A 0  
57 LYS ( 73B ) A 0  
58 VAL ( 73C ) A 0  
59 SER ( 74 ) A 0  
60 VAL ( 75 ) A 0  
63 GLU ( 78 ) A 0  
64 LYS ( 79 ) A 0

And so on for a total of 126 lines

# 51 # Error: Backbone conformation Z-score very low

A comparison of the backbone conformation with database proteins shows that the backbone fold in this structure is very unusual.

Backbone conformation Z-score : -9.841

# 52 # Warning: Average B-factor problem

The average B-factor for all buried protein atoms normally lies between 10--20. Values around 3--5 are expected for X-ray studies performed

at liquid nitrogen temperature.

Because of the extreme value for the average B-factor, no further analysis of the B-factors is performed.

Average B-factor for buried atoms : 0.000

# 53 # Warning: B-factor plot impossible

All average B-factors are zero. Plot suppressed.

Chain identifier: A

# 54 # Error: HIS, ASN, GLN side chain flips

Listed here are histidine, asparagine or glutamine residues for which the orientation determined from hydrogen bonding analysis are different from the assignment given in the input. Either they could form energetically more favorable hydrogen bonds if the terminal group was rotated by 180 degrees, or there is no assignment in the input file (atom type 'A') but an assignment could be made. If a residue is marked "flexible" the flipped conformation is only slightly better than the non-flipped conformation.

10 HIS ( 25 ) A

12 GLN ( 27 ) A

135 GLN ( 137 ) A

137 GLN ( 139 ) A

# 55 # Note: Histidine type assignments

For all complete HIS residues in the structure a tentative assignment to HIS-D (protonated on ND1), HIS-E (protonated on NE2), or HIS-H (protonated on both ND1 and NE2, positively charged) is made based on the hydrogen bond network. A second assignment is made based on which of the Engh and Huber [REF] histidine geometries fits best to the structure.

In the table below all normal histidine residues are listed. The assignment based on the geometry of the residue is listed first, together with the RMS Z-score for the fit to the Engh and Huber parameters. For all residues where the H-bond assignment is different, the assignment is listed in the last columns, together with its RMS Z-score to the Engh and Huber parameters.

As always, the RMS Z-scores should be close to 1.0 if the residues were restrained to the Engh and Huber parameters during refinement.

Please note that because the differences between the geometries of the different types are small it is possible that the geometric assignment given here does not correspond to the type used in

refinement. This is especially true if the RMS Z-scores are much higher than 1.0.

If the two assignments differ, or the "geometry" RMS Z-score is high, it is advisable to verify the hydrogen bond assignment, check the HIS type used during the refinement and possibly adjust it.

3	HIS ( 18 ) A	HIS-E	2.05	HIS-D	2.53
10	HIS ( 25 ) A	HIS-E	2.54	HIS-D	2.64
27	HIS ( 39 ) A	HIS-E	2.84		
45	HIS ( 57 ) A	HIS-D	2.84	HIS-E	2.94
54	HIS ( 71 ) A	HIS-D	2.84	HIS-E	2.96
75	HIS ( 91 ) A	HIS-D	2.08	HIS-E	2.58
225	HIS ( 202 ) A	HIS-D	2.85	HIS-E	2.89
256	HIS ( 1 ) A	HIS-E	3.02	HIS-D	3.16
261	HIS ( 6 ) A	HIS-E	2.43	HIS-D	2.73

# 56 # Warning: Buried unsatisfied hydrogen bond donors

The buried hydrogen bond donors listed in the table below have a hydrogen atom that is not involved in a hydrogen bond in the optimized hydrogen bond network.

Hydrogen bond donors that are buried inside the protein normally use all of their hydrogens to form hydrogen bonds within the

protein. If there are any non hydrogen bonded buried hydrogen bond donors in the structure they will be listed here. In very good structures the number of listed atoms will tend to zero.

1 TRP ( 16 ) A N  
1 TRP ( 16 ) A NE1  
2 GLU ( 17 ) A N  
3 HIS ( 18 ) A N  
4 ARG ( 19 ) A N  
4 ARG ( 19 ) A NE  
4 ARG ( 19 ) A NH2  
5 LYS ( 20 ) A N  
14 TRP ( 1 ) A N  
14 TRP ( 1 ) A NE1  
17 LYS ( 4 ) A N  
20 VAL ( 7 ) A N  
32 GLY ( 44 ) A N  
38 TYR ( 50 ) A N  
44 ALA ( 56 ) A N  
47 PHE ( 59 ) A N  
52 LYS ( 69 ) A N  
61 GLY ( 76 ) A N  
65 ARG ( 5 ) A N  
67 LEU ( 83 ) A N

68 GLU ( 84 ) A N

86 ALA ( 97 ) A N

92 TYR ( 99 ) A N

92 TYR ( 99 ) A OH

93 ASP ( 100 ) A N

And so on for a total of 84 lines

**# 57 # Warning: Buried unsatisfied hydrogen bond acceptors**

The buried side-chain hydrogen bond acceptors listed in the table below are not involved in a hydrogen bond in the optimized hydrogen bond network.

Side-chain hydrogen bond acceptors that are buried inside the protein normally form hydrogen bonds within the protein. If there are any not hydrogen bonded in the optimized hydrogen bond network they will be listed here.

15 GLN ( 2 ) A OE1

28 GLU ( 40 ) A OE1

120 GLU ( 1 ) A OE2

140 GLU ( 142 ) A OE1

163 GLU ( 5 ) A OE1

176 GLU ( 169 ) A OE1

212 ASN ( 189 ) A OD1

# 58 # Note: Summary report for users of a structure

This is an overall summary of the quality of the structure as compared with current reliable structures. This summary is most useful for biologists seeking a good structure to use for modelling calculations.

The second part of the table mostly gives an impression of how well the model conforms to common refinement constraint values. The first part of the table shows a number of constraint-independent quality indicators.

Structure Z-scores, positive is better than average:

1st generation packing quality : -5.622 (poor)

2nd generation packing quality : -6.790 (bad)

Ramachandran plot appearance : -5.194 (bad)

chi-1/chi-2 rotamer normality : -4.444 (bad)

Backbone conformation : -9.841 (bad)

RMS Z-scores, should be close to 1.0:

Bond lengths : 1.806 (loose)

Bond angles : 2.987 (loose)

Omega angle restraints : 3.386 (loose)

Side chain planarity : 3.505 (loose)  
Improper dihedral distribution : 4.153 (loose)  
Inside/Outside distribution : 1.243 (unusual)

## REFERENCES

=====

### WHAT IF

G.Vriend,

WHAT IF: a molecular modelling and drug design program,

J. Mol. Graph. 8, 52--56 (1990).

### WHAT\_CHECK (verification routines from WHAT IF)

R.W.W.Hooft, G.Vriend, C.Sander and E.E.Abola,

Errors in protein structures

Nature 381, 272 (1996).

### Bond lengths and angles, protein residues

R.Engl and R.Huber,

Accurate bond and angle parameters for X-ray protein structure  
refinement,

Acta Crystallogr. A47, 392--400 (1991).

### Bond lengths and angles, DNA/RNA



G.Parkinson, J.Voitechovsky, L.Clowney, A.T.Bruenger and H.Berman,  
New parameters for the refinement of nucleic acid-containing structures  
Acta Crystallogr. D52, 57--64 (1996).

#### DSSP

W.Kabsch and C.Sander,  
Dictionary of protein secondary structure: pattern  
recognition of hydrogen bond and geometrical features  
Biopolymers 22, 2577--2637 (1983).

#### Hydrogen bond networks

R.W.W.Hooft, C.Sander and G.Vriend,  
Positioning hydrogen atoms by optimizing hydrogen bond networks in  
protein structures  
PROTEINS, 26, 363--376 (1996).

#### Matthews' Coefficient

B.W.Matthews  
Solvent content of Protein Crystals  
J. Mol. Biol. 33, 491--497 (1968).

#### Protein side chain planarity

R.W.W. Hooft, C. Sander and G. Vriend,  
Verification of protein structures: side-chain planarity

J. Appl. Cryst. 29, 714--716 (1996).

#### Puckering parameters

D.Cremer and J.A.Pople,

A general definition of ring puckering coordinates

J. Am. Chem. Soc. 97, 1354--1358 (1975).

#### Quality Control

G.Vriend and C.Sander,

Quality control of protein models: directional atomic  
contact analysis,

J. Appl. Cryst. 26, 47--60 (1993).

#### Ramachandran plot

G.N.Ramachandran, C.Ramakrishnan and V.Sasisekharan,

Stereochemistry of Polypeptide Chain Conformations

J. Mol. Biol. 7, 95--99 (1963).

#### Symmetry Checks

R.W.W.Hoof, C.Sander and G.Vriend,

Reconstruction of symmetry related molecules from protein  
data bank (PDB) files

J. Appl. Cryst. 27, 1006--1009 (1994).



## 9 References

- <sup>1</sup> J.T. Edsall, P.J. Flory, J.C. Kendrew, A.M. Liquori, G. Nemethy, G.N. Ramachandran, M.A. Scheraga  
IUPAC - IUB Commission of biochemical Nomenclature  
Journal Molecular Biology 15: 399 - 407, (1966)
- <sup>2</sup> J.T. Edsall, P.J. Flory, J.C. Kendrew, A.M. Liquori, G. Nemethy, G.N. Ramachandran, M.A. Scheraga  
IUPAC - IUB Commission of biochemical Nomenclature  
Journal Biological Chemistry 241: 1004 - 1008, (1966)
- <sup>3</sup> J.T. Edsall, P.J. Flory, J.C. Kendrew, A.M. Liquori, G. Nemethy, G.N. Ramachandran, M.A. Scheraga  
IUPAC - IUB Commission of biochemical Nomenclature  
Journal Molecular Biology 52: 1 17, (1970)
- <sup>4</sup> J.T. Edsall, P.J. Flory, J.C. Kendrew, A.M. Liquori, G. Nemethy, G.N. Ramachandran, M.A. Scheraga  
IUPAC - IUB Commission of biochemical Nomenclature  
Biochemistry Journal 121: 577 - 585, (1971)
- <sup>5</sup> R.B. Corey, L. Pauling  
Proceedings Royal Society London, B. 141: 10, (1953)
- <sup>6</sup> F.A. Momany, R.F. McGuire, A.W. Burgess, H.A. Scheraga  
Journal Physical Chemistry 79: 2361 - 2381, (1975)
- <sup>7</sup> E. Baker, R. Hubbard  
Hydrogen bonding in globular proteins  
Progress Biophysical Molecular Biology 44: 97 - 179, (1984)
- <sup>8</sup> R. Sheridan, R. Lee, N. Peters, L. Allen  
Hydrogen bond co-operativity in protein secondary structure  
Biopolymers 18: 2451 - 2458, (1979)
- <sup>9</sup> E. Baker, R. Hubbard  
Hydrogen bonding in globular proteins  
Progress Biophysical Molecular Biology 44: 97 - 179, (1984)
- <sup>10</sup> S.K.A. Law, K.B.M. Reid  
Complement  
IRL Press Ltd (1988)
- <sup>11</sup> G. D. Ross (Editor)

---

Chapter 2. The alternative Pathway

Immunobiology of the complement System, 45 - 62. Academic Press (1986).

<sup>12</sup> R.A. Good, N.K. Day

Biological amplification systems in immunobiology

Plenum Publishing Corporation, (1977)

<sup>13</sup> L.T.J. Delbaere, W.L.B. Hutcheon, M.N.G. James, W.E. Theissen

Nature 257, 758 (1975)

<sup>14</sup> C.S. Wright, R.A Alden, J. Kraut

Nature 221, 235 (1969)

<sup>15</sup> W.R. Kester, B.W. Matthews

Biochemistry, 16, 2506 (1977)

<sup>16</sup> W.R. Kester, B.W. Matthews

Journal of Biological Chemistry vol. 252, p7704 (1977)

<sup>17</sup> P. Argos, R.M. Garavito, W. Eventoff, M.G. Rosman, C. -I. Brändén

Journal of Molecular Biology, vol. 126, p141 (1978)

<sup>18</sup> B.W. Matthews, S.J. Remington, M.G. Grutter, W.F. Anderson

Journal of Molecular Biology, Vol. 147, p545 (1981)

<sup>19</sup> B.W. Matthews, S.J. Remington, M.G. Grutter, W.F. Anderson

Journal of Molecular Biology, Vol. 147, p545 (1981)

<sup>20</sup> M.G. Rossman, P. Argos

Journal of Molecular Biology, vol. 105 p75 (1976)

<sup>21</sup> S. Magnusson

Thrombin and Prothombin

The Enzymes , Vol. 3, p277 (1971) Academic Press New York.

<sup>22</sup> T.J. Leighton, R.H. Doi, A.J. Warren, R.A. Kelln

The relationship of serine protease activity to RNA polymerase modifications and sporulation in *Bacillus subtilis*

Journal of Molecular Biology, Vol. 76, p103 (1973)

<sup>23</sup> R. Stambaugh, B. Brackett, L. Mastroianni

Inhibition of in vitro fertilisation of rabbit ova by Trypsin inhibitors

Biological Reproduction, Vol. 1, p 223 (1969)

<sup>24</sup> J. Fastrez, A.R. Fersht

Biochemistry Vol. 12, p 2025 (1973)

- 
- <sup>25</sup> D.M. Blow, J.J. Birktoft, B.S. Hartley  
Nature, Vol. 221, p337 (1969)
- <sup>26</sup> M.W. Hunkapiller, S.H. Smallcombe, D.R. Whitaker, J.H. Richards  
Carbon nuclear magnetic resonance studies of the histidine residue in  $\alpha$ -lytic protease  
Biochemistry Vol. 12, p 4732 (1973)
- <sup>27</sup> P.W. Inward, W.P. Jencks  
The reactivity of nucleophilic reagents with furoyl-chymotrypsin  
Journal Biological Chemistry, Vol. 240, p 1986 (1965)
- <sup>28</sup> W.P. Jencks  
Reactivity correlations and general acid - base catalysis in enzymatic transacylation reactions  
Cold Spring Harbour Symposium Vol. 36 p 1 (1971)
- <sup>29</sup> W.P. Jencks  
General acid base catalysis of complex reactions in water  
Chemical review Vol. 72, p 705
- <sup>30</sup> R. Henderson  
Structure of crystalline  $\alpha$ -Chymotrypsin. IV. Structure of indoleacryloyl- $\alpha$ -chymotrypsin and its relevance to the hydrolytic mechanism of the enzyme.  
Journal of Molecular Biology, Vol. 54, p 341 (1970)
- <sup>31</sup> D.M. Segal, J.C. Powers, G.H. Cohen, D.R. Davies, P.E. Wilcox  
Substrate binding site in bovine chymotrypsin A $\gamma$ . A crystallographic study using peptide chloromethyl ketones as site specific inhibitors  
Biochemistry vol 10, p 3728 (1971)
- <sup>32</sup> M.H. O'Leary, M.D. Kluetz  
Nitrogen isotope effects on the Chymotrypsin catalysed hydrolysis of N-acetyl-tryptophanamide  
Journal of the American Chemical Society Vol 94, p 3585 (1972)
- <sup>33</sup> A. Warshel, G. Naray-Szabo, J. Sussman, J.K. Hwang  
How Do Serine Proteases Really Work?  
Biochemistry, vol 28, No 9, p 3629 (1989)
- <sup>34</sup> M.J.S. Dewar, D.M. Storch  
Proceedings Natural Academy of Science (USA), vol 82, p 2225 - 2229
- <sup>35</sup> H. Umeyama, A. Nakagawa, T. Kudo  
Journal of Molecular Biology, vol 150, p 409 - 421

- 
- <sup>36</sup> G. Natay-Szabo  
International Journal of Quantum Chemistry, vol 23, p 723 - 728
- <sup>37</sup> A. Warshal, S. Russell  
Journal of the American Chemical Society, vol 108, p 6569 - 6579
- <sup>38</sup> S. Sprang, T. Standing, R.J. Fletterick, J. Finer-Moore, R.M. Stroud, N.H. Xuong, R. Hamlin, W.J. Rutter, C.S. Craik  
Science, vol 237, p 905 - 907
- <sup>39</sup> C. Gaboriaud, L. Serre, O. Guy-Crotte, E. Forest, J.C. Fontecilla-Camps  
Crystal structure of human Trypsin : Unexpected phosphorylation of tyrosine  
To Be Published
- <sup>40</sup> M. Marquart, J. Walter, J. Deisenhofer, W. Bode, R. Huber  
The geometry of the reactive site and of the peptide groups in Trypsin, Trypsinogen and its complexes with inhibitors  
Acta Crystallography, section B, vol 39, 480 (1983)
- <sup>41</sup> Q. Huang, Z. Wang, Y. Li, S. Liu, Y. Tang  
Refined 1.8Å resolution crystal structure of porcine ε-Trypsin  
Biochim. Biophys. Acta vol 1209, 77 (1994)
- <sup>42</sup> R.J. Read, M.N.G. James  
Refined crystal structure of Streptomyces Griseus Trypsin at 1.7Å resolution  
Journal Molecular Biology vol 200, 523 (1988)
- <sup>43</sup> H. Tsukada, D.M. Blow  
Structure of α-Chymotrypsin refined at 1.68Å resolution  
Journal Molecular Biology vol 184, 703 (1985)
- <sup>44</sup> M.M. Dixon, B.W. Matthews  
Is γ-Chymotrypsin a tetrapeptide acyl-enzyme adduct of γ-Chymotrypsin?  
Biochemistry vol 28, 7033 (1989)
- <sup>45</sup> A.Z. Wei, I. Mayr, W. Bode  
The refined 2.3Å crystal structure of human leukocyte Elastase in a complex with a valine chloromethyl ketone inhibitor  
FEBS Letters vol 234, 367 (1988)
- <sup>46</sup> X. Ding, B.F. Rasmussen, G.A. Petsko, D. Ringe  
Direct structure observation of an acyl - enzyme intermediate in the hydrolysis of an ester substrate by elastase  
To be published
- <sup>47</sup> G.I. Berglund, N.P. Willassen, A. Hordvick, A.O. Smalaas

---

Structure of native pancreatic elastase from North Atlantic salmon at 1.61Å resolution  
To be published

<sup>48</sup> R. Bone, A.B. Shenvi, C.A. Kettner, D.A. Agard  
Serin protease mechanism. Structure of an inhibitory complex of  $\alpha$  - lytic protease and a tightly bound peptide boronic acid  
Biochemistry vol 26, 7609 (1987)

<sup>49</sup> M.N.G. James, A.R. Sielecki, G.D. Brayer, L.T.J. Delbaere, C.A. Bauer  
Structures of product and inhibitor complexes of streptomyces griseus protease A at 1.8Å resolution. A model for serine protease catalysis  
Journal Molecular Biology vol 144, 43 (1980)

<sup>50</sup> R.J. Read, M. Fujinaga, A.R. Seilecki, M.N.G. James  
Structure of the complex of streptomyces griseus protease B and the third domain of the turkey ovomucoid inhibitor at 1.8Å resolution  
Biochemistry vol 22, 4420 (1983)

<sup>51</sup> G.S. Spraggon, C. Phillips, U.K. Nowak, C.P. Ponting, D. Saunders, C.M. Dobson, D.I. Stuart, E.Y. Jones  
The crystal structure of the catalytic domain of human urokinase - type plasminogen activator  
Structure (London) vol 3, 681 (1995)

<sup>52</sup> W. Bode, Z. Chen, K. Bartels, C. Kutzbach, G. Schmidt-Kastner, H. Bartunik  
Refined 2Å X-ray crystal structure of porcine pancreatic Kalikrein A, a specific Trypsin - like serine protease. Crystallization, structure determination, crystallographic refinement, structure and its comparison with bovine Trypsin  
Journal Molecular Biology vol 164, 237 (1983)

<sup>53</sup> M. Fuginaga, M.N.G. James  
Rat submaxillary gland serine protease, Tonin. Structure solution and refinement at 1.8Å resolution  
Journal Molecular Biology vol 195, 373 (1987)

<sup>54</sup> T.J. Rydel, M. Yin, K.P. Padmanabhan, D.T. Blankenship, A.D. Cardin, P.E. Correa, J.W. Fenton II, A. Tulinsky  
Crystallographic structure of human  $\gamma$  - Thrombin  
Journal Molecular Biology vol 269, 22000 (1994)

<sup>55</sup> K. Kitadokoro, H. Tsuzuki, E. Nakamura, T. Sato, H. Toraoka  
Purification, characterisation, primary structure, crystallization and preliminary crystallographic study of a serine proteinase from streptomyces fradiae ATCC 14544  
European Journal Biochemistry vol 220, 55 (1994)

<sup>56</sup> R.A. Alden, J.J. Birktoft, J. Kraut, J.D. Robertus



---

Atomic coordinates for Subtilisin / BPN

Biochem. Biophys. Res. Comm. vol 45, 337 (1971)

<sup>57</sup> T. Yamane, T. Kani, T. Hatanaka, A. Suzuki, T. Ashida, T. Kobayashi, S. Ito, O. Yamashita

Crystal structure of a new alkaline serine protease (M-Protease) from Bacillus SP. KSM-K16

To be published

<sup>58</sup> C. Betzel, G.P. Pal, W. Saenger

Synchrotron X-ray data collection and restrained least squares refinement of the crystal structure of proteinase K at 1.5Å resolution

Acta Crystallography, Section B vol 44, 163 (1988)

<sup>59</sup> A.V. Teplyakov, I.P. Kuranova, E.H. Harutyunyan, B.K. Vainshtein, C. Frommel, W.E. Hoehne, K.S. Wilson

Crystal structure of Thermitase at 1.4Å resolution

Journal Molecular Biology vol 214, 261 (1990)

<sup>60</sup> M. Laskowski, R.W. Sealock

The Enzymes, vol 3, p375 (1971)

<sup>61</sup> M. Rigby

Proceedings of the International Conference on Proteinase Inhibitors

Walter de Gruyter, Berlin, p 117 (1971).

<sup>62</sup> T.A. Steitz, R. Henderson, D.M. Blow

Journal of Molecular Biology, vol 46, p 337 (1969)

<sup>63</sup> J.J. Peronna, C.S. Craik

Structural basis of substrate specificity in the serine proteases

Protein Science, vol 4, p 337 - 360 (1995)

<sup>64</sup> L. Graf C.S. Craik, A. Pathy, S. Roczniak, R.J. Flaterrick, W.J. Rutter

Selective alteration of substrate specificity by replacement of Asp 189 with Lys in the binding pocket of Trypsin

Biochemistry, vol 26, p 2616 - 2623 (1987)

<sup>65</sup> L. Graf, A. Jansco, L. Szilagyi, G. Hegyi, K. Pinter, G. Naray-Szabo, J. Hepp

Electrostatic complementarity in the substrate binding pocket of Trypsin

Proceedings of the Natural Academy of Science U.S.A., vol 85, p 4961 - 4965 (1988)

<sup>66</sup> J.J. Perona, L. Hedstrom, R. Wagner, W.J. Rutter, C.S. Craik, R.J. Flatterik

Exogenous acetate reconstitutes the enzymatic activity of Asp 189 Ser Trypsin

Biochemistry vol 33, p 3252 - 3259 (1994)

- 
- <sup>67</sup> J.J. Perona, L. Hedstrom, W.J. Rutter, R.J. Fletterick  
Structural origins of substrate discrimination in Trypsin and Chymotrypsin  
Biochemistry vol 34, p 1489 - 1499 (1995)
- <sup>68</sup> A.R. Fersht, D.M. Blow, J. Fastrez  
Biochemistry, vol 12, p 2035 (1973)
- <sup>69</sup> D.H. Andrews  
Physics Review, 1930, 36, 544
- <sup>70</sup> S.J. Weiner, P.A. Kollman, D.A. Case, U.C. Singh, C. Ghio, G. Alagona, S. Profeta Jr,  
P. Weiner  
A new force field for molecular mechanics simulation of nucleic acids and proteins  
Journal American Chemical Society 106: 765 - 785, (1984)
- <sup>71</sup> D.H.R. Barton  
Experimentia 1950, 6, 316
- <sup>72</sup> J.A. Pole, D.L. Beveridge  
"Approximate Molecular Orbital Theory", McGraw-Hill, New York, 1970
- <sup>73</sup> A.T. Hagler, S.T. Stern, S. Lifson, S. Ariel  
Journal American Chemical Society, 1979, 101, 813
- <sup>74</sup> N.L. Allinger, Y.H. Young, L. Jenn Huei  
Journal American Chemical Society, 1989, 111, 8551
- <sup>75</sup> I. Dostrovsky, E.D. Hughes, C.K. Ingold  
Journal Chemistry Society, 1946, 173
- <sup>76</sup> T.L. Hill  
Journal Chemical Physics, 1946, 14, 465
- <sup>77</sup> F.H. Xestheimer, J.E. Mayer  
Journal Physical Chemistry, 1946, 14, 733
- <sup>78</sup> D.H. Wertz, N.L. Allinger  
Tetrahedron 1974, 30, 1579
- <sup>79</sup> E.A. Mason, M.M. Kreevoy  
Journal American Chemical Society, 1955, 77, 5808
- <sup>80</sup> J.M. Lehn, G. Ourisson Bull  
Soc. Chim. Fr., 1963, 1113
- <sup>81</sup> N.L. Allinger

- <sup>82</sup> J.E. Lennard - Jones  
Procedures Royal Society London, Ser A, 1924, 106, 463
- <sup>83</sup> A.D. Buckingham, B.D. Witting  
Annual Review, Physical Chemistry, 1970, 21, 287
- <sup>84</sup> H. Hargenau, N.R. Kestner  
"Theory of Intermolecular Forces", Pergamon, Oxford, 1969
- <sup>85</sup> D.E. Williams  
Journal Chemical Physics, 1965, 43, 4424
- <sup>86</sup> D.E. Williams, T.L. Starr  
Computational Chemistry, 1977, 1, 173
- <sup>87</sup> E.K. Davies, N.W. Murrall  
Computational Chemistry 1989, 2, 149
- <sup>88</sup> J.G. Vinter, A. Davies, M.R. Saunders  
Journal Computational Aided Molecular Design, 1987, 1, 31
- <sup>89</sup> D.N.J. White, M.J. Bovill  
Journal Chemistry Society, Perkins, 1977, II, 1610
- <sup>90</sup> D.N.J. White, J.N. Ruddock, P.R. Eddington  
Molecular Simulation, 1989, 3, 71
- <sup>91</sup> S. Lifson, A. Warshel  
Journal Chemical Physics, 1968, 49, 5116
- <sup>92</sup> E.J. Jacob, H.B. Thomson, L.S. Bartell  
Journal Chemical Physics, 1967, 47, 3736
- <sup>93</sup> D.N.J. White G.A. Sim  
Tetrahedron, 1973, 29, 3933
- <sup>94</sup> M.O. Dayhoff et al  
Atlas of Protein Sequence and Structure 5: 345 - 362, (1978)
- <sup>95</sup> S. Henikoff, J.G. Henikoff  
Proceedings National Academy Science 89: 10915 - 10919, (1992)
- <sup>96</sup> Needleman, Wunsch  
Journal Molecular Biology 48: 443 - 453, (1970)

- 
- <sup>97</sup> F.C. Fernstein, T.F. Koeltz, G.J.B. Williams, E.F.R. Meyer, M.D. Brice  
The Protein Databank: A Computer Based Archival File For Macromolecules  
Journal Molecular Biology 112: 535 - 542, (1971)
- <sup>98</sup> T.F. Smith, M.S. Waterman  
Journal Molecular Biology 147: 195 - 197, (1981)
- <sup>99</sup> A. D. McLachlan, D.M. Shotton  
Structural similarities between  $\alpha$ -lytic protease of Myxobacter 495 and elastase.  
Nature 229: 202 - 205, (1971)
- <sup>100</sup> J. Greer  
Comparative model building of the mammalian serine proteases  
Journal Molecular Biology 153: 1027 - 1042, (1981)
- <sup>101</sup> J. Greer  
Comparative modelling Methods: Application to the family of the Mammalian serine proteases  
Proteins 7: 317 - 334, (1990)
- <sup>102</sup> W.J. Brown, A.C.T. North, D.C. Philips, K. Brew, T.C. Vanaman, R.L.A. Hill  
A possible three dimensional structure of bovine  $\alpha$ -lactalbumin based on that of hen's egg white lysozyme  
Journal Molecular Biology 42: 65 - 86, (1969)
- <sup>103</sup> C. Chothia, A.M. Lesk  
The Relation Between The Divergence Of Sequence And Structure In Proteins  
EMBO Journal 5: 823 - 826. (1986)
- <sup>104</sup> T.J.P. Hubbard, T.L. Blundell  
Comparison Of Solvent Inaccessible Cores Of Homologous Proteins  
Protein Engineering 1: 159 - 171. (1987)
- <sup>105</sup> S.S. Zimmermann, M.S. Pottle, G. Nemethy, H.A. Scheraga  
Macromolecules 10, 1
- <sup>106</sup> J. Christie, D.L. Gagnon  
Isolation, Characterisation and N - Terminal Sequence of the CNBr - Cleavage Peptides from Human Complement Factor B  
Journal of Biochemistry 201, 555 - 567, 1982
- <sup>107</sup> J. Christie, D.L. Gagnon  
Amino Acid Sequence of the Bb Fragment From Human Complement FB  
Journal Biochemistry 209, 61 - 70, 1983

- 
- <sup>108</sup> D.L. Christie, J. Gagnon, R.R. Porter  
Partial Sequence of Human Complement Component FB: Novel Type of Serine Protease  
Proceedings National Academy Science USA 77, 4923 - 4927 1980
- <sup>109</sup> A. Sager, W.E. Mayer, J. Klein  
A complement Factor B like cDNA clone from zebrafish (brachydanio-rerio)  
Molecular Immunology 33: 511 - 520, (1996)
- <sup>110</sup> J. Greer  
Comparative model building of the mammalian serine proteases  
Journal Molecular Biology 153: 1027 - 1042, (1981)
- <sup>111</sup> J. Greer  
Comparative modelling methods: application to the family of the mammalian proteases  
Proteins 7: 317 - 334, (1990)
- <sup>112</sup> P. Chou, G. Fasman  
Biochemistry (74) 13 222 - 245
- <sup>113</sup> Biosym Technologies,  
9685 Scranton Road,  
San Diego,  
California 92121 - 2777, USA
- <sup>114</sup> Computational results obtained using software from Biosym Technologies of San Diego  
- mechanics / dynamics calculations were done with the Discover ® program, using the  
CVFF force field, graphical displays were printed out from the InsightII ® molecular  
modelling system
- <sup>115</sup> G. Vriend  
WHAT IF: A Molecular Modelling And Drug Design Program  
Journal Molecular Graphics 8: 52 - 56, (1990)
- <sup>116</sup> R.W.W. Hooft, G. Vriend, C. Sander, E.E. Abola  
Errors In Protein Structures  
Nature 381: 272, (1996)
- <sup>117</sup> G.K. Farber, G.A. Petsko  
The Evolution of a/b Barrel Enzymes  
Trends in Biochemistry Science 15: 228 - 234. (1990)
- <sup>118</sup> W. Basch, H.G. Mannherz, D. Suck, E.F. Pai, K.C. Holms  
Atomic Structure Of The Actin : DNase I Complex  
Nature 347: 37 - 44. (1990)

- 
- <sup>119</sup> C. Chothia, A.M. Lesk  
The Relation Between The Divergence Of Sequence And Structure In Proteins  
EMBO Journal 5: 823 - 826. (1986)
- <sup>120</sup> T.J.P. Hubbard, T.L. Blundell  
Comparison Of Solvent Inaccessible Cores Of Homologous Proteins  
Protein Engineering 1: 159 - 171. (1987)
- <sup>121</sup> C. Chothia  
One Thousand Families For the Molecular Biologist  
Nature 357: 543 - 544. (1992)
- <sup>122</sup> C. Chothia  
One Thousand Families For the Molecular Biologist  
Nature 357: 543 - 544. (1992)
- <sup>123</sup> A.V. Finkelstein, A.M. Gutun, A.Y. Badretelinov  
Why Are The Same Protein Folds Used To Perform Different Functions  
Febs Letters 325: 23 - 28, (1993)
- <sup>124</sup> J. Greer  
Comparative Model Building Of The Mammalian Serine Proteases  
Journal Molecular Biology 153: 1027 - 1042, (1981)
- <sup>125</sup> J. Greer  
Comparative modelling Methods: Applications To The Family Of The Mammalian Serine  
Proteases  
Proteins 7: 317 - 334, (1990)
- <sup>126</sup> C.M.R. Lemer, M.J. Rooman, S.J. Wodak  
Protein Structure Prediction By Threading Methods: Evaluation Of Current Techniques  
Proteins 23: 337 - 355, (1995)
- <sup>127</sup> M.S. Johnson, J.P. Overington, T.L. Blundell  
Alignment And Searching For Common Protein Folds Using A Databank Of Structural  
Templates  
Journal Molecular Biology 231: 735 - 752, (1993)
- <sup>128</sup> S.J. Wodak, M.J. Rooman  
Generating And Testing Protein Folds  
Current Opinion In Structural Biology 3: 247 - 259, (1993)
- <sup>129</sup> C.A. Orengo, T.P. Flores, W.R. Taylor, J.M. Thornton  
Identification and Classification Of Protein Family Folds  
Protein Engineering 6: 485 - 500. (1993)

- 
- <sup>130</sup> J.U. Bowie, N.D. Clarke, C.O. Pabo, R.T. Sauer  
Identification Of Protein Folds: Matching Hydrophobicity Patterns Of Sequence Sets With Solvent Accessibility Patterns Of Known Structure  
Proteins 7: 275 - 264, (1990)
- <sup>131</sup> R. Luthy, A.D. McLachlan, D. Eisenberg  
Secondary Structure Based Profiles: Use Of Structure Conserving Scoring Tables In Searching Protein Sequence Databases For Structure Similarities  
Proteins 10: 230 - 239, (1991)
- <sup>132</sup> M.J. Sippl, S. Weitkus  
Detection Of Native Like Models For Amino Acid Sequences Of Unknown Three Dimensional Structure In A Database Of Known Protein Structures  
Proteins 13: 258 - 271, (1992)
- <sup>133</sup> D.T. Jones, W.R. Taylor, J.M. Thornton  
A New Approach To Protein Fold Recognition  
Nature 358: 86 - 89, (1992)
- <sup>134</sup> J.S. Fetrow, S.H. Bryant  
New Programs For Protein Tertiary Structure Prediction  
Biotechnology 11: 479 - 484, (1993)
- <sup>135</sup> D.E. Goldberg  
Genetic Algorithms In Search, Optimisation, And Machine Learning  
Addison - Wesley, Reading MA, (1989)
- <sup>136</sup> R. Unger, J. Moult  
Genetic Algorithms For Protein Folding Simulations  
Journal Molecular Biology 231: 75 - 81, (1993)
- <sup>137</sup> N. Metropolis, A.W. Rosenbuth, M.N. Rosenbuth, A.H. Teller, E. Teller  
Equation Of State Calculations By Fast Computing Machines  
Journal Chemical Physics 21: 1087 - 1092, (1953)
- <sup>138</sup> R. Calabretti, S. Nolfi, D. Parisi  
An Artificial Life Model For Predicting The Tertiary Structure Of Unknown Proteins That Emulate The Folding Process  
Lecture Notes in A.I. 929: 862 - 875, (1995)
- <sup>139</sup> T. Dandeker, P. Argos  
Folding The Main Chain Of Small Proteins With The Genetic Algorithm  
Journal Molecular Biology 236: 844 - 861, (1994)
- <sup>140</sup> D.J. Jones  
De - Novo Protein Design Using Pairwise Potentials And A Genetic Algorithm

---

Protein Science 3: 567 - 574, (1994)

<sup>141</sup> R.J. Read, G.D. Brayer L. Jurasek, M.N.G. James  
Critical Evaluation Of Comparative Model Building Of Streptomyces Griseus Trypsin  
Biochemistry 23: 6570 - 6575, (1984)

<sup>142</sup> F.C. Fernstein, T.F. Koeltz, G.J.B. Williams, E.F.R. Meyer, M.D. Brice  
The Protein Databank: A Computer Based Archival File For Macromolecules  
Journal Molecular Biology 112: 535 - 542, (1971)

<sup>143</sup> K. Fidelis, P.S. Stern, D. Bacon, J. Moulton  
Comparison Of Systematic Search And Database Methods For Constructing Segments Of  
Protein Structure  
Protein Engineering 7: 953 - 960, (1994)

<sup>144</sup> R.E. Bruccoleri, M. Karplus  
Prediction Of The Folding Of Short Polypeptide Segments By Uniform Conformational  
Sampling  
Biopolymers 26: 137 - 168 (1987)

