CANADIAN THESES

THÈSES CANADIENNES

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30.

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30.

THIS DISSERTATION HAS BEEN MICROFILMED EXACTLY AS RECEIVED

LA THÈSE A ÉTÉ MICROFILMÉE TELLE QUE NOUS L'AVONS REÇUE

NL-339 (r 86/06)

Canada

A Distance Constraint Model for the Prediction of Tertiary Structures of Globular Proteins

by
E.A.D. Foster
Department of Physiology and Biophysics

Submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

at

Dalhousie University

Halifax, Nova Scotia, Canada

November, 1986

ISBN 0-315-35403-8

Contents

Table of Contents. iv

List of Figures. vii

List of Tables. viii

Abstract. ix

List of Symbols. x

Acknowledgements. xii

1 A General Introduction to Protein Chemistry and Physics. 1
- 1.1 Introduction. . . . 1
- 1.2 Protein Physicochemistry and Structure. . . . 1
  - 1.2.1 Fibrous versus Globular Proteins. . . . 1
  - 1.2.2 Protein Manufacture. . . . 2
  - 1.2.3 Twenty Amino Acids. . . . 3
  - 1.2.4 Characteristics and Polarities of Amino Acids. . . . 6
  - 1.2.5 Peptide Bond Formation and Geometry. . . . 7
  - 1.2.6 Side Chain Chemistry. . . . 8
  - 1.2.7 Disulfide Bonds, Salt Bridges, Prosthetic Groups. . . . 9
  - 1.2.8 Hydrophobicity. . . . 10
  - 1.2.9 Secondary Structure. . . . 12
  - 1.2.10 Tertiary Structure. . . . 13
  - 1.2.11 Quaternary Structure. . . . 17
  - 1.2.12 Fluctuations. . . . 17

2 Previous Approaches to Tertiary Structure Prediction. 19
- 2.1 Energy Minimization Models. . . . 19
- 2.2 Secondary Structure Based Models. . . . 21
- 2.3 Distance Constraint Models. . . . 23
- 2.4 The Present Model. . . . 26

3 The Basic Parameters of Distance Constraint Models. 27
- 3.1 Near Neighbour Distances. . . . 27
- 3.2 Distances Between Far Neighbour Residues. . . . 28
- 3.3 Hydrophobicity Constraints. . . . 30
- 3.4 Chemically Derived Constraints. . . . 34

4 Suitably Constrained Systems. 37
- 4.1 Choice of Coordinate System. . . . 37
  - 4.1.1 Distance Geometry. . . . 40
- 4.2 The Mapping $\psi : Y \rightarrow X$. . . . 47
  - 4.2.1 Linear Criteria for a Suitably Constrained System. . . . 51
  - 4.2.2 Explicit Form of the Mapping $\psi : Y \rightarrow X$. . . . 54

4.2.3 Nonlinear Criteria for a Suitably Constrained System. 55
4.2.4 A Suitably Constrained Model. 57

5 The Mathematical Model **59**

6 The Optimization Algorithms. **61**

7 Results. **63**
7.1 Numerical Results for Structural Predictions of Three Globular Proteins. 65
7.2 Repeated Optimization of BPTI from Different Initial Configurations. . 88
7.3 Comparison to Previous Distance Constraint Models. 94
7.3.1 X-ray Diffraction Technique. 94
7.3.2 Goel, Yčas *et al.* 96
7.3.3 Kuntz, Crippen *et al.* 101
7.3.4 Wako and Scheraga. 106

8 Discussion. **110**
8.1 Improvements for the Present Model. 110
8.2 Alternative Algorithms for Solution. 112
8.3 Implications for Future Study. 114

9 Appendix: The Mathematical Model. **117**
9.1 Parameter Notation. 117
9.2 The Nonlinear Programming Formulation. 120
9.3 The Values of the Parameters. 124
9.3.1 Near Neighbour Parameters. 124
9.3.2 Far Neighbour Parameters. 127
9.3.3 Hydrophobicity Parameters. 129
9.3.4 Disulfide Bond Parameters. 133
9.3.5 Summary of Parameter Values. 134

10 Appendix: The Algorithms. **136**
10.1 Notation. 137
10.2 Newton's Method. 137
10.3 Steepest Descent Method. 138
10.4 Truncated-Newton Method. 140
10.5 Conjugate Gradient Method. 141
10.6 Solving the Nonlinear Programming Problem. 143
10.7 Outer Loop Algorithm. 144
10.8 Inner Loop Algorithm. 145
10.9 Compact Storage of the Hessian. 145

11 Appendix: On Theoretical Near Neighbour Distance Parameters. **151**
11.1 Ramachandran Angle and Virtual Bond Descriptions of a Polypeptide. . 152
11.2 Theoretical Calculation of Near Neighbour Distances as Functions of Virtual Bond Angles. 156
11.2.1 Finding $d_{i,i+2} = f(\theta)$. 156
11.2.2 Finding $d_{i,i+3} = f(\gamma)$. 157

11.2.3 Finding $d_{i,i+4} = f(d_{i,i+j})$. 160
11.3 Theoretical Calculation of Near Neighbour Distances as Functions of the Ramachandran Angles. 162
11.4 Numerical Results: Theoretical Near Neighbour Distances. 168
11.4.1 Minimum and Maximum Distance Parameters. 169
11.4.2 Mean Value Parameters: Secondary Structure Distances. 177

References. 182

List of Figures

1 General Structure of an Amino Acid. 3
2 Amino Acid Side Chains (from Schulz and Schirmer 1979). 5
3 The Peptide Bond (from Creighton 1983). 8
4 Tertiary Structure of BPTI Molecule (from Scheraga 1983). 15
5 Tertiary Structure of Lysozyme Molecule (from Stryer 1981). 16
6 The Tetrangle Inequality for a System of Four Points. 44
7 Contact Maps for BPTI, Including Disulfide Bond Constraints. 72
8 Contact Maps for BPTI, Excluding Disulfide Bond Constraints. 73
9 Distances of the Residues from the Centroidal Point for Optimized Structure of BPTI. 76
10 Contact Maps for Rubredoxin. 78
11 Contact Maps for Lysozyme, the Four Disulfide Bonds Included. I. The Real Structure. 82
12 Contact Maps for Lysozyme, the Four Disulfide Bonds Included. II. Initial Configuration. 83
13 Contact Maps for Lysozyme, the Four Disulfide Bonds Included. III. First Outer Loop. 84
14 Contact Maps for Lysozyme, the Four Disulfide Bonds Included. IV. Optimized Structure. 85
15 Contact Maps for BPTI. Optimized Structures from Dissimilar Initial Configurations. 93
16 Flowchart of the Inner Loop Algorithm. 146
17 The Nonzero Hessian Matrix Elements. 147
18 Required Hessian Matrix Elements, Symmetry Included. 148
19 Definition of Ramachandran Bond Angles for a Polypeptide Chain (from Schulz and Schirmer 1979). 153
20 Definition of Virtual Bond Angles for a Dipeptide (from Nishikawa *et al.* 1974). 155
21 Second Neighbour Distance: Virtual Bond Residues. 157
22 Third Neighbour Distance: Virtual Bond Residues. 158
23 Fourth Neighbour Distance: Virtual Bond Residues. 161
24 Theoretical Distributions of $d_{i,i+2}$ Distances. 171
25 Theoretical Distributions of $d_{i,i+3}$ Distances. 172
26 Theoretical Distributions of $d_{i,i+4}$ Distances. 173

List of Tables

1 Notations and Hydrophobicity Classifications for the 20 Common Amino Acid Residues of Proteins. . . . 4
2 Hydrophobicity Classification with Respect to the Centroidal Point. . . 31
3 Empirical Hydrophobicity Classification of Residues (from Goel and Yčas 1979). . . . 34
4 Numerical Results: Rubredoxin. . . . 68
5 Numerical Results: BPTI, No Disulfide Bond Constraints. . . . 68
6 Numerical Results: BPTI, Including Disulfide Bond Constraints. . . . 69
7 Numerical Results: Lysozyme. . . . 69
8 Numerical Results for BPTI: Near Neighbour and Centroidal Point Distance Statistics. . . . 75
9 Numerical Results for Rubredoxin: Near Neighbour and Centroidal Point Distance Statistics. . . . 79
10 Distances Between the Cys Residues in Rubredoxin (Real Structure). . . 80
11 Distances Between the Cys Residues in Rubredoxin (Optimized Structure). . . . 81
12 Numerical Results for Lysozyme: Near Neighbour and Centroidal Point Distance Statistics. . . . 86
13 Numerical Results: BPTI, from Initial Configuration A. . . . 89
14 Numerical Results: BPTI, from Initial Configuration B. . . . 90
15 Numerical Results: BPTI, from Initial Configuration C. . . . 90
16 Numerical Results: BPTI, from Initial Configuration D. . . . 90
17 Numerical Results: BPTI, from Initial Configuration E. . . . 91
18 Comparison of Optimized RMS_y Structures for BPTI. . . . 91
19 Comparison of Optimized $RMS_y(4)$ Structures for BPTI. . . . 91
20 Further Comparison of RMS_y and $RMS_y(4)$ Final Structures for BPTI. 92
21 Near Neighbour Parameters for Distance Constraint Models. . . . 125
22 Far Neighbour Parameters for Distance Constraint Models. . . . 128
23 Regression Statistics for Hydrophobicity Classes of the Amino Acids (from Goel and Yčas 1979). . . . 130
24 Hydrophobicity Parameters for the Distance Constraint Model. . . . 131
25 Disulfide Bond Parameters for Distance Constraint Models. . . . 134
26 Parameter Values for the Present Model. . . . 135
27 Standard Values for the Peptide Bond Angles. . . . 163
28 Theoretical Near Neighbour Distance Statistics. . . . 169
29 Differences in the Near Neighbour Distance Bounds: (ψ, ϕ) Angles Unrestricted. . . . 170
30 Limits on the Ramachandran Angles (ψ, ϕ). . . . 175
31 Theoretical Near Neighbour Distance Statistics for the Standard Limits on (ψ, ϕ) Angle Sets. . . . 175
32 Ramachandran Angle Sets Corresponding to the Theoretical Near Neighbour Distance Maxima. . . . 176
33 Secondary Structure Examples: BPTI. . . . 180
34 Theoretical Near Neighbour Distances: Secondary Structures. . . . 180

Abstract.

Requiring only the one-dimensional primary structure as input, the positions of the constituent residues of a globular protein are predicted in three-dimensional space by a model using current mathematical programming techniques.

Semi-empirically derived parameters in the form of distances between points are utilized. The residues are positioned by minimization of a simple distance function of their hydrophobicity classes, given constraints on their near neighbour distances and bounds on their far neighbour distances. Disulfide bonding information or extra-primary substructures may also be used, where appropriate The objective function and constraints are combined into a nonlinear penalty function, which is minimized by a new low-storage optimization technique. The optimization method employs a combination of steepest descent and a truncated-Newton method.

The model is designed to be suitably constrained, in that the predicted structures are not overly dependent upon initial conditions and the solution space is small with respect to both Cartesian coordinate and distance coordinate space. The model is capable of predicting tertiary structures for all single strand globular proteins, with no restriction on length.

The tertiary structures calculated are found to have global structures similar to those found by experimental crystal X-ray diffraction techniques. Using the distance space root-mean-square (RMS_y) as a measure, RMS_y differences from the diffraction structures are found of 4.88 Å, 4.45 ± 0.43 Å and 5.75 Å for rubredoxin (54 residues), BPTI (58 residues) and lysozyme (129 residues), respectively.

List of Symbols.

Å — Ångström unit of length, equal to 10^{-10} meter.

R^n — Euclidean n-dimensional space.

C_α-atom — The central carbon atom of an amino acid residue.

C_β-atom — The sidechain carbon atom nearest to the C_α-atom in a residue.

C_i^α — The C_α-atom of the ith residue in the primary sequence of a protein.

$d_{i,i+j}$ — The Euclidean distance between the C_α-atom locations of residues i and $i+j$.

d_{ij} — The Euclidean distance between the C_α-atom locations of specific residues with primary sequence positions of i and j.

$d_{i,cp}$ — The distance of residue i from the centroidal point of the protein. The centroidal point is defined as the average Cartesian coordinate location of the C_α-atoms of the residues.

J_1 — The set of residue types classified as hydrophobic, tending toward the centroidal point of the protein.

J_2 — The set of residue types classified as hydrophilic, tending toward the outer surface of the protein.

J_3 — The set of residue types classified as ambivalent, having no tendency with respect to the centroidal point of the protein.

$\overline{d}_i$ — The mean distance between residues separated in primary sequence by i residues.

$\overline{d}_S$ — The mean distance between residue pairs joined by disulfide bonds.

L_j — Minimum distance of approach between residues separated by j residues in primary sequence.

U_j — Maximum distance between residues separated by j residues in primary sequence.

L_N — An absolute minimum distance between residues that are widely separated in primary sequence.

U_N — An absolute maximum distance between residues that are widely separated in primary sequence.

D — A parameter representing an ideal distance from the centroidal point for hydrophilic residues.

$\nabla f(x)$ — The gradient vector of partial derivatives of a function f evaluated at a point x.

$H(x)$ — The Hessian matrix of second order mixed partial derivatives of f evaluated at x.

T_σ^z — A clockwise rotation of an angle σ about the z-axis of the orthogonal Cartesian coordinate frame (x,y,z).

(ψ, ϕ) — Dihedral Ramachandran angle description of polypeptide chain.

(θ, γ) — Virtual bond angle description of a polypeptide chain.

Y-space — Euclidean space using the distances between points as the primary coordinate system.

X-space — Euclidean space using a set of Cartesian coordinates as the primary coordinate system.

RMS_y — Root-mean-square measure with respect to distance coordinates.

RMS_x — Root-mean-square measure with respect to orthogonal Cartesian coordinates.

Acknowledgements.

I would like to express gratitude to my supervisor, Dr. Robert Rosen, for his support throughout my degree program. I value highly the biomathematical discussions we have had at the old Red House and the new biophysics annex.

Also, I wish to thank Dr. Phillip O'Neill for his interest and enthusiasm for the project, and for his time and effort spent in the joint development of the optimization algorithms.

I would like to extend my appreciation to my colleagues, my fellow graduate students, for their support and helpful discussions. In particular, I would like to thank Janet Gregory, Aloisius Louie and Mikhail Youssef in this regard

I wish to express heartfelt thanks to Kate Fletcher and Evelyn Nugent for their encouragement and their critical reading of the manuscript.

I want to especially thank my wife, Judy, for her encouragement, great assistance and unflagging support throughout the lengthy process

1 A General Introduction to Protein Chemistry and Physics.

1.1 Introduction.

A central concern in molecular biophysics is the study of the three-dimensional conformations of proteins. Proteins are crucial in virtually all biological processes The elucidation of the three-dimensional structure of proteins aids in our understanding of these processes, because a protein's function is determined entirely by its structure.

This problem is simple in principle. It is believed that the primary structure of a protein, the one-dimensional sequence of its constituent amino acid residues, uniquely determines its tertiary structure, namely the locations of its atoms in R^3 [3,4,48,54]. That is, the protein will fold spontaneously into a unique stable three-dimensional structure in a suitable environment, without the necessity of an additional energy or information input. The problem then, is to define an algorithm that produces the tertiary structure from a given primary structure. Once tertiary structures can be accurately predicted, the causal relationship of structure to function can be properly addressed.

The determination of the native three-dimensional conformation of proteins is a major unresolved problem in molecular biology. All theoretical approaches to this problem are based on the accepted dogma that the tertiary structure of a protein is the direct result of its primary structure in the native environment.

1.2 Protein Physicochemistry and Structure.

1.2.1 Fibrous versus Globular Proteins.

Proteins can be classified into two groups according to their macrostructure. Fibrous proteins are those associated with structural elements in the cell, and are largely

insoluble in an aqueous environment. They have high molecular weights and are capable of stretching and contracting. In general, their overall conformations are either long fibers or sheets.

In contrast to fibrous proteins, the globular proteins are generally soluble in water, smaller and less symmetrical. Nearly all enzymes are globular proteins. Other globular proteins perform a remarkably diverse range of functions, acting as antibodies, hormones and receptors, growth and differentiation controllers, and ion and molecule transporters. In this study, only globular proteins will be examined.

For detailed accounts of protein chemistry and structure, the reader is referred to Dickerson and Geis [34], Schulz and Schirmer [101], or Creighton [28].

1.2.2 Protein Manufacture.

According to currently held ideas on protein synthesis [122], the amino acid sequence in a polypeptide chain of a protein is a colinear and unique representation of the nucleotide sequence of the nucleic acid which codes it. Three adjacent nucleotides constitute a codon, and specify a single corresponding amino acid. Accordingly, the polypeptides are similar to nucleic acids in that they are *linear*, unbranched chain molecules with standard elements and one standard linkage. This arrangement allows for a simple and universal nucleic acid reading and polypeptide synthesizing mechanism.

In proteins, as in nucleic acids, not only the linkages but also the atomic groups forming the backbone of the chain are uniform; in polypeptides all 20 common amino acids are of the α-type and have the L-configuration at their central C_α-atoms. All differences, and therefore all information, are restricted to the rather short sidechains of the amino acids.

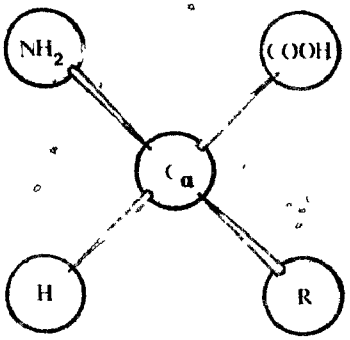


Figure 1: General Structure of an Amino Acid.

1.2.3 Twenty Amino Acids.

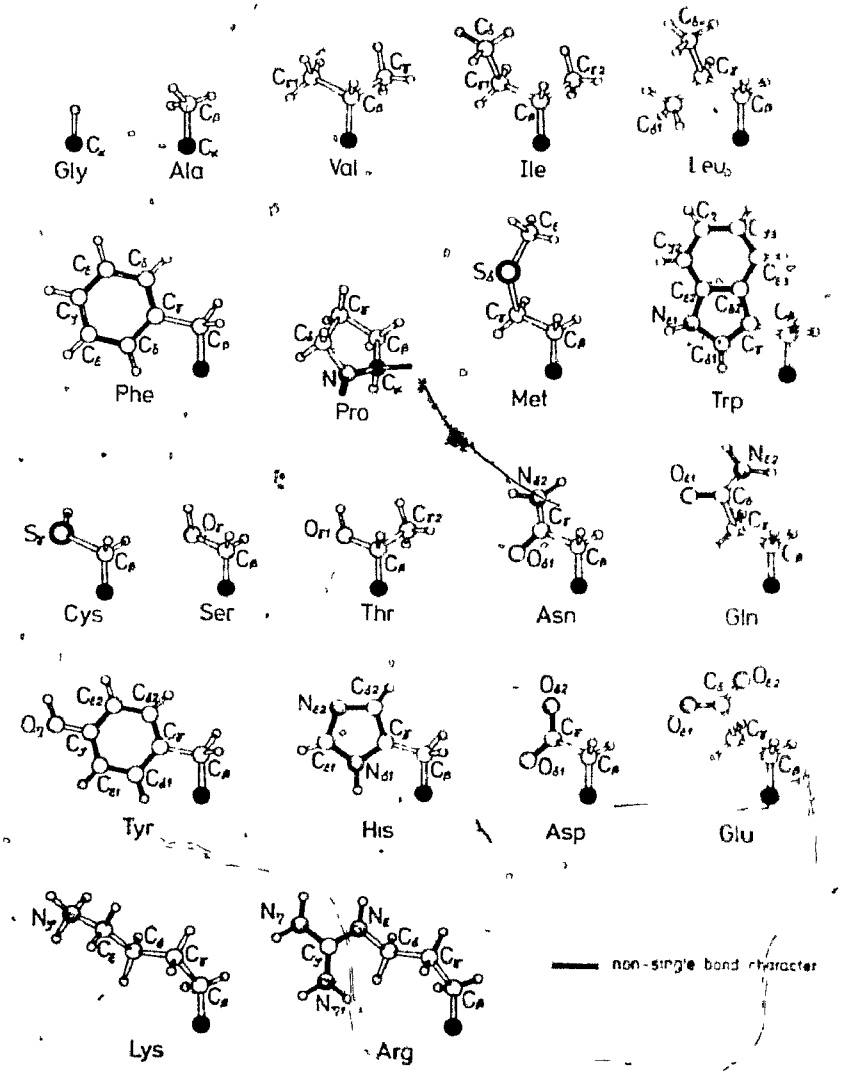
There are 20 standard amino acids residues occurring in natural proteins. These are listed in Table 1 along with their hydrophobicity classifications from various sources. The displayed hydrophobicity classifications will be discussed in Chapter 3.3 Table 1 also gives the commonly used three-letter and one-letter abbreviations for the amino acids. The three-letter abbreviations will be utilized throughout this thesis. The molecular weights of the amino acid residues range from 57 to 186 daltons with the mean, weighted by relative abundance, being about 110.

The general structure of an amino acid is shown in Figure 1. All amino acids, with the exception of proline, have an amino ($-NH_2$) group, a carboxyl ($-COOH$) group, a hydrogen atom, and a distinguishing R group (called the sidechain), all bonded to a central (C_α) carbon atom. The amino acids can be classified with respect to their sidechains as either polar or nonpolar. The polar sidechains can be further subdivided into neutral, basic, or acidic. Figure 2 illustrates the 20 common amino acid sidechains. In this figure, amino acids with similar properties are grouped near one another.

Amino Acid or Residue	Three-letter Abbreviation	One-letter Symbol	Ref # 1	Ref # 2	Ref # 3	Ref # 4	Ref # 5	Ref # 6
Alanine	Ala	A	a	a	a	a	l	a
Arginine	Arg	R	l	l	a	l	a	l
Asparagine	Asn	N	a	a	l	l	l	l
Aspartic Acid	Asp	D	l	l	l	l	l	l
Cysteine	Cys	C	a	a	b	b	l	b
Glutamine	Gln	Q	l	a	l	a	l	l
Glutamic Acid	Glu	E	l	l	l	l	l	l
Glycine	Gly	G	l	a	l	a	a	a
Histidine	His	H	a	l	b	l	l	a
Isoleucine	Ile	I	b	b	b	b	b	b
Leucine	Leu	L	b	b	b	b	b	b
Lysine	Lys	K	l	l	l	l	a	l
Methionine	Met	M	b	b	b	b	b	b
Phenylalanine	Phe	F	b	b	b	b	b	b
Proline	Pro	P	l	b	l	a	b	l
Serine	Ser	S	a	a	a	l	a	l
Threonine	Thr	T	a	a	l	a	l	l
Tryptophan	Trp	W	a	b	b	a	b	b
Tyrosine	Tyr	Y	a	l	a	l	b	a
Valine	Val	V	b	b	b	b	a	b

Table shows hydrophobicity classifications by various authors for the twenty amino acids commonly found in natural proteins. The three hydrophobicity classes used are b = hydrophobic, l = hydrophilic, and a = ambivalent. The sources are Goel and Yčas [46] (Ref #1), Dickerson and Geis [34] (Ref #2), Wertz and Scheraga [124] (Ref #3), Charton and Charton [16,17] (Ref #4), Lawson *et al.* [64] and Jones [55] (Ref #5), and Meirovitch *et al.* [74,75,76] (Ref #6).

Table 1: Notations and Hydrophobicity Classifications for the 20 Common Amino Acid Residues of Proteins.



Shown are the sidechains for the 20 common amino acids. For proline, part of the main chain is inserted. The other sidechains are shown as they emerge from the C_α-atom of the residue.

Figure 2: Amino Acid Side Chains (from Schulz and Schirmer 1979).

1.2.4 Characteristics and Polarities of Amino Acids.

The standard amino acids differ only with respect to their sidechains. Each sidechain is so specific that it cannot be easily substituted with another one without altering the gross properties of the protein. Glycine has only a hydrogen as its sidechain. With no sidechain hindrance, Gly residues can adopt unusual dihedral angles, giving rise to kinks in the main chain. Therefore, the presence of these amino acids will increase the flexibility of the polypeptide chain. Gly and Ala are so small that they can apparently be accommodated in the interior of a protein or on its surface with equal ease. The nonpolar sidechains of Val, Ile and Leu are branched. Branched sidechains are stiffer, making them easier to fix in specific positions. Met has a rather flexible sidechain containing one sulfur atom. The nonpolar amino acids are predominantly found on the inside of protein molecules. Pro, the only amino acid in which the sidechain reattaches itself to the main chain, has the unique property of disrupting an α-helix and forcing a bend in the main chain.

The aromatic amino acids Phe, Trp and Tyr all contain one methylene group that acts as a spacer between the C_α-atom and the aromatic ring. Without this group the main chain would be extremely stiff due to the steric hindrance at the C_α-atom.

Typical polar and neutral sidechains are those of Cys, Ser, Thr, Asn, Gln and Tyr. They tend to form hydrogen bonds and to be found on the outside of the molecule. Most of the active centers of enzymes contain His amino acids. Asp and Glu are negatively charged amino acids at physiological pH and are both found at protein surfaces. Positively charged Lys and Arg residues also tend to be found at the surface.

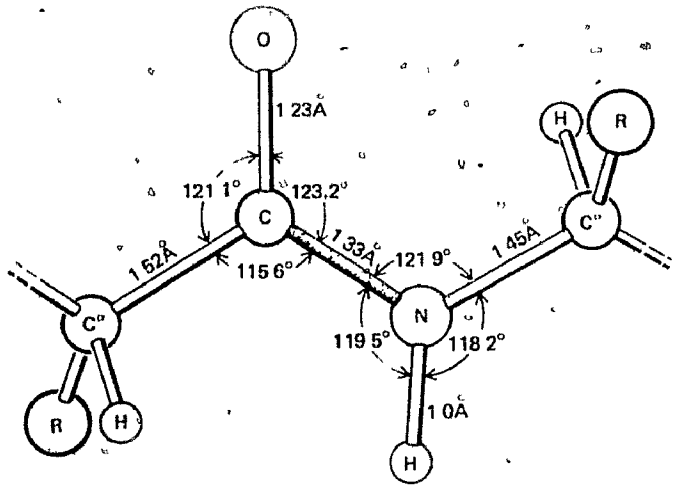
Cys can help to stabilize protein structures because of the ability of two such amino

acids to combine to form a disulfide bond within a protein. These disulfide bridges are the only common covalent cross-links in proteins.

1.2.5 Peptide Bond Formation and Geometry.

Amino acids are polymerized into a polypeptide chain on ribosomes in the cell. The polymerization is based on the formation of substituted amide bonds, usually called "peptide" bonds. Once polymerized, the individual amino acids are referred to as "residues". Typically, a single chain of a protein will contain 50–1000 residues. Globular proteins are composed of one or more residue chains, or *strands*. The chain direction is defined as pointing from the amino end (N-terminus) to the carboxyl end (C-terminus), coinciding with the direction of chain synthesis *in vivo*.

The geometry and the dimensions of the peptide bond are shown in Figure 3. These data have been derived by Marsh and Donohue [73] and Ramachandran *et al.* [89] as a refinement of the pioneering studies of Pauling *et al.* [25,81], using crystal structures of small polypeptides. The peptide linkages are predominantly *trans* (Figure 3) so that the hydrogen of the $(-NH)$ group is as far as possible away from the oxygen of the $(-C=O)$ group; the alternative *cis* peptides, wherein the $(-NH)$ group hydrogen is as close as possible to the $(-C=O)$ oxygen, occur rarely. Rotation around the peptide bond is inhibited by resonance, causing a partial double bond of $(O=C-N)$. This makes the peptide bond essentially planar. The rotational freedom of the backbone is thus localized in the two single bonds $(C_\alpha - N)$ and $(C_\alpha - C)$. With a stiff peptide bond and with rather rigid bond lengths and bond angles, the distance between the C_α-atoms of two adjacent residues in the chain is found to be essentially constant, equal to 3.80 Å.



Standard angles and distances for the usual *trans* peptide bond as given by Ramachandran *et al.* (1974).

Figure 3: The Peptide Bond (from Creighton 1983).

1.2.6 Side Chain Chemistry.

The local chemistry within the protein is affected by the peptide bond, which is stiff (restricting the chain flexibility) and rather bulky (giving rise to substantial steric hindrance). Except for Gly and Pro, the sterically allowed regions for all residues are essentially the same and rather small [90,91]. This causes residues that are close together in primary sequence (the "near neighbour" residues) to have strict and specific limitations on their maximum and minimum pairwise distances in R^3.

The packing density of a molecule is defined as the ratio of the van der Waals radii of its atoms to the volume it actually occupies in space [94]. Because globular proteins possess a high packing density (close to the density of crystals of small molecules that are held together by van der Waals forces), the final structure of a protein is very dependent upon noncovalent forces. The noncovalent forces drive the spontaneous folding process, and later act as the mediators of enzyme-substrate reaction mechanisms

or other biological activities. Noncovalent forces in the protein include dispersion forces and electrostatic interactions between partially charged residue sidechains (van der Waals forces), hydrogen bonding between two residues or between a single residue and a water molecule, and hydrophobic forces from the nonpolar residues. Polar residues in the interior of the protein help stabilize the protein by the formation of numerous hydrogen bonds.

1.2.7 Disulfide Bonds, Salt Bridges, Prosthetic Groups.

Salt bridges are weak ionic interactions between oppositely charged sidechain groups. There are only a few salt bridges in proteins. They are usually located on the exterior of the protein, although interior salt bridges would be much more useful in stabilizing the structure of a protein.

Disulfide bridges can be formed between pairs of Cys residues. These covalent bonds can serve to cross-link different parts of a protein chain. As a rule, these bonds form spontaneously.

Historically, it was thought that disulfide bonds determine the three-dimensional structure of the protein. However, it was found that the disulfide bonds in most proteins can be fully reduced and denatured, and the denatured protein will refold into its native structure with correct disulfides upon reoxidation (*cf.,* Haber and Anfinsen [49]). It was also determined from denaturing-renaturing experiments that some S-S bonds are transient during the folding process of a protein, and the disulfide pairings of the final folded structure may actually be formed after the secondary and the tertiary structure of the protein has been achieved [27]. Furthermore, a great many S-S links of proteins can be broken without the loss of the protein's structure or function. For example,

all three disulfide bonds of α-amylase can be reduced without impairing its enzymatic activity [105].

Apparent exceptions to the hypothesis of spontaneous S-S bond formation occur with insulin and α-chymotrypsin, which cannot be renatured once their disulfide bonds are broken [44]. However, both of these proteins are formed from larger precursor molecules by proteolytic cleavage, and both of their precursors (proinsulin and chymotrypsinogen) reform their native structures and S-S bond pairings upon reduction and reoxidation. This indicates that insulin and α-chymotrypsin require the energy contributions from their native set of S-S bonds for stability. The common function of disulfide bonds then, is not to determine the three-dimensional structure but to give extra stability to otherwise properly folded proteins [4,101]. These bonds are consequences of folding, and not the driving forces.

Prosthetic groups, although often noncovalently bound to the polypeptide chain, may in some cases be linked to the sidechains of protein residues. It is possible in many instances to remove the prosthetic group without damage (*e.g.*, the heme in globins), whereas in other cases the protein becomes denatured (*e.g.*, the heme in catalase).

1.2.8 Hydrophobicity.

One of the principal driving forces of protein folding results from the energetically unfavorable interactions between nonpolar sidechains and water. This "hydrophobicity" causes the majority of nonpolar residues in native proteins to cluster inside the molecule, away from the aqueous environment, forming a tightly packed solvent-inaccessible hydrophobic core. As first observed by Danielli [31] and Kauzmann [58], this hydrophobicity is a major factor in determining the three-dimensional shape of proteins, and hence their

activity. Nonpolar sidechains that do remain at the surface are frequently found to be oriented so that their contact with water is minimized.

Since globular proteins have diameters of about 30 angstroms, sidechains cannot be buried in the protein interior without also burying part of the backbone, the polar amide and carbonyl groups. However, the polar groups coexist well with water. Burying them in the interior without loss of free energy is achieved only by the formation of hydrogen bonds. Regular hydrogen bonding patterns are commonly observed among residues in the interior of a protein; these give rise to what are termed "secondary structures". It has been found that the fraction of buried nonpolar groups increases with a protein's size, whereas the fraction of buried polar groups remains relatively constant [18].

The removal of charged groups from water is energetically very unfavorable. The vast majority of charged sidechains are at the protein surface. These types of residues are referred to as "hydrophilic".

The hydrophobicity rule of "nonpolar in, charged groups out" helps to stabilize a protein in aqueous solution, giving proteins their globular shape. The arrangement of the internal sidechains is remarkably efficient. If the internal volume is compared to the sum of the volumes of the constituent sidechains, the interior of the protein is found to be packed at about the same density as solid crystalline amino acids [94].

Hydrophobicity indices or classifications have been proposed by several authors, and several of these are listed in Table 1. These consist of rating or classifying the residue types on their preference to be situated within or away from the aqueous environment. They are discussed in detail in Chapter 3.3.

1.2.9 Secondary Structure.

As suggested by the terms *primary* structure and *tertiary* structure, there exists a hierarchy of identifiable geometrical structures in proteins. This hierarchy will be described in the following three sections.

The secondary structure of a protein, intermediate between primary and tertiary structure, can be defined as the arrangement of its main chain atoms without regard to the types or conformation of its sidechains or its relationship with other chain segments. Secondary structures are stabilized by hydrogen bonds between the peptide amide and carbonyl groups. Four types of secondary structures are commonly found in globular proteins:

1. The α-helix [83], a regularly repeating structure containing 3.6 residues per helical turn, resulting in small uniform distances between near neighbour residues;
2. The β-strand [82], a helical structure such that the polypeptide chain is nearly fully extended, resulting in large uniform distances between near neighbour residues;
3. The 3_{10}-helix [35], an intermediate helical structure, occurring less frequently in proteins, which contains 3.0 residues per turn;
4. Reverse, or hairpin, turns [115], sharp turns containing four residues and usually stabilized by a single hydrogen bond.

The α-helix is the most abundant secondary structure in proteins, effecting rather stable rods through the interiors of globular proteins. The stability of normal helical conformations is affected by both the polypeptide length and the residue sequence. Gly and Pro have the characteristic of destabilizing any hydrogen bonding pattern and

are termed helix disrupters, albeit for different reasons. A Pro residue does not have a hydrogen atom on its peptide nitrogen atom and, therefore, is unable to contribute to the hydrogen bonding patterns of a helix. With only a hydrogen atom for its sidechain, a Gly residue can destabilize a helix due to its extensive flexibility. Residues are theoretically capable of forming helices of types other than the ones listed above. However, none has been found with any significant frequency in real proteins.

β-strand arrangements differ from other regular helical structures because they involve hydrogen bonding between sequentially distant residues. β-strand systems are observed of two types: parallel and antiparallel, in which adjacent β-strands run in the same or in opposite directions, respectively. In these, the stabilizing hydrogen bonding pattern occurs between residues of opposing β-strands These arrangements are termed β-pleated sheets.

Reverse turns are usually located on the surface of a protein. They are quite flexible, and susceptible to changes in environment. They generally have distinctly recognizable conformations and are often restricted in their residue composition [5], with Gly being the major participant. About one quarter of all protein residues are involved in turns.

1.2.10 Tertiary Structure.

Tertiary structure refers to the three-dimensional conformation of the atoms in a single strand of a protein. The function of a protein is dependent upon its tertiary structure. Tertiary structure is stabilized not only by the covalent bonding of the atoms in the main chain and disulfide bonds, but also by essential noncovalent forces such as hydrogen bonding, van der Waals forces and hydrophobic interactions.

After synthesis on the ribosome, general physical principles imply that the polypep-

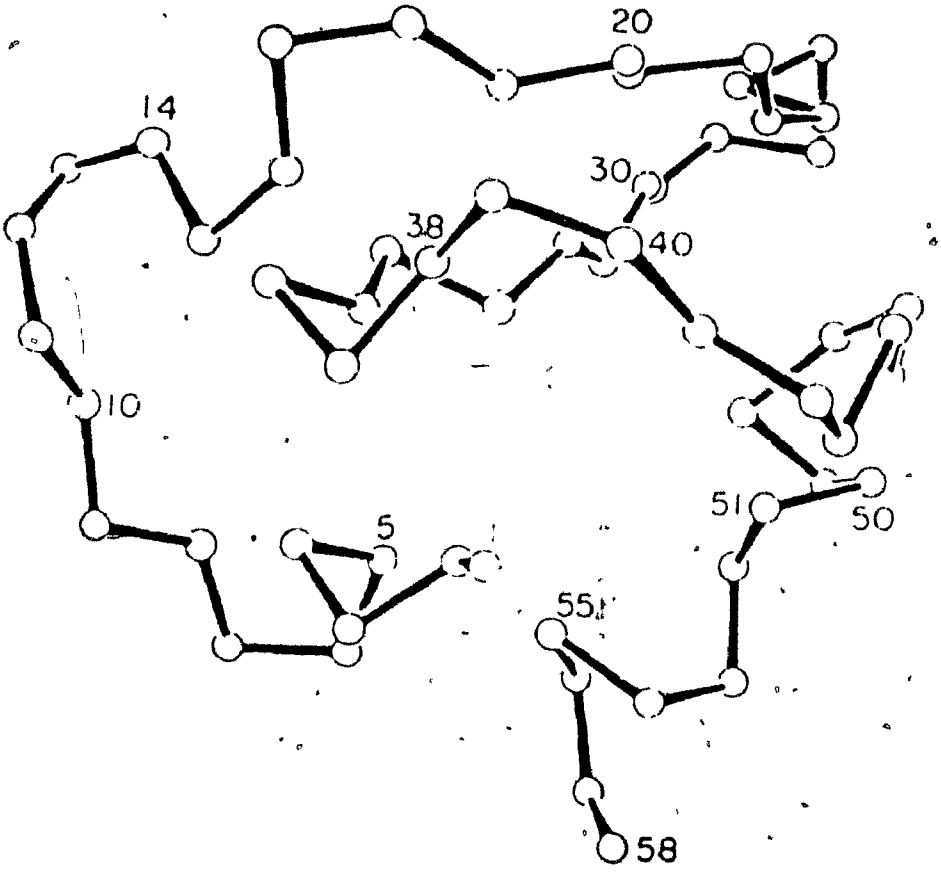
tide chain will fold spontaneously into a primary sequence-dependent globular protein by adopting a state of minimal free energy. It is probable that folding already starts during the synthesis. The resulting folded structure determines the biological activity of the protein.

Experimental evidence supports the hypothesis that, under native conditions, a protein will fold into a three-dimensional structure that is unique [4,48,54]. This complicated structure (*cf.*, Figures 4 and 5) is dictated only by the amino acid sequence and the chemical environment. However, the relationship between sequence and structure is highly degenerate [97]. That is, many primary sequences can give similar folded structures and biological activity [113].

There appear to be only a limited number of amino acid sequences that can provide a unique structure in a given environment [97]. Artificially constructed random polypeptide strings tend not to have unique configurations, but instead behave as random coils that continually shift from one structure to another [15,91]. This implies the natural proteins may be a small subset of polypeptides, selected partly for their stability of structure [36].

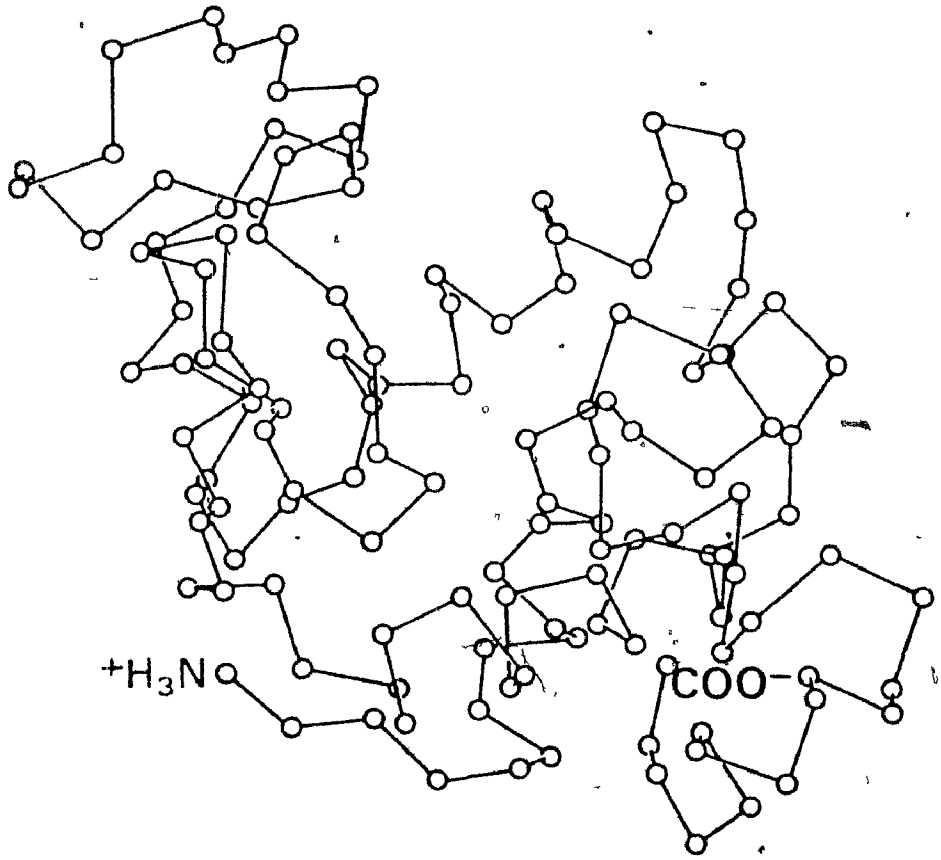
The tertiary structures of natural proteins are not only unique, but they are also specific in a given environment. Only a few specific residue substitutions are possible that will allow the molecule to retain some activity.

As will be discussed in Chapter 2.2, the knowledge of the secondary structures of a protein unfortunately does not greatly aid the elucidation of tertiary structure. The tertiary structure of a protein thus cannot adequately be described as a simple aggregate of connected secondary structures.



The residues are numbered sequentially from the amino end to the carboxyl end. Only the central C_α-atoms of the residues are shown, connected by virtual bonds.

Figure 4. Tertiary Structure of BPTI Molecule (from Scheraga 1983).



Only the C_α-atoms of the polypeptide backbone are shown, connected by virtual bonds.

Figure 5: Tertiary Structure of Lysozyme Molecule (from Stryer 1981).

1.2.11 Quaternary Structure.

Many proteins exist as large molecules formed by the specific aggregation of several identical or non-identical protein strands. These strands may be held together by disulfide bonds, hydrogen bonds or hydrophobic bonds. The quaternary structure deals with the arrangement of the constituent strands. For example, hemoglobin is a protein with a quaternary structure consisting of two pairs of single-chain subunits. Each of these four strands is folded into a shape similar to a myoglobin molecule, which is a single strand protein. The four separate strands are consolidated into a single stable structure by a great many hydrophobic interactions, along with a few hydrogen bonds and charged-group interactions. Proteins that have molecular weights in excess of 50,000 are likely to involve two or more polypeptide chains.

1.2.12 Fluctuations.

A protein molecule is not a static structure. Protein atoms are in a constant state of motion. The fluctuations of the atoms are possible from bond rotations that are available in the polypeptide backbone and sidechains, and separated by only small energy barriers.

These high-frequency fluctuations may be a factor in the functions of proteins, such as enzyme catalysis [57,123]. However, the scale of the fluctuations is only of the order of 0.5 Å for an individual atom, with larger movements being prohibited by the tight packing of the molecule. This means that the fluctuation magnitude is smaller than the resolution of the present theoretical models, as well as being beyond the resolving power of current experimental methods for tertiary structure generation. Therefore,

fluctuations will not be considered further in this thesis.

2 Previous Approaches to Tertiary Structure Prediction.

This chapter is concerned with reviewing the methods that are currently being employed for theoretical prediction of the tertiary structures of globular proteins. The starting point for all of these methods is the hypothesis that all the information needed to define the three-dimensional structure of a protein is inherent in its amino acid sequence.

2.1 Energy Minimization Models.

The existence of a unique stable conformation under native conditions indicates that the Gibbs free energy of the system consisting of the protein and the surrounding solvent must correspond to a minimum state. Therefore, the most straightforward approach for prediction of tertiary structure is to write the equations that describe the free-energy state of the molecule in its solvent, and solve for a global minimum. In principle, this method should always be successful, and the structures of simple chemical systems have been determined by such direct energy minimization. However, the numerous attempts to determine protein structures by this method have met with limited success in practice, mainly due to the enormous size of these molecules (*cf.*, Nemethy and Scheraga [78]).

The set of equations describing even a moderate-sized protein is so large that the equations would be virtually impossible to write down, let alone solve. The computation of the potential energy of a conformation is based on the assumption that the energy can be expressed as a sum of interactions between the atoms of the protein, with individual interaction terms required for all possible pair combinations of the atoms.

This amounts to a computation in the order of n^2 nonlinear distances and energy terms for every evaluation of the energy function for a molecule containing n atoms; for example, over a million terms would be calculated for the small enzyme lysozyme at each iteration.

The equations are nonlinear and the free-energy surface is characterized by an extremely large number of local minima. This "multiple-minima" problem is considered a major obstacle in the solution of free-energy equations for any large chemical system [100]. Mathematical procedures developed to solve the multiple-minima problem by passing from one potential well to another consume far too much computer time, and it is not practical to apply them to a polypeptide longer than a few residues [29]. Currently, energy minimization models are being developed that attempt to circumvent the multiple-minima problem; for example, the real-space renormalization group technique familiar to the world of theoretical physics can be used [100].

A further complication for this method is that it is unclear whether a protein's free-energy state *in vivo* corresponds to a global minimum or to a strong local minimum. If the native conformation does not correspond to a global minimum, this would mean that certain conformations are energetically inaccessible because of high potential energy barriers and that there is a limited number of possible pathways along which a protein can fold. This has led researchers away from the exploration of the entire conformational space in search of a free-energy minimum, and instead to consider only directed pathways of folding, leading through intermediate conformations involving stable near neighbour and secondary structure arrangements. In effect, this amounts to the consideration of a problem that is different from the energy minimization of the true protein system, but much more amenable to solution.

The complexities of direct free-energy minimization have stimulated a search for alternate solution techniques. One method is to grossly simplify, modify or approximate the free-energy function in order to make it tractable [67,68,107,108,109,110,111,112,129]. This often involves treating each residue as a single point instead of as a group of atoms, or using information such as homologies between proteins to choose initial conformations that are already close to the real structure. Also, empirical or simplified functions are used in place of the original free-energy equations. These simplification approaches have also proved difficult. Either the equations remain highly nonlinear and unwieldy, with a multitude of local minima, or the original functions become so distorted that they are barely recognizable.

It appears unlikely that realistic free-energy equations can be solved in any real sense in the near future. Therefore, one must search for other means of tackling this problem.

2.2 Secondary Structure Based Models.

One alternative technique is to first predict the secondary structures of a protein by statistical methods such as those of Chou and Fasman [19,20], and then to compose the structural elements into a suitable globular structure [21,37,38,85,104].

Most approaches to protein structure prediction have concentrated on secondary structure, since most protein residues participate in some type of α-helix, β-strand or hairpin turn. The object of these approaches is to first accurately predict secondary structure elements and then learn how to pack them together to generate the correct tertiary structure. The empirical tendencies of the amino acids to form various secondary structures have been studied on an individual basis [66] for conformational

correlations, with marginally significant results. A more successful empirical scheme to predict the occurrence of secondary structures is based upon first ranking the amino acids with respect to forming, breaking, or being ambivalent to each type of secondary conformation. Then the relative positions of near neighbour residues are observed with respect to the primary sequence of a protein, and the prediction proceeds by an elaboration on the number and kind of residues required to nucleate and terminate a given structural element. The best known of these methods is the Chou-Fasman technique [19,20], perhaps because of its ease in implementation [26]. Secondary structure predictions from basic nonempirical considerations, such as statistical mechanics [60] or stereochemistry [69] have also been proposed.

However, none of these prediction methods has been found to be highly accurate. At best, only about 50 per cent of the residues in a given protein are correctly classified as elements with respect to the four secondary structure categories: α-helix, β-strand, reverse turn or irregular conformation [28]. It appears that residues widely separated in primary sequence have a substantial effect on secondary structure determination, but these far neighbour interactions are generally not included in the prediction techniques.

Even when the secondary substructures along the residue chain are known, the final stage of assembling the resultant secondary structures into a reasonable tertiary structure is far from straightforward. The algorithms developed for combining the secondary structures of a protein into tertiary structure [21,37,38,85,104] have not met with much success. The reverse turns and irregular segments are "flexible"; they are not fixed structures in evolutionary related proteins. Even though their geometry is local, they are determined by nonlocal interactions. They can thus been seen as points of least resistance in folding, and not as active folding elements. This makes the

relative orientations of the secondary structures difficult to predict. Small differences in the (ψ, ϕ) angular orientations of residues involved in a turn can result in completely different global structures.

Therefore, it appears that the information on secondary structure locations is not sufficient to predict an accurate tertiary structure [50]. One aspect where the techniques for predicting secondary structure locations do show much promise is in refining the resolution of structures whose tertiary characteristics are roughly known. Thus a complementary approach may be in order, where first a distance constraint model gives the correct global tertiary structure, and then is refined by a secondary structure prediction model and further refined by a model concerned with direct local energy minimization.

2.3 Distance Constraint Models.

Another alternative approach from direct free-energy minimization modelling is to use empirical and statistical methods to exploit the information available from the X-ray diffraction studies of crystalline proteins (*cf.*, Chapter 7.3.1) These folded structures can be investigated empirically for common structural restrictions that give rise to universal characteristics in the geometry. The tertiary structure of an unresolved protein may then be predicted by forcing its residues into a conformation sharing the properties of the known structures.

The geometry of all globular proteins with known tertiary structures have been found to contain common structural restrictions, which are naturally expressed as constraints on distances between pairs of atoms. For example, the distance between C_α-atoms of adjacent residues is a constant 3.80 Å. Tertiary structure prediction models

using this type of observation have shown much promise in recent years. These modelling approaches are variously referred to as distance constraint, distance geometry or semi-empirical models.

With distance constraint models, a small number of simple controls force the protein into its final tertiary structure. These models do not attempt to follow the folding process in any way, but when the constraints are well chosen, they do reflect the underlying dynamics of the folded state of the protein. In other words, the free-energy equations are implicit in the empirical constraints. Also, by using known native protein structures as their basis, the question of whether the real folded protein lies at a local or global free-energy minimum is irrelevant for distance constraint models.

The major contributions to this type of model will be outlined here. The specifics of these models are discussed in Chapter 7.3, wherein their prediction results are compared to the results generated by the present model.

The model of Goel, Yčas *et al.* [14,45,46,128] attempts to satisfy a set of distance constraints identically by writing the constraints in the form of a weighted penalty function (*cf.*, Chapter 7.3.2). Various constraint combinations are presented to be solved exactly with the constraints being either pairwise distance constraints, minimum or maximum bounds on distances, or set average constraints. The set averages are weak constraints requiring a set of residues to attain an average distance, with no restrictions on individual distances. The penalty function is solved by minimization in the corresponding Cartesian coordinates. A sequential optimization is performed where each residue is selected in turn and its position optimized while keeping all other residues fixed. Their method is quite successful at the prediction of final structures with various constraint combinations used as input. However, this approach seems to

be overly dependent on the choice of the initial configuration, which implies that their models may not be suitably constrained in a sense to be described in Chapter 4.

The approaches of Kuntz, Crippen *et al.* [30,50,51,52,61,62,63] are summarized in Chapter 7.3.3. Their more recent approach [50,51,52,62] first imposes a set of distance constraints directly on the matrix of all pairwise distances between the residues, limiting specified entries in this distance matrix to be within upper and lower bounds. Thus the model works directly with a geometry of pairwise distances (*cf.*, Chapter 4.1). The system is easily solved with respect to this coordinate system by simply assigning values in the distance matrix. Unfortunately, since it requires four, not three, pairwise distance coordinates to specify a point in distance space, an arbitrary distance matrix will correspond to a structure in R^n, where $n > 3$. Hence, the difficulty arises in making the nonlinear transformation of the optimized structure from the distance space, a space of higher than three dimensions in general, into R^3. There is no obvious way to perform this embedding process optimally, and the system behaves essentially as an overconstrained one. The most difficult step of this approach is to decide in some rigorous fashion which distance constraints to relax so that the distance matrix can be embedded in R^3, whether the embedding process occurs during or after the optimization step. In spite of this, the method shows very promising results in prediction of tertiary structure, and is improving as the properties of the transformation become more familiar.

The models of Wako and Scheraga [117,118,119,120], discussed in greater detail in Chapter 7.3.4, combine distance constraint bounds similar to those of Kuntz, Crippen *et al.* with algorithms for free-energy minimization. Mean distance constraints obtained from such sources as primary structure, secondary structure, hydrophobicity and hy-

drophilicity ratings of the residues, as well as estimated possible sites from chemical cross-linking studies, are used to force the protein into an energy conformation space close to that of the real structure, so that free-energy minimization can refine the structure. In addition, semi-empirical estimates are derived for possible candidates of residue pairs that are involved in short-, medium- or long-range contact with respect to tertiary structure. The protein structures are optimized for these distance constraints on two-dimensional or three-dimensional lattices. Scheraga [100] extends this model for tertiary structure prediction by outlining a procedure which consists of repeated cycling between the method of distance constraint optimization utilized above and techniques for free-energy minimization.

2.4 The Present Model.

The distance model to be presented in this thesis has employed the model of Goel and Yčas [46] as its starting point. However, it differs from the previous models in several important aspects, apart from the number and type of geometrical constraints involved. These differences mainly deal with the mathematical form of the objective function and constraints. The model is formulated as a standard nonlinear optimization problem. It was deliberately designed to be suitably constrained with respect to both Cartesian coordinates and the coordinates defined by pairwise distances between points (*cf.*, Chapter 4). Furthermore, the solution algorithm was expressly designed for this problem, using current ideas in nonlinear programming. As with the other distance constraint models, it does not yet attempt a high resolution prediction of the tertiary structure, only the correct global characteristics.

3 The Basic Parameters of Distance Constraint Models.

In this chapter, the general geometrical characteristics of the chain conformation of globular proteins are explored. These characteristics lend themselves naturally to description in terms of distance geometry coordinates.

In order to simplify the discussion of the geometry of the protein, a protein will be represented by the locations of the central C_α-atoms of its residues. This representation corresponds to a "virtual bond" description of the molecule [79], which is explained in Chapter 11.1.

3.1 Near Neighbour Distances.

A detailed discussion of the near neighbour distances of proteins is given in Chapter 11, where theoretical near neighbour parameters are calculated and analyzed.

The local peptide geometry determines the near neighbour distances for the residues. First of all, the distance between C_α-atoms of adjacent residues in a protein will generally be a constant, equal to 3.80 Å. These first neighbour distances are found to be essentially constant both from theoretical consideration of the basic geometry and from empirical data. It is this constraint that gives the protein its chain structure.

Pairs of residues that are not adjacent but close together in the primary structure are found to have separations that are nonconstant, but lie within strict minimum and maximum bounds. The minimum and maximum bounds on these near neighbour distances are determined by steric hindrance and van der Waals forces. These minimum bounds are best estimated by empirical methods (*cf.*, Chapter 11), and the present model employs the empirical results of Goel and Yčas [46] from X-ray diffraction data of 21 globular proteins to obtain these parameters. Maximum bounds less than $k \cdot 3.80$

also exist for the near neighbour distances. These maximum bounds are determined by theoretical methods in Chapter 11, to avoid using measurement errors from the X-ray diffraction data. Analysis of empirical distributions of these near neighbour distances reveals unimodal peaked distributions for both second neighbour and fourth neighbour distances [46]. Therefore, it is justifiable to use the mean values of these distributions as parameters. The mean values for these distances are best determined empirically, since X-ray diffraction errors will tend to be averaged by this calculation and any theoretical procedure would necessarily involve an empirical evaluation of secondary structure proportions.

The distribution of the third neighbour distances is bimodal [46], and it follows that the third neighbour average distance may not be a useful parameter for distance constraint models.

Usually, only first to fourth neighbour distances are included for near neighbour constraints in distance models. Pairwise distances for residues farther apart in the chain show larger variability, and therefore, the additional information gained by the inclusion of mean value constraints for these residues would be small.

Possible near neighbour distance parameters for use in distance constraint models are given in Table 21 of Chapter 9.3, including the relevant parameters used in the present model.

3.2 Distances Between Far Neighbour Residues.

Let the distance between two residues that are far apart with respect to primary sequence (*i.e.*, separated by more than about 8 residues) be referred to as "far neighbour" distances, to conform with the "near neighbour" terminology of the previous section.

Any exact knowledge of pairwise distances for residues that are far apart in the chain is extremely valuable in that the information gained from such data would be global in nature, not local [119]. However, these far neighbour distances show large variability empirically, and so their usefulness is limited. The twentieth neighbour distances, for example, are found to show a substantial variability within a single protein and their mean values also vary irregularly when compared over a set of proteins. No correlation was found between far neighbour mean distances or standard deviations for far neighbour distances, and therefore these statistics are not incorporated into the present model.

If mean values for far neighbour distances are to be used in distance constraint modelling, they should be best determined empirically, not theoretically. The coordinates obtained from X-ray diffraction studies would be used for this empirical evaluation, and errors in these coordinate values would tend to be averaged out by the calculation of mean values. Also, any theoretical procedure would have the disadvantage of involving some empirical evaluation of the secondary structure proportions.

All residue pairs that are far apart in the primary structure do lie within absolute minimum and maximum distances with respect to the tertiary structure. The minimum bound on the distances between far neighbour residues is controlled by steric hindrance and van der Waals forces. It is an absolute number, not dependent upon either the size of the molecule or the separation in primary sequence of the residues involved. This bound is difficult to determine theoretically because it depends upon the orientation of the sidechains of the various residues that are nearby in tertiary structure. The minimum bound is best estimated by empirical methods, and the value given by Havel *et al.* [50] is used in the present study.

Residues far apart with respect to primary structure also show a definite maximum distance of separation, which is a function of the length n of the primary sequence, but has a value much less than $3.80 \times n$. This characteristic is due to the tight hydrophobic packing of the residues, resulting in the *globular*, or roughly spherical, shape of the molecule.

The parameters used for the far neighbour distance constraints of the present model are given in Table 22 of Chapter 9.3. These include a maximum bound for the far neighbour distance between any two residues in a chain of length n (derived in Chapter 9.3) and a semi-empirically obtained parameter for the minimum bound. No far neighbour mean value parameters are employed.

3.3 Hydrophobicity Constraints.

Globular proteins conform to a "hydrophobicity" rule in an aqueous environment [31,58,128]. Some amino acids have hydrophilic sidechains, that are preferentially located on the outside surface of the molecule. Other types have hydrophobic sidechains, that tend to bury themselves beneath the surface. The hydrophobicity rule (*cf.*, Chapter 1.2.8) separates the twenty common amino acids by the tendencies of their residues to lie in the interior or on the exterior of the globule, essentially due to the chemical properties of their sidechains. The rule is only approximate, but it is valuable in providing global information about the tertiary structure.

The hydrophobicity rule can be expressed in terms of distance measurements between each residue C_α-atom and the centroidal point of the molecule, where the centroidal point is defined as the average Cartesian coordinate location of the C_α-atoms of the

residues:

$$(x_{cp}, y_{cp}, z_{cp}) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i, \frac{1}{n}\sum_{i=1}^{n} y_i, \frac{1}{n}\sum_{i=1}^{n} z_i \right). \quad (1)$$

The rule can be stated as a tendency for the C_α-atom of each amino acid type to fall into exactly one of three classes:

J_1 — hydrophobic — tends toward the centroidal point of the configuration.

J_2 — hydrophilic — tends toward the surface of the configuration, away from the centroidal point.

J_3 — ambivalent — has no tendency.

Table 2: Hydrophobicity Classification with Respect to the Centroidal Point.

Although the hydrophobicity rule is caused by sidechain chemistry, the volumes of globular proteins are large enough that the burying of residue sidechains will also result in the burying of a corresponding part of the backbone of the protein. In fact, there is found to be a strong correlation between backbone and sidechain orientations [74]. Therefore, it is acceptable to express the hydrophobicity rule in terms of the backbone C_α-atoms.

An important consideration for the implementation of the hydrophobicity restrictions in a distance geometry model is the method of rating each of the amino acid types with respect to their relative preferences for the inside and outside of globular proteins. There are numerous hydrophobicity classification indices for the amino acids [16,17,34,46,55,59,64,74,75,76,80,124], and six of these classifications are given in Table 1 of Chapter 1.

The residues are sometimes rated on hydrophobicity index scales based on local

energy considerations or chemical properties of the sidechains [16,17,34,55,59,64,80]. Numerical hydrophobicity scales have been developed [55] by experimental methods, in which the free energy of transfer of amino acid sidechains from ethanol to water [80] or from cyclohexylpyrrolidine to water [64] is taken as a measure of the contribution of each sidechain to the total hydrophobic effect. Hydrophobicity has also been investigated experimentally by neutron diffraction and the hydrogen exchange technique [59] and by statistical-chemical methods [16,17]. The statistical methods define hydrophobicity as a function of chemical properties, such as the presence or absence of $(-OH)$ or $(-NH)$ groups capable of forming hydrogen bonds, the presence of chemically basic groups, or the numbers of atoms other than hydrogen bonded to the first and second carbon atoms of the sidechains.

Alternatively, hydrophobicity can be defined empirically by examination of the known tertiary structures of globular proteins. This approach is different in principle from the experimental approach, which does not reflect the influence of secondary structures, chain connectivity or long-range interactions. Empirical methods can involve geometrically defining a "surface" for a protein, which can then be used to assign residues to the inside or the outside of the structure [124]. In this way, a relative hydrophobicity index can be compiled by observing the overall fraction of each residue type that is found inside the surface. Simpler empirical methods involve observation of the average relative distances of the residues from the centroidal point of the globule [46,74], the distribution of relative distances of the residues from the centroidal point [84], or the average orientations of the sidechains of the residues with respect to the centroidal point [74].

The hydrophobicity measures are often presented as numerical indexing scales,

resulting in twenty different hydrophobicity classes, one for each residue type. Due to the imprecision of the classification techniques, perhaps a more realistic scheme is to employ only three categories: hydrophobic, hydrophilic and ambivalent, as in Table 2 at the beginning of this section. Alternatively, a separate and fourth category can be justified [46], containing the residues of Gly and Pro exclusively. These two types do not behave as chemically hydrophilic residues but can be classified as empirically hydrophilic. This is because they tend to participate in hairpin turns, which usually occur toward the surface of the protein molecule.

As shown in Table 1 of Chapter 1, the various methods of measuring hydrophobicity give similar results, but with some notable differences. The experimental classifications arise from investigating chemical properties of the individual amino acids from small peptide studies (not studies of complete proteins) or from elaborate physicochemical weighted functions evaluating the non-covalent forces of a protein. None of the experimental classifications appears to be highly correlated with respect to the known protein structures. Since the hydrophobicity rule just reflects tendencies for residues to prefer inside or outside, and is not yet expressible in terms of physics, statistical results obtained from real protein structures will presently provide the best data for parameter estimation.

For the present model, the simple empirical classification of Goel and Yčas [46] is used. They divided the twenty naturally occurring amino acids into three hydrophobicity categories by semi-empirical observation, as in Table 3.

For this hydrophobicity classification, the centroidal point distances were measured for all residues in twenty-one globular proteins. The amino acids were then classified

Hydrophobics — Val, Leu, Ile, Phe, Met.

Hydrophilics — Arg, Asp, Glu, Gln, Gly, Lys, Pro.

Ambivalent — Ala, Asn, Cys, His, Ser, Thr, Trp, Tyr.

Table 3: Empirical Hydrophobicity Classification of Residues (from Goel and Yčas 1979).

geometrically with respect to their observed average distances from the respective centroidal points. Whenever an amino acid showed an inconsistent behaviour, presumably as measured by the standard deviation, it was classed as ambivalent.

Given these three hydrophobicity classes, the model implements the hydrophobicity condition as a set of radial distance tendencies from the centroidal point for the individual C_α-atoms of the residues. The numerical hydrophobicity parameters used in this model are presented in Table 24 of Chapter 9.3, and a full explanation of these hydrophobicity parameters is given in that chapter.

3.4 Chemically Derived Constraints.

There are other distance constraints that may be available from the chemistry of a specific protein, the most obvious ones being the location of disulfide bonds. Disulfide bonds are cross-links connecting pairs of Cys residues, which may be far apart in the primary sequence. These covalent bonds are probably not integral to the folding process but certainly aid in the stability of the folded protein. Disulfide bond locations are not strictly part of the primary structure information. However, they are stable covalent bonds and can easily be found by the same techniques as those used to determine primary sequence. For example, the proteins can be cleaved into small polypeptides,

and the peptides can be separated under chemical conditions such that the disulfides remain intact. The primary structure identities of the Cys residues linked by the disulfides can then be determined [28].

The model has an option to use a mean value parameter for the distance between Cys residues that are known to be connected by a disulfide bond. The numerical value for the parameter used in the present model is obtained from the empirical studies of Thornton [114]. The possible parameters for this type of constraint in distance constraint models are given in Table 25 of Chapter 9.3.

Using the techniques of bifunctional reagent bonding [125,126], nonradiative excitation energy transfer [2], fluorescence energy-transfer [11,100], proton nuclear magnetic resonance [7,116,127], nuclear Overhauser measurements [10] or other physicochemical techniques [23], alternate chemically derived constraints are possible.

In theory, cross-linking experiments using bifunctional reagents can derive information such as the location of medium-range pairs of Asp, Glu, Lys or Tyr residues [50]. However, these studies require considerable effort in practice and are often unreliable, due to the possibility of protein distortion or multiple effects from the reaction. This type of study holds promise for the future in the extra-primary prediction of distances between specific residue pairs.

The residues Phe, Tyr and Trp have aromatic sidechains. These residues tend to interact in pairs or larger networks within the hydrophobic cores of globular proteins, at pairwise distances of about 4.5 to 7.0 Å [13]. These medium-range aromatic-aromatic interactions are not well correlated, but may help to stabilize protein structure. They may be of some value in providing far neighbour constraints for distance geometry models if reliable prediction algorithms become available.

Chemical and physicochemical methods have been used sucessfully in distance constraint predictions for protein tertiary structures. For the protein ribonuclease, Scheraga and his co-workers [100] found by chemical techniques the locations of the four disulfide bonds, the proximity of His12, His119 and Lys41 in the active site, and the pairings of carboxyl groups with tyrosyl groups, which were all expressed in the form of distance constraints. They then used these constraints in an energy-minimization model to predict the tertiary structure. Cohen and Sternberg [23] used chemically derived distance constraints in the form of the locations of potential candidates for interacting residues in the central fold and the proximities of the His64 and His93 residues to the heme iron in order to predict the tertiary structure of myoglobin.

As observed by several authors [45,51,119], the exact knowledge of only a few short-range or long-range distances between far neighbours in the chain can greatly facilitate the final resolution of a tertiary structure prediction. On the other hand, the approximate knowledge of a great many medium-range distances may not be as effective in determining the final conformation. Thus, the use of extra-primary distance information is potentially very valuable, and may even be crucial for this type of model to be successful.

4 Suitably Constrained Systems.

This chapter will treat distance constraint models as a single class of problems. It involves formulating, and then putting into practice, objectives for distance constraint modelling.

In Section 4.1, general geometric aspects for describing a one-dimensional chain within a three-dimensional Euclidean space will be considered. The most natural choice of coordinate system for empirical constraint modelling is found to be the pairwise distances between points, and this geometry is explored in some detail. The complicated relation between these coordinates and the canonical Cartesian coordinates are outlined in Section 4.2. The rest of the chapter is concerned with deriving conditions such that a distance constraint model is suitably constrained; these conditions involve pointwise continuity properties of the mapping from distance coordinate space to Cartesian coordinate space, under the geometrical restrictions imposed by the model.

Some ideas concerning the suitable constraining of distance geometry models have been presented in Foster [40], where it was shown that the present model contains a necessary and sufficient number and type of constraints so that the solution space of optimized structures for a given protein will be small with respect to both distance coordinates and Cartesian coordinates. These ideas are continued and expanded in this chapter.

4.1 Choice of Coordinate System.

As in most distance constraint models, the tertiary structure of a protein will be approximated by the positions of the central C_α-atoms of its constituent residues. Various primary coordinate systems have been chosen in recent literature [70,71,79,86,87,88] to

elicit the relative positions of a protein's C_α-atoms in R^3. Since the distance between any two adjacent residues is effectively a constant, proteins are usually envisaged as open polygonal arcs of length n, where n is the number of residues. Still, knowledge of any one of the following sets of data is sufficient to uniquely determine the conformation of a protein represented by n residue points:

1. $3n$ Cartesian coordinates $\{(x_i, y_i, z_i) | \, i = 1, \ldots, n\}$.

2. $3n$ Cartesian coordinates with respect to the canonical orthogonal basis for R^3, given by $e_1 = (1,0,0)$, $e_2 = (0,1,0)$ and $e_3 = (0,0,1)$. Note that there are only $3n-6$ degrees of freedom in (#1) and (#2), but the remaining 6 coordinates are necessary to fix the molecule in R^3 with respect to rigid rotations and translations.

3. (κ, τ) – the curvature and torsion of the arc length [86,87,88]. These are the usual descriptions of a local reference frame from differential geometry.

4. (b, w) – parameters controlling the shape of a regular parametrized surface, where b is a size or "bulkiness" parameter and w controls the amount of twist [70,71].

5. (ψ, ϕ)– the dihedral Ramachandran angles for each residue (*cf.*, Chapter 11.1).

6. (θ, γ)– virtual bond angles connecting the C_α-atoms of the residues into a chain (*cf.*, Chapter 11.1).

7. $3n-6$ coordinates, consisting of $n-1$ distance coordinates $d_{i,i+1}$, plus $n-2$ values of $d_{i,i+2}$ and $n-3$ values of $d_{i,i+3}$, provided that the sign of $d_{i,i+3}$ is specified with respect to the plane determined by the points C^α_i, C^α_{i+1} and C^α_{i+2} [119].

8. $4n-10$ coordinates, consisting of independent elements of the set $d_{i,i+j}$ [119].

Each of these data sets can be seen to be coordinate systems on a manifold locally homeomorphic to Euclidean space R^3.

In Chapter 3, possible parameters for tertiary structure prediction models were discussed that could be estimated by utilizing the semi-empirical and theoretical results realizable from known tertiary structures. These parameters are based solely on the observed common geometrical characteristics of globular proteins.

The most striking geometrical characteristic of proteins is that the distance between first neighbour C_α-atoms is effectively constant. Most coordinate systems implemented for the description of protein configuration (*e.g.*, coordinate systems #3 - #8 above), are chosen so as to take advantage of this property. Cartesian coordinate systems (#1 - #2) do not easily incorporate this basic property.

Characteristic features of proteins involving residues that are close together but not adjacent in the primary structure, such as the geometry of secondary structures (*cf.*, Chapter 11.4.2), are also easily described in the various local coordinate systems (#3 - #6) or in distance coordinates (#7 - #8). Yet, those coordinate systems with a local frame of reference (#3 - #6) are very poor in dealing with non-local phenomena, such as a disulfide bond occurring between two Cys residues that are far apart in primary structure. However, observed "global" characteristics of proteins, such as disulfide bonding and the hydrophobic close-packing of residues, are important geometrical restrictions on the folded molecule which cannot be adequately described by any simple combination of local geometry rules. Distance coordinates (#7 - #8) then, are the "natural" coordinates for expressing the empirical geometric properties discussed in Chapter 2.

4.1.1 Distance Geometry.

Distance geometry may be defined as the study of Euclidean configurations using the distances between points as the primary coordinate system. The set of all distances between pairs of points for a collection of m points forms a matrix D with elements d_{ij}. The present model uses such distance matrices to display the results of tertiary structure predictions in Chapter 7. In these displayed results, the matrix elements are in the form of coded symbols for the distances between the C_α-atoms of the protein. When the distance matrix is used as a representation of the coordinates of a protein, it has the advantage of containing all the structural information of the C_α-atom positions (up to translations, rotations and reflections) in a two-dimensional form.

The distance matrix D has several obvious properties:

1. it is a symmetric $m \times m$ matrix ($d_{ij} = d_{ji}$);

2. all elements on the main diagonal are equal to zero ($d_{ii} = 0$);

3. all elements off the main diagonal are strictly greater than zero ($d_{ij} > 0, i \neq j$).

When the distance matrix is used to represent a protein configuration, elements of the first diagonal above (or below) the main diagonal represent first neighbour distances $d_{i,i+1}$, which are effectively constant. The elements close to the main diagonal represent local structures, whereas elements far from the main diagonal represent long-range, or *global*, structures. Any available global geometric constraints such as disulfide bond locations can immediately be incorporated into the distance matrix by specifying the distance d_{ij} between the two residues involved. A chain can be protected from self-intersecting by the requirement that all elements must be strictly greater than some

positive number, representing a minimum distance of approach.

It is not difficult to find a distance matrix that satisfies all the properties given above, and yet not have a realizable conformation in R^3. As an example [30], note that there is no arrangement of four points in two dimensions that satisfies the following distance matrix representation:

$$D = \begin{pmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \tag{2}$$

This distance matrix (2) represents a tetrahedron, requiring three dimensions.

Proceeding one step further, it is observed that the following distance matrix represents a four dimensional structure, and cannot be realized by any arrangement of five points limited to R^3:

$$D = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \tag{3}$$

Any four points of this configuration can be positioned at unit pairwise distances, resulting in the tetrahedral structure. However, the fifth point cannot be added in R^3 such that its distance from each of the other points is one unit. It is important to note that furthermore, there is no method of slightly perturbing the elements of this distance matrix (3) such that the modification would result in a three-dimensional configuration. It is not at all obvious which configurations of five points in R^3 would be most "similar" to this structure.

These examples emphasize related problems concerning the use of distance geometry for modelling:

1. how to generate a distance matrix that represents a three-dimensional structure;

2. how to test a distance matrix for correspondence to a three-dimensional structure;

3. how to optimally embed the points of an arbitrary distance matrix from R^n into R^3.

The necessary and sufficient conditions for being able to embed k points in R^n, for any given n, are derived in a theorem due to Blumenthal [9], which is here specialized to three dimensions, as in Crippen [30]. This theorem does not utilize distance matrices directly, but instead is concerned with bordered matrices of squared distances. The matrix of squared distances $D^{(2)} = \{d_{ij}^2\}$ is defined as the matrix whose elements are the squares of the elements of D. The bordered matrix of squared distances, $D_b^{(2)}$, is a matrix consisting of $D^{(2)}$ augmented by an additional row and column consisting of all ones except for their common diagonal element, which is given the value zero. A Cayley-Menger determinant is the determinant of a bordered matrix of squared distances $\det(D_b^{(2)})$, as follows:

$$\det(D_b^{(2)}) = \begin{vmatrix} 0 & d_{01}^2 & d_{02}^2 & \dots & 1 \\ d_{10}^2 & 0 & d_{12}^2 & \dots & 1 \\ d_{20}^2 & d_{21}^2 & 0 & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 1 & \dots & 0 \end{vmatrix}. \tag{4}$$

Now the theorem can be given. Assuming points to be uniquely distinguished by their distances to other points, the theorem can be employed to test for the embeddability of any set of k points in Euclidean three-space. This theorem should also prove valuable in the attempt to modify any general matrix of distances into one that is embeddable in R^3.

THEOREM (Blumenthal). *A necessary and sufficient condition that a semimetric k-tuple may be irreducibly embeddable in R^3 for $k \geq 4$ is that the sign of all nonvanishing Cayley-Menger determinants of M points be given by $(-1)^M$ for all $M \leq 4$,*

at least one Cayley-Menger determinant of 4 points is nonzero, and the value of all Cayley-Menger determinants of more than 4 points is zero.

For the proof of this theorem, the reader is directed to Blumenthal [9] or Havel *et al* [51]. In Havel *et al.* [51], this theorem is used to generate algorithms for testing the dimensionality of distance matrix configurations.

Havel *et al.* [51] show that the conditions for embeddability in R^3 can be reduced to a series of tests on the Cayley-Menger determinants. For instance, the condition on the Cayley-Menger determinant for two points in Euclidean space is

$$\det(D_b^{(2)}\{p_0,p_1\}) = \begin{vmatrix} 0 & d_{01}^2 & 1 \\ d_{10}^2 & 0 & 1 \\ 1 & 1 & 0 \end{vmatrix} = 2d_{01}^2 > 0. \tag{5}$$

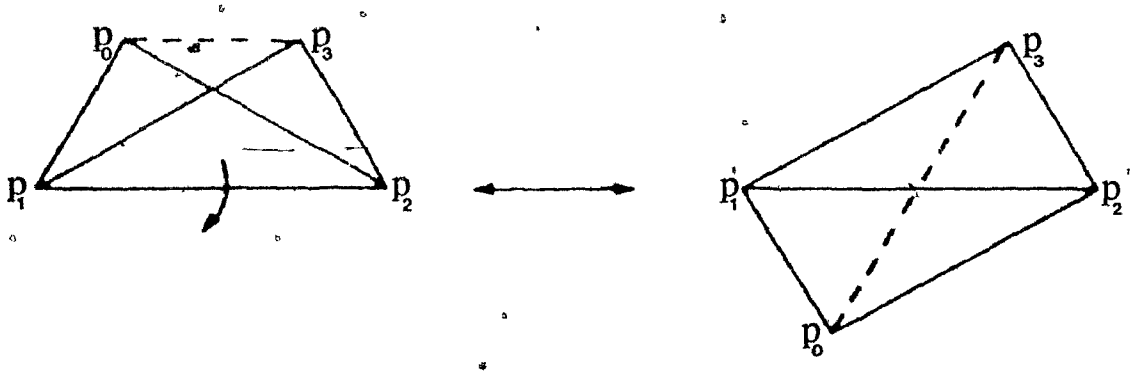
This merely states that the Euclidean distances must be real numbers.

For three points, the condition on the Cayley-Menger determinant is

$$\det(D_b^{(2)}\{p_0,p_1,p_2\}) = \begin{vmatrix} 0 & d_{01}^2 & d_{02}^2 & 1 \\ d_{10}^2 & 0 & d_{12}^2 & 1 \\ d_{20}^2 & d_{21}^2 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{vmatrix} \leq 0, \tag{6}$$

which reduces to restrictions on the distances d_{01}, d_{02} and d_{12}. The distance d_{02} must be less than or equal to $d_{01} + d_{12}$ (the usual triangle inequality), and must be greater than or equal to $|d_{01} - d_{12}|$ (an "inverse triangle inequality" lower bound). Similar restrictions must hold for the other permuted distances d_{01} and d_{12}. The points are also constrained to be not all collinear.

The Cayley-Menger determinant restrictions on four points results in a "tetrangle inequality". With reference to Figure 6, this inequality requires that d_{03} be restricted to distances attainable when the dihedral angle of rotation about the line segment connecting points p_1 and p_2 is between $-180°$ and $+180°$. A formula relating d_{03} to the angle of rotation is derived in [51]. From Figure 6 it is seen that the first configuration



The two extreme solutions of the tetrangle inequality for a configuration of four points are shown. The first configuration cannot be continuously deformed into the second without passing through a three-dimensional configuration.

Figure 6: The Tetrangle Inequality for a System of Four Points.

cannot be continuously deformed into the second without passing through a three-dimensional configuration, given that the other distances remain constant.

Finally, two equality relations, the "pentangle equality" and the "hexangle equality" must also be satisfied to ensure embeddability in R^3. These relationships arise from Cayley-Menger determinant restrictions on five and on six points, respectively. For these relationships, sign relations for various dihedral angles need to be compared for each set of five and of six points.

The Cayley-Menger restriction for five points has a similar interpretation as that for four points, except in three dimensions. There are found to be two possible distances d_{04} that result in realizable configurations in R^3, and it is not possible to pass from one configuration to the other without either altering some of the other distances or passing through a four-dimensional configuration. This results in a multiple minimum problem for distance geometry optimization, because two structures with similar

distance matrices can be expected to be separated by a barrier when restricted to R^3.

The Cayley-Menger restriction for six points is necessary to ensure that the relative chiralities for each combination of four points contained in any R^3 configuration of six points are mutually consistent. Any configuration of six points is uniquely determined if all but one of the distances (*e.g.* d_{05}) are known. The distance d_{05} must correspond to one of the two solutions of the pentangle equality for each of the four sets of five points that contain p_0 and p_5. Changing to the other solution of the pentangle equality in any of these sets of five points will reverse the relative chiralities of its two quartets of points that do not include d_{05}. Thus the hexangle equality creates an additional barrier between R^3 configurations, prohibiting continuous transformations of the distances that alter the relative chiralities.

Testing for each of these equality and inequality relationships for each subset of points in a large structure can be seen to be very time consuming and quite difficult computationally. However, the use of distance coordinates is natural for the description of a polygonal arc. Two contrasting approaches have been attempted to optimize protein tertiary structures by using distances between the points as the primary coordinates.

Crippen [30] was the first to consider the direct application of distance space coordinates for the protein folding problem. His work has since been substantially extended by Kuntz, Crippen and co-workers [50,52,62,63].

In the approach of Crippen [30], a distance matrix is initially optimized with respect to all required constraints on distances between residues, with no restrictions on the Euclidean dimensionality of the resultant matrix. Then a three-dimensional configuration is sought containing a set of pairwise distances that are similar to those

of the optimized matrix. In practice, the latter stage consists of an exhaustive application of the triangle inequalities followed by an exhaustive search of the tetrangle inequality from the optimized matrix in order to find *any* three-dimensional structure with distance properties similar to that of the optimized matrix. Due to the enormous computer time involved, this method could only be tested on very small systems (8 points). Even if more efficient algorithms were introduced for embedding a general R^n distance matrix into R^3, the criteria for optimally embedding this matrix would remain even more difficult to implement. Optimal embedding would probably involve choosing the element from the set of all R^3-corresponding distance matrices such that the RMS_y (*cf.*, equation (8) of Section 4.2) with respect to the optimized matrix is minimized. The multiple-minimum problem arising from the pentangle and hexangle equalities make this optimal embedding computationally intractable.

The second approach [51] attempts to maintain three-dimensionality at each step of the optimization process. This method writes the set of distance constraints as a penalty function, to be minimized directly in distance space. The penalty function is augmented with a set of equality and inequality constraints, as follows:

$$\text{Minimize } p(x) \tag{7}$$

subject to:

$$\begin{aligned} g_1(x) &\leq 0 \\ g_2(x) &\leq 0 \\ &\vdots \\ g_m(x) &\leq 0 \\ h_1(x) &= 0 \end{aligned}$$

$$h_2(x) = 0$$

$$\vdots$$

$$h_n(x) = 0.$$

The constraints represent the triangle inequalities, tetrangle inequality, and pentangle and hexangle equalities for each subset of points. The amount of computation involved in this approach has also proved to be formidable. Havel *et al.* [51] offer several suggested approaches for implementing this type of system (7).

4.2 The Mapping $\psi : Y \to X$.

In this section, the relationship between the two principal coordinate representations for protein tertiary structures will be investigated. Subsection 4.2.1 contains an outline of a derivation due to Rosen [96] of linear criteria concerning the continuity of the mapping $\psi : Y \to X$ from the space of distance coordinates to the space of Cartesian coordinates. In [96], Rosen found conditions in the linear case under which near-optimal values of a function expressed in terms of distance coordinates will yield correspondingly small errors with respect to Cartesian coordinates.

However, the transformation from distance geometry coordinates into Cartesian coordinates can be given in explicit form for the case when the distance matrix can be embeddable in R^3. When an explicit calculation of this coordinate transformation is derived as in Subsection 4.2.2, the nonlinear version of the Rosen criteria can be found. This is accomplished in Subsection 4.2.3, with the final nonlinear criteria for pointwise continuity of the mapping $\psi : Y \to X$ being found to involve expressions relating the Cartesian coordinates to perturbations of squared distances. Finally, in Subsection 4.2.4 it is demonstrated that structures that are close with respect to distance

coordinates under the present model will necessarily be close with respect to their Cartesian coordinates. Thus the present model is shown to be suitably constrained with respect to the number and type of constraints.

It is clear that the objective function and constraints are essentially defined on a space (call it Y) of distances between residues. Details of the geometry of this distance space, as well as problems associated with working directly in these coordinates, have been described in Section 4.1.

A basic goal of the present research is to clarify the properties of distance models in general. The present model will be solved by transforming the distance coordinates into the coordinates of the usual orthogonal Cartesian system (call this X), although comparisons of the final structure to other structures will be made in Y. These comparisons are made in order to study the efficiency of the model as one of a class of distance models, and to judge the relative merits of the different functions that comprise the function space of distance constraint models.

The relationship between these distance coordinates and the usual Cartesian coordinates is complicated. Nevertheless, their mapping $\psi : Y \rightarrow X$ and the comparison of structures with respect to X-space cannot be ignored. Suppose that a point $y^* \in Y$ is a solution of the constrained problem, and let $x^* \in X$ be the same point expressed in Cartesian coordinates. It is necessary that y^* be in a neighbourhood of the global minimum not only in Y, but also in X. However, a study of the mapping $\psi : Y \rightarrow X$ reveals that a small arbitrary perturbation $y^0 \rightarrow y'$ in Y can result in a large change in the corresponding coordinates $\psi(y^0) \rightarrow \psi(y')$ in X [98]. In other words, structures that are close together according to some metric in Y-space may turn out to be dissimilar with respect to a metric in X-space. The minimization of a function F on Y

can be regarded as a way to restrict the perturbations $y^* \rightarrow y'$ that are allowable, by the requirement that $F(y')$ be close to the minimum $F(y^*)$. Rosen [95,96] derives linear conditions on a function F such that if $F(y')$ is close to $F(y^*)$ then the distance in X between $\psi(y')$ and $\psi(y^*)$ is small (*cf.*, Subsection 4.2.1). It is demonstrated in Subsection 4.2.4 that similar conditions are met by the present model, and thus when the model reaches a neighbourhood of the solution to the constrained problem, the X coordinates of the resulting configuration must necessarily be close to those of the optimized structure. This property of the model ensures that the algorithm will not be an underconstrained one in which the optimized structures are highly dependent upon the initial pre-folded configurations. On the other hand, the model is not so highly constrained as to become trapped in a myriad of local minima.

Along with the obvious goal of being able to obtain folded configurations similar to the *in vivo* structures, any mathematical model formulated in distance geometry coordinates (Y-space) must satisfy three conditions [40] in order to yield useful results:

I. Accurate solutions y^* of the mathematical model (*i.e.*, tertiary structure predictions) must be achievable in practice.

II. The space of all possible solutions $y^* \in Y$ must be small, in a sense to be made precise below.

III. The domain of the solution space must be small with respect to the usual Cartesian coordinates in R^3 (X-space). In other words, all of the solutions $x^* \in X$ must be "close together" with respect to some X-space metric.

Condition I simply states that there must exist an algorithm that is guaranteed to systematically find a solution if there is one, given an initial configuration. This con-

dition may seem trivial, but is decidedly nontrivial with large-scale nonlinear systems such as these.

Assuming that the primary structure folds, if condition II does not hold, then the algorithm is underconstrained and the predicted tertiary structures will be found to be dependent upon the initial configurations. Different initial configurations will produce y^* values which are far apart in the solution space.

Condition III refers to local continuity properties of the mapping $\psi : Y \rightarrow X$ [96]. Differences in structures with respect to Y-space are generally measured by the distance root-mean-square metric:

$$RMS_y = \left[\frac{1}{N}\sum_{(i,j)}(d - d')^2\right]^{\frac{1}{2}} \tag{8}$$

where d and d' are the set of respective $d_{i,i+j}$ distances for the two structures, N is the total number of residue pairs and the summation is over all pairs of residues. Differences in structures with respect to X-space coordinates will be measured by the X-space root-mean-square metric:

$$RMS_x = \left[\frac{1}{3n}\sum(x - x')^2\right]^{\frac{1}{2}} \tag{9}$$

where x and x' are the usual Cartesian coordinates for the two structures, $3n$ is the number of coordinates in a protein of length n, and the summation is over all coordinates. These two metrics are discussed in more detail in Chapter 7.

It has been found that small perturbations of a structure in Y-space as measured by RMS_y can produce large changes in the structure with respect to the X-space coordinates as measured by RMS_x. For instance, Sanati [98] has shown an example of a small change in RMS_y causing a large change in RMS_x for an algorithm similar to that of Goel and Yčas [46] in the protein bovine pancreatic trypsin inhibitor. One

purpose of a distance geometry model must be to eliminate all such perturbations since, clearly, the existence of dissimilar optimized structure predictions x^* for a single protein is unacceptable.

4.2.1 Linear Criteria for a Suitably Constrained System.

It is essential that all models formulated in distance space should demonstrate restrictions on the pointwise continuity of the function $\psi : Y \rightarrow X$. If not, neighbouring structures calculated in Y may turn out to be only remotely related with respect to X.

Any configuration C of points in Euclidean space can be represented with respect to a Cartesian frame (X-space) by a particular choice of coordinates $\mu(C) = (x_1, x_2, \ldots, x_m)$, and in distance space (Y-space) by a corresponding set of coordinates $\nu(C) = (y_1, y_2, \ldots, y_k)$. Each of these two coordinate sets will totally represent the same configuration; therefore, a mapping $\psi : Y \rightarrow X$ can be established between them. This mapping expresses the fact that, if the complete set of pairwise distances are known, it is possible to generate a corresponding Cartesian coordinate representation. Similarly, an inverse mapping $\phi : X \rightarrow Y$ is calculable.

The mapping $\psi : Y \rightarrow X$ can be determined as a set of m functions:

$$\psi = (f_1, \ldots, f_m),$$

where $f_i(y_1, \ldots, y_k) = x_i$. The function f_i may be interpreted as the means of calculating a single Cartesian coordinate of some point p from the overall set of distances y_k.

Given any specific point $y^\circ = (y_1^\circ, \ldots, y_k^\circ) \in Y$, the pointwise continuity properties of $\psi : Y \rightarrow X$ can be investigated by determining the amount of error induced in X

when y° is replaced by a neighbouring point $y' \in Y$. This is of considerable importance in the comparison of configurations. If the mapping $\psi : Y \rightarrow X$ is not sufficiently smooth, structures that are neighbouring with respect to Y may turn out to be far apart in X. For protein tertiary structure prediction it is required that, when a conformation is generated with pairwise distances similar to those of the real structure, the generated conformation must also be similar to the real structure in X-space. This condition must hold in order for a distance constraint model to produce meaningful results. This condition is especially pertinent since distance constraint models by nature do not specify a protein configuration exactly in Y, but only to the extent of satisfying a set of general average distance characteristics.

Consider some definite configuration C°, whose unique representation in Y is given by:

$$\nu(C^\circ) = (y_1^\circ, \ldots, y_k^\circ).$$

The properties of $\psi : Y \rightarrow X$ within a vicinity of the configuration C° are here of concern. In this subsection, the situation will be investigated by means of a linear analysis due to Rosen [96].

In the Y-space neighbourhood of the configuration C°, each of the functions f_i may be approximated by a linear Taylor's expansion as follows:

$$\begin{aligned} \eta_i &= f_i(y_1', \ldots, y_k') - f_i(y_1^\circ, \ldots, y_k^\circ) \\ &\approx \sum_{j=1}^{k} \frac{\partial f_i(y_1^\circ, \ldots, y_k^\circ)}{\partial y_j^\circ}(y_j' - y_j^\circ). \end{aligned}$$

This expresses the error η_i introduced into the ith coordinate function $f_i = x_i$ in terms of the errors $\epsilon_j = (y_j' - y_j^\circ)$ produced by replacing the given point y° by a nearby point $y' \in Y$.

From the above, it is seen that the RMS_x error between $\psi(y')$ and $\psi(y^\circ)$ is given by

$$RMS_x = \left[\sum_{i=1}^{m} \frac{1}{m}(x_i' - x_i^\circ)^2\right]^{\frac{1}{2}} = \left[\sum_{i=1}^{m} \frac{1}{m}\eta_i^2\right]^{\frac{1}{2}}.$$

If the RMS_x is to be small, then each η_i must be individually small.

If each η_i is required to be small for every sufficiently small perturbation $y^\circ \rightarrow y'$ in Y, then this condition can only be satisfied for those points y° which simultaneously come close to extremizing all of the functions f_i. That is, the set of η_i can only be small simultaneously if all the partial derivatives $\partial f_i/\partial y_j$ are individually close to zero at y°. However, while this extremely strong condition is sufficient, it is not necessary. This is because the space of allowable perturbations $y^\circ \rightarrow y'$ will be restricted to those which are compatable with the condition that $F(y')$ is small.

For the linear case, Rosen [96] has derived the necessary condition such that all of the η_i will be small simultaneously. This is given as follows. The point y° must have the property that, for each function f_i, the gradient vector:

$$\nabla f_i = \left(\frac{\partial f_i}{\partial y_1}, \ldots, \frac{\partial f_i}{\partial y_k}\right) \tag{10}$$

evaluated at y° must be approximately orthogonal to the perturbation vector $(\epsilon_1, \ldots, \epsilon_k)$ for every perturbation $y^\circ \rightarrow y'$ such that $F(y')$ is small. To test a particular distance constraint model for this condition, it only remains to ascertain the subspace of allowable perturbations. This will vary according to the composition of each particular function $F(y)$.

4.2.2 Explicit Form of the Mapping $\psi : Y \rightarrow X$.

To quantify conditions II and III, the mathematical form of the transformation $\psi : Y \rightarrow X$ and the effect of distance geometry models on ψ must be investigated. In order to derive the mapping $\psi : Y \rightarrow X$, assume that the complete set of Y-space coordinates for a protein of length n is known. It is sufficient to know $4n - 10$ coordinates, consisting of independent elements of a set of d_{ij}, in order to determine the complete coordinate set. A mapping ψ was derived originally by Crippen [30], and can be written as follows:

$$
\begin{aligned}
(x_1, y_1, z_1) &= (0,0,0) \\
(x_2, y_2, z_2) &= (d_{12}, 0, 0) \\
(x_3, y_3, z_3) &= \left(\frac{{d_{13}}^2 - {d_{23}}^2 + {d_{12}}^2}{2d_{12}}, \left[{d_{13}}^2 - {x_3}^2 \right]^{\frac{1}{2}}, 0 \right) \\
(x_4, y_4, z_4) &= \left(\frac{{d_{14}}^2 - {d_{24}}^2 + {d_{12}}^2}{2d_{12}}, \frac{{d_{14}}^2 - {d_{34}}^2 + {d_{13}}^2 - 2x_3x_4}{2y_3}, \right. \\
&\qquad \left. \left[{d_{14}}^2 - {x_4}^2 - {y_4}^2 \right]^{\frac{1}{2}} \right) \\
(x_i, y_i, z_i) &= \left(\frac{{d_{1i}}^2 - {d_{2i}}^2 + {d_{12}}^2}{2d_{12}}, \frac{{d_{1i}}^2 - {d_{3i}}^2 + {d_{13}}^2 - 2x_3x_i}{2y_3}, \right. \\
&\qquad \left. \pm \left[{d_{1i}}^2 - {x_i}^2 - {y_i}^2 \right]^{\frac{1}{2}} \right) ; \; i = 5, \ldots, n.
\end{aligned}
\tag{11}
$$

Crippen also notes that the mapping is not stable. For example, if $d_{13} \simeq d_{23} >> d_{12}$, then a small perturbation in any of these Y-space coordinates would cause a large change in the value of x_3. A distance geometry model will act to reduce the effective degrees of freedom in Y-space and a good model must eliminate the instabilities in the mapping.

It must be noted here that Sippl and Scheraga [103] have produced a stable version of the mapping $\psi : Y \rightarrow X$, effectively resolving the troublesome $\pm$ term for z_i in

Crippen's mapping. In their derivation, it is assumed from the outset that the distance matrix corresponds to a structure in R^3, and so the essential problem of the embedding of a general distance coordinate set into R^3 is not considered. However, this section is more concerned with the local continuities of ψ, to show that conditions II and III hold under the present model. For this exercise, Crippen's mapping will serve as a better example.

4.2.3 Nonlinear Criteria for a Suitably Constrained System.

It is proven here that conditions II and III are met by the present model, and thus when the model reaches a neighbourhood of the solution to the constrained problem, the X coordinates of the resulting configuration must necessarily be close to those of the optimized structure. This proof is attained by explicitly calculating the f_i and examining the effect of an arbitrary perturbation vector.

Assume that an optimum point $y^* \in Y$ has been reached with respect to the model. A small perturbation $y^* \rightarrow y'$ should serve to push the configuration to a non-optimum one (condition II). If the model is suitably constrained, the change in structure with respect to X-space should also be small (condition III).

Unfortunately, under most models a perturbation $y^* \rightarrow y'$ will affect the X-space coordinates of the first four residues, as well as those of (x_i, y_i, z_i). However, since Crippen's equations (11) will hold for the choice of any four points for the coordinate references, any set of four virtual residues may be chosen, as follows:

1. (0,0,0) — the centroidal point of the molecule.

2. $(x_2,0,0)$.

3. $(x_3, y_3, 0)$.

4. (x_4, y_4, z_4).

Under an arbitrary perturbation, the resulting change in the point p_i will be given in X-space coordinates by:

$$\begin{aligned} x_i' - x_i^* &= \frac{(d_{1i}'^2 - d_{1i}^{*2}) - (d_{2i}'^2 - d_{2i}^{*2})}{2d_{12}} \\ y_i' - y_i^* &= \frac{(d_{1i}'^2 - d_{1i}^{*2}) - (d_{3i}'^2 - d_{3i}^{*2})}{2y_3} - \frac{x_3}{y_3}(x_i' - x_i^*) \\ z_i' - z_i^* &= \frac{(d_{1i}'^2 - d_{1i}^{*2}) - (d_{4i}'^2 - d_{4i}^{*2})}{2z_4} - \frac{y_4}{z_4}(y_i' - y_i^*) - \frac{x_4}{z_4}(x_i' - x_i^*). \end{aligned} \tag{12}$$

Using this explicit transformation $\psi : Y \to X$ between the X and Y coordinates, an explicit nonlinear version of the Rosen criterion can now be stated.

A sufficient condition for η_i to be small is that $(d_{ji}'^2 - d_{ji}^{*2})$ be nearly zero for $j = 1, 2, 3$ and 4 for every $i = 1, \ldots, m$. This condition is almost equivalent to requiring the perturbation vector $(d_{1i}' - d_{1i}^*, d_{2i}' - d_{2i}^*, d_{3i}' - d_{3i}^*, d_{4i}' - d_{4i}^*)$ to be small with respect to every Y-space coordinate of the basis set. This is, of course, a very strong condition.

If the distances d_{1i} can be considered to be constant for every i (for example, by representing d_{1i} by the first neighbour distances), then a necessary condition for η_i to be small reduces to

$$\begin{aligned} d_{2i}'^2 - d_{2i}^{*2} &\ll x_2 \\ d_{3i}'^2 - d_{3i}^{*2} &\ll y_3 \\ d_{4i}'^2 - d_{4i}^{*2} &\ll z_4. \end{aligned} \tag{13}$$

This condition may be equivalently stated by requiring the perturbation vector to be approximately orthogonal to the vector

$$\left[\frac{d_{2i}' + d_{2i}^*}{x_2}, \frac{d_{3i}' + d_{3i}^*}{y_3}, \frac{d_{4i}' + d_{4i}^*}{z_4}\right] \tag{14}$$

for every perturbation such that $F(y')$ is small.

As in the linear case, the event of this condition being satisfied for a particular distance constraint model will depend upon the individual form of the function $F(y)$ that is employed. Note that the vector given in expression (14) is quite dissimiliar from its counterpart expression (10) in the linear case. It is thus found that the linear does not give a truly necessary condition for small arbitrary perturbations η_i. However, it may hold for all practical cases, in which the magnitudes of d'_{ji} and d^*_{ji} are similar and the distance d^*_{ji} is not close to zero.

4.2.4 A Suitably Constrained Model.

If, as in the present model, exact conditions are required on the centroidal point, first neighbour and second neighbour distances at equilibrium, then the X-space coordinates of the four reference points will be invariant under any perturbation $y^* \rightarrow y'$. Under the requirements of this model, it follows that:

$$\begin{aligned} x_i' - x_i^* &= 0 \\ y_i' - y_i^* &= 0 \\ z_i' - z_i^* &= \frac{d_{4i}^{*2} - d_{4i}'^2}{2z_4} \end{aligned} \tag{15}$$

The last expression simply reduces to $z_i' = \pm z_i^*$. Thus any perturbation $y^* \rightarrow y'$ under this model will limit the perturbations possible in X-space to a set of measure zero. The addition of further minimum or maximum constraints to the model may limit the allowable perturbations in Y-space to zero. Therefore, the model as written in Chapter 9.2 is suitably constrained.

When the model does not include exact second neighbour constraints, analytical

results are extremely difficult to obtain. The argument that conditions II and III are satisfied must proceed heuristically. A test for such a model is to optimize the same protein from several random starting configurations (*cf.*, Chapter 7), and to further test the model by optimizing with slightly changing parameter values from the same initial configuration.

This concludes the discussion of distance coordinate geometry and its special relationship to Cartesian coordinates in the representation of protein tertiary structure.

5 The Mathematical Model

A protein will be modelled by the coordinate locations of the central C_α-atoms of its amino acid residues. The initial data used will be the primary structure of the protein, namely, the number and sequence of its residues. For the model, no other data is needed *a priori* from the specific protein to be folded. However, other data such as the locations of disulfide bonds or chemically derived information may be incorporated into the model as required.

In order to predict the tertiary structure of single-strand globular proteins, known structures found by X-ray diffraction are examined empirically for geometric properties that are universal. The properties take the form of distance restrictions between pairs of points: between various residue pairs or between single residues and the centroidal point of the structure. A discussion of possible distance constraints to be used in the model is given in Chapter 3. The actual numerical values for the parameters of the present model are derived in an appendix (Chapter 9.3). A set of empirically found distances from Goel and Yčas [46] is used to develop many of the parameters in the model.

The mathematical model employed in the thesis is formulated to meet several objectives. The model incorporates the information from the semi-empirical distance constraints in such a way that it remains simple in form and easy to modify. Equally important, it is capable of methodically obtaining solutions (*i.e.*, predicted tertiary structures) for an entire class of primary structures. This class consists of all single strand globular proteins, with virtually no restriction on primary sequence length. The model is suitably constrained, in the sense described in Chapter 4, so that the predicted

tertiary structures are not dependent upon the imposed initial prefolded conditions, and the range of the solution space is small with respect to both Cartesian coordinate space and distance coordinate space.

The model is written as a nonlinear programming problem, which is described in detail in an appendix (Chapter 9).

The hydrophobicity condition is presented as the objective function of the programming problem. The constraints consist only of first neighbour mean distances, second neighbour mean distances (although another constraint subset may be substituted for these distances in a subsequent model), and minimum and maximum bounds for other near neighbour and far neighbour distances. The constraints also include disulfide bond distances where applicable. The model can easily take advantage of other chemically derived distance constraints that may be available, for instance the location of medium-range residue pairs of Lys, Tyr, Glu or Asp residues [50] or interaction distances between pairs of aromatic residues [13]. The model is formulated in terms of distance geometry coordinates, since this formulation most naturally reflects the empirical constraints. However, it is more efficient to solve the model in the space of Cartesian coordinates (Chapter 10).

6 The Optimization Algorithms.

The distance constraint model is formulated as a nonlinear programming problem which in turn is solved by employing a penalty function. The programming formulation and the conversion into a penalty function are explained rigorously in Chapter 9. The overall problem is solved by combining the objective function (hydrophobicity conditions) and the constraints into a quadratic loss penalty function, alternately minimizing the penalty function for a fixed value of the constraint weights and strengthening the constraint weights. This effectively transforms the constrained problem into a sequence of unconstrained problems. The expression for the penalty function is nonconvex; therefore, an overall solution may correspond to a strong local minimum, and not to a global minimum.

The overall optimization method for solution of the nonlinear programming problem is explained fully in Chapter 10. It was designed specifically for this problem by the author, with the technical assistance of P.F. O'Neill of the Department of Mathematics, Statistics and Computing Science at Dalhousie University [41]. The minimization of the penalty function is accomplished by employing a strategy based on the steepest descent algorithm and Newton's method. Steepest descent is the principal gradient search algorithm for nonlinear optimization of a continuously differentiable function, whereas Newton's method is the fundamental nonlinear algorithm exploiting the second order information from the variables [39]. The optimization technique further employs a refinement strategy for large-scale systems, called the truncated-Newton method [33]. The truncated-Newton algorithm saves on computer execution time and memory space by calculating only an approximate solution to the Newton equations at each step,

using a conjugate gradient iterative scheme. Hence, the penalty function is minimized by an appropriate combination of steepest descent steps and truncated-Newton steps. The overall solution method was designed in conjunction with the mathematical model, and fully exploits the sparsity of the Hessian matrix of second derivatives.

There is no standard procedure for solving large-scale nonlinear systems of equations, and there are not as yet standardized test problems for evaluating solution techniques. In the present case, the best method was to design an algorithm which could utilize explicit second order information of a continuously differentiable function, while maintaining low storage space from the sparse Hessian.

7 Results.

The results of the present model will be described in this chapter, both as numerical results in tabular form and graphically as two-dimensional contact maps. These results will be compared to the published results of other distance constraint models in Section 7.3. Some of the results shown, specifically those of Figures 7 and 8 and Tables 8 and 12 for the proteins BPTI (initial configuration A) and lysozyme, have previously been published in Foster [40].

The model results are obtained from a FORTRAN implementation of the algorithm (Chapter 10) on three globular proteins, performed on a CYBER 170-730 computer at Dalhousie University, in Halifax, Nova Scotia, Canada. The mathematical form to be optimized is the penalty function $p(x;\mu)$ given in Chapter 9.2, using the parameters itemized in Chapter 9.3. The computed protein structures are compared to known X-ray diffraction coordinates at each stage of the optimization process (*i.e.*, for each value μ of the penalty function $p(x,\mu)$). The coordinate structures obtained from X-ray diffraction studies of crystalline proteins are regarded as the "real" structures. These X-ray diffraction coordinates were supplied by the Protein Data Bank of the Brookhaven National Laboratory [1].

The initial configurations used are sets of n points, generated at random in R^3, but scaled approximately to the volume of the protein. These initial sets of points are thought to provide suitably unbiased structures, in that no actual tertiary structure information is provided *a priori*. The execution time required for the nonlinear optimization can be substantially decreased by generating initial configurations with correct first neighbour characteristics (*i.e.*, random chain conformations), and this will

be implemented in future studies. However, it was decided that this preprocessing could be viewed as influencing the optimization process and therefore was not included for the present study.

The metric widely used for comparing protein structures is the RMS_y, which is defined as follows:

$$RMS_y = \left[\frac{1}{N}\sum_{(i,j)}(d-d')^2\right]^{\frac{1}{2}}. \qquad (16)$$

Here d and d' are the sets of respective $d_{i,i+j}$ distances for the two structures being compared, $N = n(n-1)/2$ is the total number of residue pairs for a protein containing n residues, and the summation is over all residue pairs. This distance (Y-space) root-mean-square value is the most commonly used metric for measuring the similarity in protein structures, mainly due to the difficulty in calculating root-mean-square values with respect to Cartesian (X-space) coordinates. Differences in structures with respect to X-space coordinates are measured by the RMS_x metric, defined by:

$$RMS_x = \left[\frac{1}{3n}\sum(x-x')^2\right]^{\frac{1}{2}}. \qquad (17)$$

In equation (17), x and x' are the usual Cartesian coordinate representations for the two structures, $3n$ is the number of coordinates in a protein of length n, and the summation is over the set of all coordinates. The problem in calculating RMS_x arises because this value as shown in equation (17) is dependent upon relative rotations or reflections of the two structures. In actual evaluations for protein comparisons between the coordinates of the crystal structure $\{x\}$ and the generated structure $\{x'\}$, a unique value for RMS_x is obtained by rotation of the primed frame so as to minimize the value of the expression (17). This optimal rotation has an essential singularity, and therefore is difficult to compute.

Cohen and Sternberg [22] state that the RMS_x provides a more effective and significant method of comparing structures than the RMS_y. However, it is shown in Chapter 4 that the present model is suitably constrained so that small RMS_y differences will also be small with respect to RMS_x. Also, the RMS_y measure was chosen for the present study in order to provide a meaningful comparison to the results of other distance constraint models, where the RMS_y was often used exclusively.

As a general guideline, an RMS_y value of 1–3 Å implies that the two structures are very similar, whereas an RMS_y value greater than 6 Å indicates that the two proteins may have dissimilar global structures [22,93]. There exists a weak direct relationship between the RMS_y and the length n of the chain; therefore, it is suggested both by Cohen and Sternberg [22] and by Remington and Matthews [93] that any structural comparisons be judged in the light of the expected value for a random structure of the same size.

7.1 Numerical Results for Structural Predictions of Three Globular Proteins.

The globular proteins used for this study include rubredoxin, bovine pancreatic trypsin inhibitor (referred to as "BPTI") and lysozyme.

Rubredoxin is a carrier molecule, a very small non-heme iron-sulfur protein that functions as an electron transporter. It is found in many anaerobic bacteria and is similar in structure to bacterial ferrodoxin, which it can replace in certain enzymatic reactions. The molecule is composed of 54 amino acid residues and has a molecular weight of approximately 6100. The sulfur atoms of its four Cys residues are coordinated to a single iron atom in a tetrahedral Fe-S_4 complex. From the crystallographic evidence, it is believed to be a pliable, readily deformable molecule [53,121].

Bovine pancreatic trypsin inhibitor (BPTI) has become a standard test case for tertiary structure prediction models because it is a very small protein containing several important structural elements of globular proteins: α-helix, antiparallel β-strands and hairpin turns. One of numerous inhibitor proteins, whose physiological function is to inhibit the digestive enzymes, it acts to inhibit trypsin and other proteases by binding with them as a pseudosubstrate. BPTI contains 58 amino acid residues (MW = 6500), which are cross-linked by three disulfide bridges. These disulfide bonds may account for its high stability against denaturation. The molecule has dimensions of approximately $43 \times 23 \times 49$ Å. It contains two domains of secondary structure: a double stranded antiparallel β structure, composed of residues 16 through 36, and an α-helix composed of residues 47 through 56. The β structure is considerably distorted, but the α-helix is very regular between residues 48 and 54. BPTI also contains four internal water molecules, occupying space in crevasses of the outer surface. This unusually high number of internally-bonded water molecules is probably a consequence of the small size of the BPTI chain. The peptide bond Lys15 - Ala16, which appears at the binding site of BPTI to trypsin, is strained and deviates significantly from planarity [32].

Lysozyme was the first true enzyme to have its tertiary structure determined by X-ray diffraction [8]. Its function is to dissolve certain bacteria by hydrolyzing the polysaccharide component of their cell walls, causing the cell to lyse. Lysozyme is a relatively small enzyme (129 amino acids, MW = 14600), crosslinked by four disulfide bridges, which contribute to its high stability. It is a compact molecule, roughly ellipsoidal in shape, with dimensions of $45 \times 30 \times 30$ Å. Residues Asp101, Trp63 and Trp62 are the main binding participants in the active site. The active site also involves residues Asn59 and Ala107, as well as Glu35 and Asp52. There is an α-helix contain-

ing residues 5 through 15, two helices intermediate between α-helices and 3_{10}-helices at residues 24 through 34 and residues 88 through 96, and a 3_{10}-helix containing residues 80 through 85. There is an incompletely-formed antiparallel β-sheet involving residues 41 through 48 versus residues 49 through 54 [8,77].

The numerical results from the optimizations for the three proteins are summarized in Tables 4 - 7. Each row of Tables 4 - 7 represents one call to the algorithm Inner Loop (see Chapter 10). The column "RMS_y" indicates the root-mean-square deviation of pairwise distances between the tertiary structure of the protein found from X-ray diffraction techniques and the structure returned by Inner Loop. It is noted that the X-ray diffraction structure is comprised of coordinates of the crystalline form of the protein and is at best an averaging of *in vivo* states of the true fluctuating structure. However, it is at present the best available representation of the actual structure for purposes of comparison.

The RMS_y differences found between the optimized structures and their real counterparts were determined to be 4.88 for rubredoxin (no disulfide bond constraints), 5.58 for BPTI (no disulfide bond constraints included), 4.22 for BPTI (all three disulfide constraints included) and 5.75 for lysozyme (all four disulfide constraints included). According to the generalized probability distribution result of Remington and Matthews [93], these results represent structural agreements that are better than average (*i.e.*, random chain) by approximately 2.7, 2.1, 2.9 and > 4.0 standard deviations, respectively. The frequencies with which these levels of agreement are expected to occur by chance in a random population are approximately 0.3%, 1.8%, 0.2% and $< 0.001\%$, respectively. For these calculations, the RMS_x and RMS_y measures are

assumed to be approximately linearly related by

$$RMS_y = 0.75 RMS_x + 0.19 \qquad (18)$$

(Cohen and Sternberg [22]).

Rubredoxin (54 amino acids)
(contains no disulfide bonds)

Steepest Descent	Negative Curvature	Newton	CPU (sec)	$RMS_y(4)$	RMS_y	ΔRMS_y
322	0	0	119	2.36	5.06	6.70
94	0	0	126	2.17	4.98	1.11
79	0	0	101	2.14	4.98	0.27
0	19	21	501	2.02	4.88	0.99
			847			

Table 4: Numerical Results: Rubredoxin.

Bovine Pancreatic Trypsin Inhibitor (58 amino acids)
(disulfide bonds not included)

Steepest Descent	Negative Curvature	Newton	CPU (sec)	$RMS_y(4)$	RMS_y	ΔRMS_y
342	0	0	147	2.20	5.99	7.41
106	0	0	175	2.07	5.78	1.17
95	0	0	154	2.02	5.74	0.33
0	12	17	383	1.86	5.58	1.41
			859			

Table 5: Numerical Results: BPTI, No Disulfide Bond Constraints.

The column "ΔRMS_y" indicates the root-mean-square difference between the structure input at the beginning of Inner Loop and the structure which is returned at the end of Inner Loop. For example, the first numerical entry in column ΔRMS_y represents the RMS_y difference between the initial structure of randomly generated points and the structure returned after the first call to Inner Loop.

Bovine Pancreatic Trypsin Inhibitor (58 amino acids)
(3 disulfide bonds included)

Steepest Descent	Negative Curvature	Newton	CPU (sec)	$RMS_y(4)$	RMS_y	ΔRMS_y
511	0	0	203	1.84	4.96	6.65
106	0	0	171	1.68	4.73	1.05
87	0	0	131	1.64	4.68	0.46
0	9	12	193	1.65	4.22	1.54
			698			

Table 6: Numerical Results: BPTI, Including Disulfide Bond Constraints.

Lysozyme (129 amino acids)
(4 disulfide bonds included)

Steepest Descent	Negative Curvature	Newton	CPU (sec)	$RMS_y(4)$	RMS_y	ΔRMS_y
546	0	0	475	2.58	6.61	9.54
69	0	0	557	2.46	6.37	0.98
184	0	0	1338	2.35	6.21	1.04
0	17	8	749	2.26	5.75	2.36
			3119			

Table 7: Numerical Results: Lysozyme.

Also, supplementary RMS_y measures are calculated where only the first to fourth neighbour distance terms are included in the RMS_y formula:

$$RMS_y(4) = \left[\frac{1}{4n-10}\sum_{j=1}^{4}\sum_{i=1}^{n-j}(d_{i,i+j} - d'_{i,i+j})^2\right]^{\frac{1}{2}} \qquad (19)$$

These "$RMS_y(4)$" values indicate the degree to which the near neighbour distances correspond in the real and optimized structures. The $RMS_y(4)$ results corresponding to the four optimized structures above were 2.02, 1.86, 1.65 and 2.26 Å, respectively.

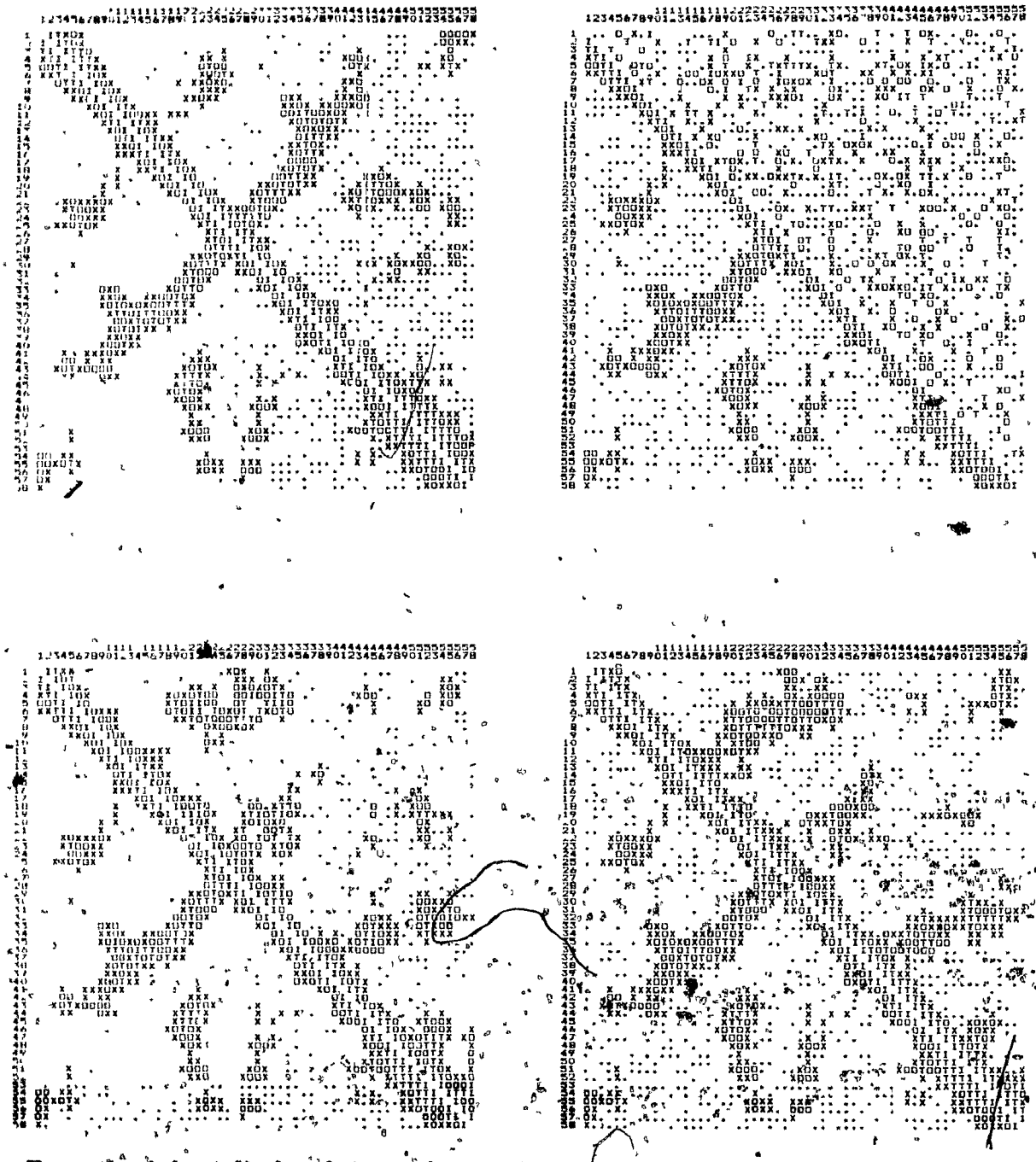
The columns "Steepest Descent", "Negative Curvature" and "Newton" indicate the number of iterations applied for each kind of descent direction. Note that the Newton steps were not attempted until the final call to Inner Loop in the present algorithm. Column "CPU" indicates the number of CPU seconds required by a CYBER 170-730 computer to perform the calculations. The minimum and maximum bound constraints for far neighbour residues were not applied until the second and subsequent calls to Inner Loop. This results in the CPU time per steepest descent iteration being much smaller during the first Inner Loop than in the other Inner Loops. From Tables 4 - 7, it is seen that the CPU time increases nonlinearly with n, as is expected.

The results of the optimizations are also displayed in the form of distance matrices, or "contact maps" [67,128]. In these displays, the elements of the Y-space coordinate matrix for a protein are shown symbolically. In Figures 7 - 15, if $d_{i,i+j} < 3.0$ Å for some i and j, then the corresponding position for $d_{i,i+j}$ in the matrix will be denoted by a blank space; if $d_{i,i+j} > 10.0$ (or $d_{i,i+j} > 12.5$ for lysozyme), the corresponding position will be denoted by a period; if $3.0 < d_{i,i+j} < 10.0$ (or $3.0 < d_{i,i+j} < 12.5$ for lysozyme), the position will be denoted by a coded letter symbol. Note that the main diagonal will always be blank, since it represents the set of distances $d_{i,i} = 0$.

Each protein structure shows a characteristic pattern on a contact map. Secondary structures are readily identifiable. An α-helix configuration shows as a broad band along the diagonal because it results in small near neighbour distances. Along parallel β-strands, the residues separated by some j positions in primary sequence will be close together in tertiary structure. This gives rise to a band of close contacts on a diagonal parallel to the main diagonal, but offset from it by j elements. The hydrogen-bonding patterns of antiparallel β-strands appear as a band running perpendicular to the main diagonal.

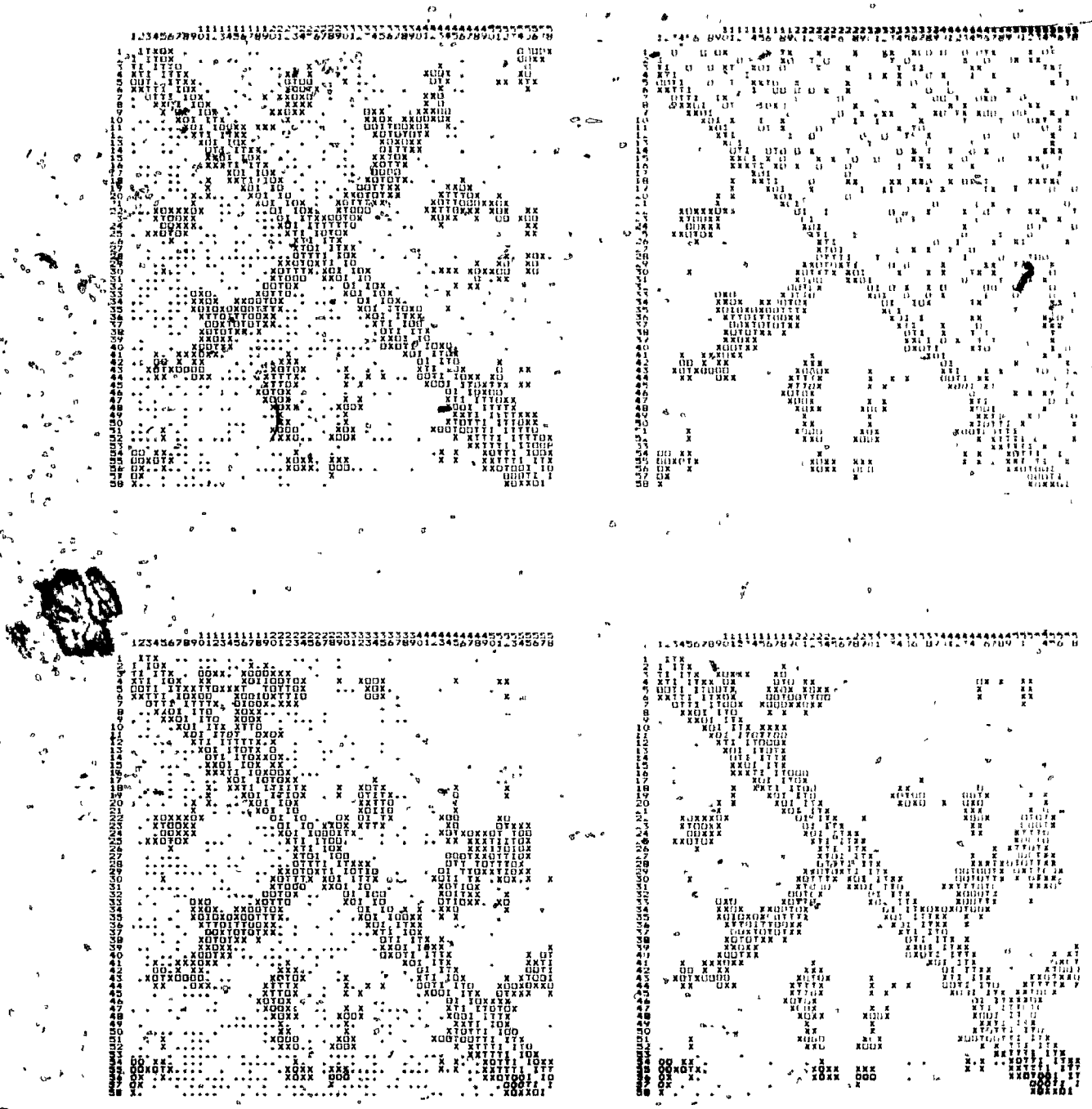
Since the distance matrix is symmetric, only one half of the matrix needs to be shown. For comparison purposes, the contact maps of Figures 7 - 15 will always show the real structure below the main diagonal, and the structure to be compared above the main diagonal.

In Figure 7 the contact maps for the optimization of BPTI are given, for the case in which the three disulfide bond constraints have been included. The optimized structure can be compared with the real one by observation of the diagram at the lower right of Figure 7. Both the real and the optimized structures show a well-formed α-helix for residues 45–58. Both structures also exhibit close contact structures containing the residues approximately 2–14 versus 14–26, although the optimized structure has many more close contacts in this region and may be considered to possess an approximate antiparallel β-structure here. The optimized structure does not predict the magnitude of the antiparallel structure for residues 4–26 versus 26–43, and also contains an extraneous close-contact substructure involving residues 31–36 versus 40–55. Therefore, although the two structures show similarities, there are also some differences with re-



For each of the 4 displayed contact maps, the real structure is shown below the main diagonal of the matrix. The following different representations are shown above the main diagonals in the contact maps: (i) the real structure, (ii) the initial configuration, (iii) the structure returned by the first call to Inner Loop, and (iv) the optimized structure. In the contact map, the distances $d_{i,i+j}$ between residues i and j (in Å) are letter coded as follows: *blank* = 0.0 to 3.0, I = 3.0 to 4.4, T = 4.4 to 6.0, O = 6.0 to 8.0, X = 8.0 to 10.0, and *period* = greater than 10.0.

Figure 7: Contact Maps for BPTI, Including Disulfide Bond Constraints.



For each of the 4 displayed contact maps, the real structure is shown below the main diagonal of the matrix. The following different representations are shown above the main diagonals in the contact maps: (i) the real structure, (ii) the initial configuration, (iii) the structure returned by the first call to Inner Loop, and (iv) the optimized structure. Letter codes for the distances $d_{i,i+j}$ are identical to those of Figure 7.

Figure 8: Contact Maps for BPTI, Excluding Disulfide Bond Constraints.

spect to secondary structures and close contact surfaces. This is not surprising since no secondary structural information was included *a priori* in the model.

Further details of the optimized structure for BPTI are given in Table 8. These statistics show that although general parameters nonspecific to the BPTI molecule were employed, the optimized configuration still conforms well to the real structure with respect to both the near neighbour distances and the distances of the individual residues from the centroidal point. The effectiveness of the hydrophobicity condition as represented by the objective function of the present algorithm is further illustrated in Figure 9, where the distances of the individual residues from the centroidal point are compared for the optimized and real structures. It is seen that the centroidal point distances for the optimized structure closely follow the pattern of centroidal point distances shown by the real structure.

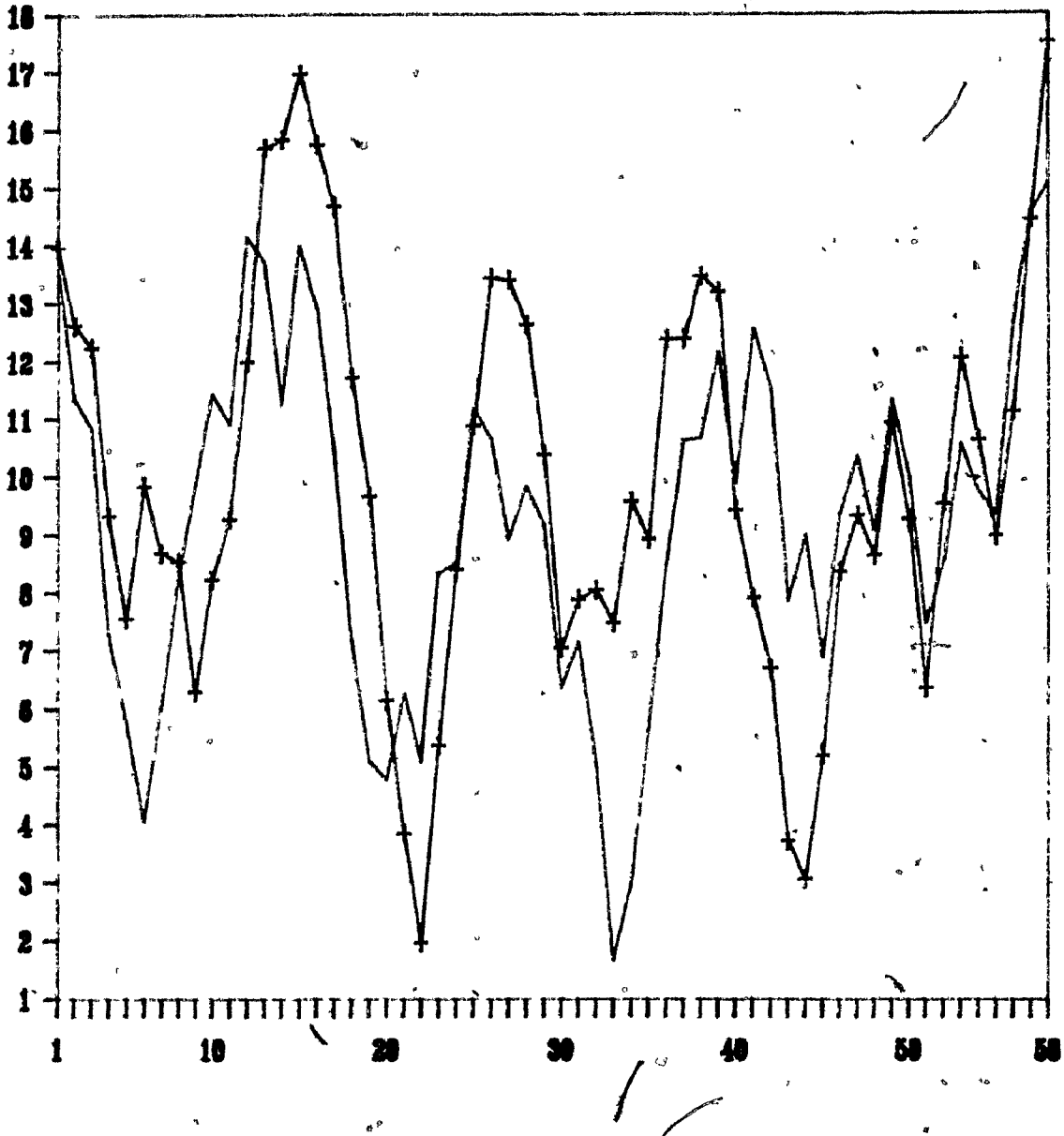
Figure 8 shows contact maps for the optimization of BPTI where the disulfide bond constraints have been removed. In the optimized structure, many of the secondary structures are now absent. The α-helix at 45-58 and the antiparallel structure at 4-26 versus 26-43 are both missing, although the short antiparallel structure at 2-14 versus 14-26 is present.

From Figures 7 and 8, as well as Tables 5 and 6, it is seen that the optimized structure is improved considerably by the inclusion of the disulfide bond locations. This demonstrates the importance of far neighbour constraints for these distance constraint models, even if the constraints are not known precisely. It may be that some chemically derived far neighbour distance information may be necessary for these models to consistently generate accurate predictions. Alternatively, it may be speculated

Distance	Real		Optimized		
	Mean	S.D.	Mean	S.D.	
$d_{i,i+1}$	3.80	0.03	3.80	0.003	
$d_{i,i+2}$	6.12	0.69	5.96	0.02	
$d_{i,i+3}$	8.12	1.71	7.85	1.25	
$d_{i,i+4}$	9.94	2.48	9.53	1.72	
Distance	**Real**		**Optimized**		**Mean Difference**
	Mean	S.D.	Mean	S.D.	
$d_{i,cp}(i \in J_1)$	8.48	2.88	5.81	2.44	3.63
$d_{i,cp}(i \in J_2)$	11.23	3.02	10.75	2.50	2.00
$d_{i,cp}(i \in J_3)$	9.30	3.85	9.19	2.45	2.21
$d_{i,cp}$(all)	9.99	3.48	9.28	2.99	2.36

Statistical comparison of the optimized structure for BPTI (3 disulfide bond constraints included) to the structure from X-ray diffraction results. The X-ray diffraction structure is considered "real". The amino acids are divided into three classes J_1 (hydrophobic), J_2 (hydrophilic) and J_3 (ambivalent). The notation $d_{i,cp}$ denotes the distance of residue i from the centroidal point of the molecule.

Table 8: Numerical Results for BPTI: Near Neighbour and Centroidal Point Distance Statistics.



The distances of the residues from the centroidal point (in Å) are plotted for the molecule BPTI. The individual centroidal point distances for the 58 residues of the optimized configuration of BPTI (3 disulfide bonds included) are represented by points (.), connected by a solid line. The centroidal point distances for the real structure are denoted by (+) signs, also connected by a solid line.

Figure 9: Distances of the Residues from the Centroidal Point for Optimized Structure of BPTI.

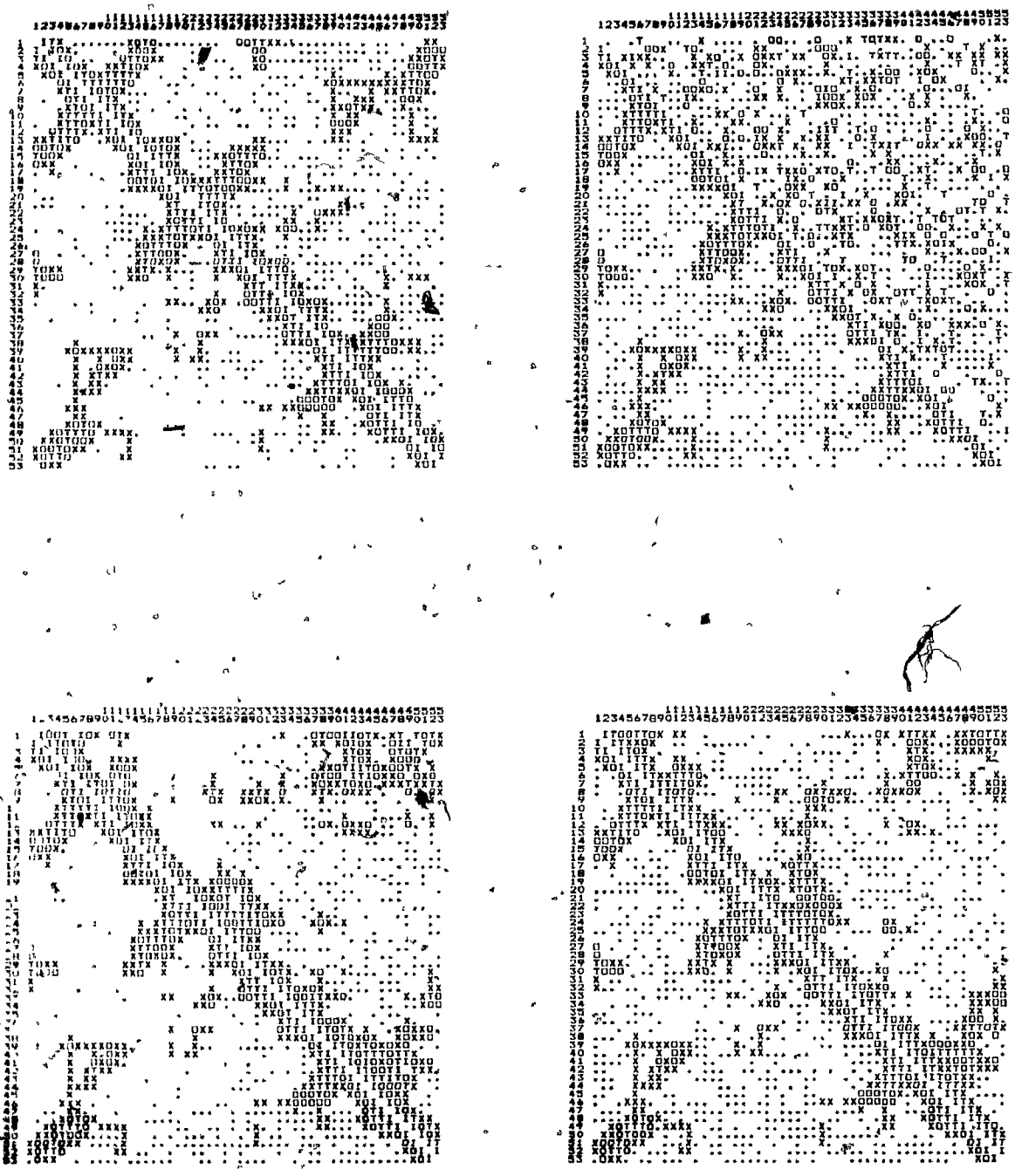
that the disulfide bond locations are an integral part of the stable tertiary structure of BPTI and a feature that cannot be predicted accurately by the model without explicit addition of the disulfide constraints.

The rubredoxin molecule was folded using the set of general parameters given in Chapter 9, with no disulfide bond constraints included. The four Cys residues of rubredoxin do not specifically form disulfide bonds in the real structure, but are found to be in close contact due to the formation of a prosthetic Fe-S_4 complex in the interior of the molecule.

The contact map for the real structure of rubredoxin (Figure 10, upper left) exhibits a noticeable lack of well-formed secondary structures. The three very short antiparallel substructures of the pattern are probably due to the major tertiary structure of rubredoxin: the Fe-S_4 tetrahadral complex involving residues Cys6, Cys9, Cys39 and Cys42. The contact map of the optimized structure (Figure 10, lower right) is very similar to that of the real structure. Both the real and optimized structures show the three small antiparallel regions and close contact between the *N*-terminal and *C*-terminal residues.

In Table 9, it is seen that, as with BPTI, the optimal structure closely conforms with the real structure in its near neighbour and centroidal point statistics, even though no parameters specific to rubredoxin were used in the optimization model.

Tables 10 and 11 give the pairwise distances for the four Cys residues involved in the tetrahedral Fe-S_4 complex of rubredoxin for the real structure and the optimized structure, respectively. No explicit constraints for these distances were included in the model for this chemical structure. The optimized structure seems to account for the proximities of the Cys residues despite this lack of information. An average pairwise



For each of the 4 displayed contact maps, the real structure is shown below the main diagonal of the matrix. The following representations are shown above the main diagonals in the contact maps: (i) the real structure, (ii) the initial configuration, (iii) the structure returned by the first call to Inner Loop, and (iv) the optimized structure. Letter codes for the distances $d_{i,i+j}$ are identical to those of Figure 7.

Figure 10: Contact Maps for Rubredoxin.

Distance	Real		Optimized	
	Mean	S.D.	Mean	S.D.
$d_{i,i+1}$	3.79	0.38	3.80	0.001
$d_{i,i+2}$	6.08	0.66	5.95	0.003
$d_{i,i+3}$	7.97	1.75	7.50	1.37
$d_{i,i+4}$	9.70	2.47	8.85	2.13

Distance	Real		Optimized		Mean Difference
	Mean	S.D.	Mean	S.D.	
$d_{i,cp}(i \in J_1)$	8.96	2.01	6.26	1.80	2.94
$d_{i,cp}(i \in J_2)$	9.66	1.72	10.68	1.59	1.91
$d_{i,cp}(i \in J_3)$	8.53	2.45	7.53	1.48	2.87
$d_{i,cp}$(all)	9.44	2.02	9.25	2.62	2.38

Statistical comparison of the optimized structure for rubredoxin (no disulfide bond constraints) with the structure from X-ray diffraction results. The X-ray diffraction structure is considered "real". The amino acids are divided into three classes J_1 (hydrophobic), J_2 (hydrophilic) and J_3 (ambivalent). Let $d_{i,cp}$ represent the distance of residue i from the centroidal point of the molecule.

Table 9: Numerical Results for Rubredoxin: Near Neighbour and Centroidal Point Distance Statistics.

distance of 9.87 Å was obtained in the optimized structure for all pairs of Cys residues; this is significantly smaller than the average pairwise distance of 12.88 Å found for all residues in the rubredoxin optimized structure (12.96 Å for all residues in the real structure). Also the maximum Cys-Cys distance of 13.06 Å in the optimized structure was much smaller than the maximum pairwise distance of 24.96 Å found for all residues in the optimized structure (25.90 Å in the real structure). The distances between Cys9 - Cys39 and Cys9 - Cys42 cannot be considered true close contacts. On the other hand, they do not represent global structural errors either, with the pairwise distances of nearby Cys9 - Val38 (9.90 Å), Val8 - Cys39 (9.86 Å), Val8 - Cys42 (8.87 Å) and Cys9 - Leu41 (10.36 Å) being respectably small in the optimized structure.

	Rubredoxin (Real Structure) (distances between the Cysteine residues)			
Residue	#6	#9	#39	#42
#6	0	5.77	6.37	8.44
#9		0	8.45	5.96
#39			0	5.83
#42				0

Distances for all six pairs of Cys residues are given in Å. The calculated mean is a separation of 6.80 Å, with a standard deviation of 1.29 Å.

Table 10: Distances Between the Cys Residues in Rubredoxin (Real Structure).

In future studies, the model will be used to predict the structure of rubredoxin, with the explicit inclusion of the proposed chemical constraints on the four Cys residues. It is expected that these constraints will improve the RMS_y error due to the addition of this extra-primary information. However, the improvement to the final structure may not be substantial, as evidenced by the correct juxtaposition of the Cys residues in the

Rubredoxin (Optimized Structure)
(distances between the Cysteine residues)

Residue	#6	#9	#39	#42
#6	0	8.76	10.69	5.00
#9		0	13.06	12.46
#39			0	9.23
#42				0

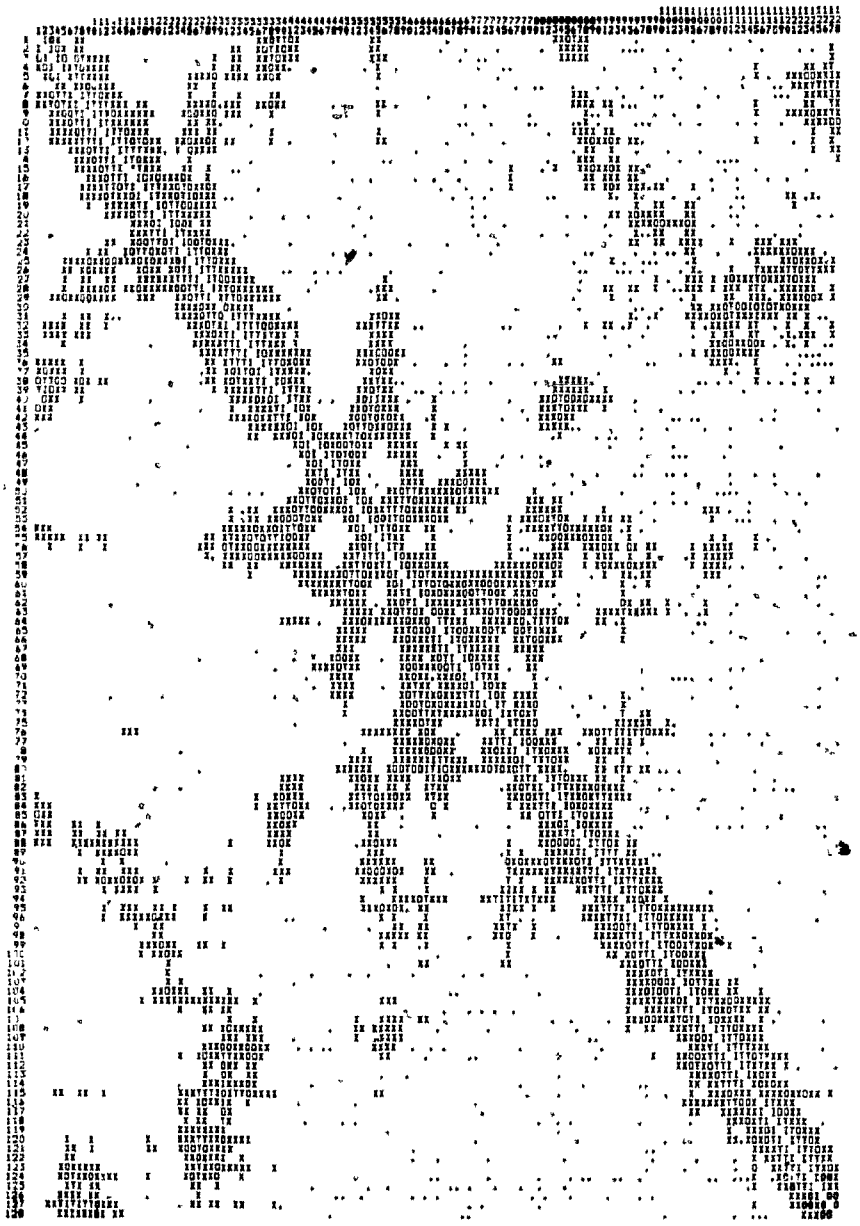
Distances for all six pairs of Cys residues are given in Å. The calculated mean is a separation of 9.87 Å, with a standard deviation of 2.93 Å.

Table 11: Distances Between the Cys Residues in Rubredoxin (Optimized Structure).

present optimized version. It appears that the Fe-S_4 complex may not be integral to the actual folding of rubredoxin but only serve to stabilize the final structure, and that the tertiary structure may be essentially attainable without it.

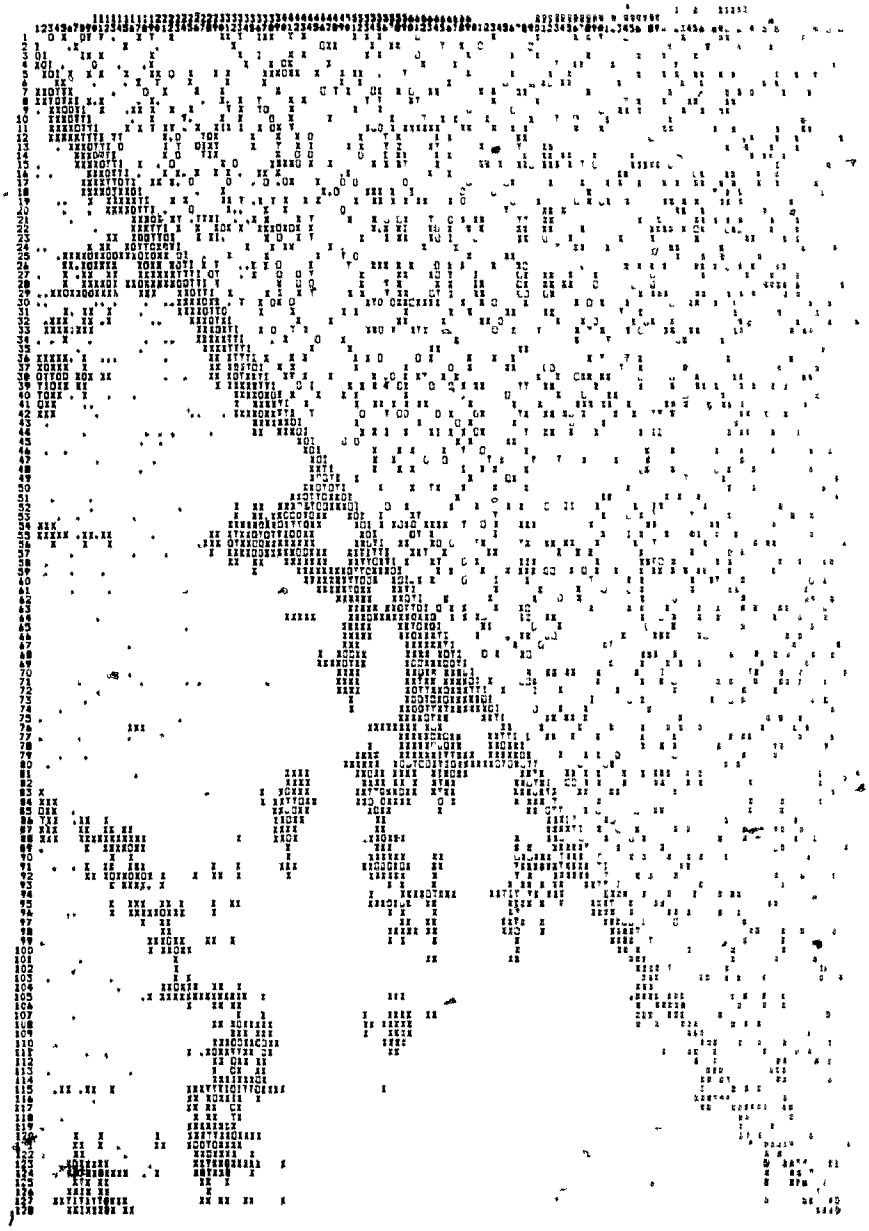
The lysozyme molecule contains 129 residues; therefore, it is much larger and more structurally complex than either BPTI or rubredoxin. Nevertheless, with the penalty function in its present form, the optimizing algorithm is capable of handling second order information to find strong optima for proteins several times the size of lysozyme. The tertiary structure of lysozyme was optimized with the inclusion of constraints for the four disulfide bonds. The contact map of the optimized structure, shown in Figure 14, has a pattern generally similar to the real structure, with some notable differences in the proximal residues near the active site.

The contact map for the optimized structure of lysozyme (Figure 14) is found to resemble the real structure in many ways; the structure of the last 30 *C*-terminal residues are similar, as are the patterns for the residues far apart in primary structure. Most of the close contacts in lysozyme are small and local in primary structure, implying



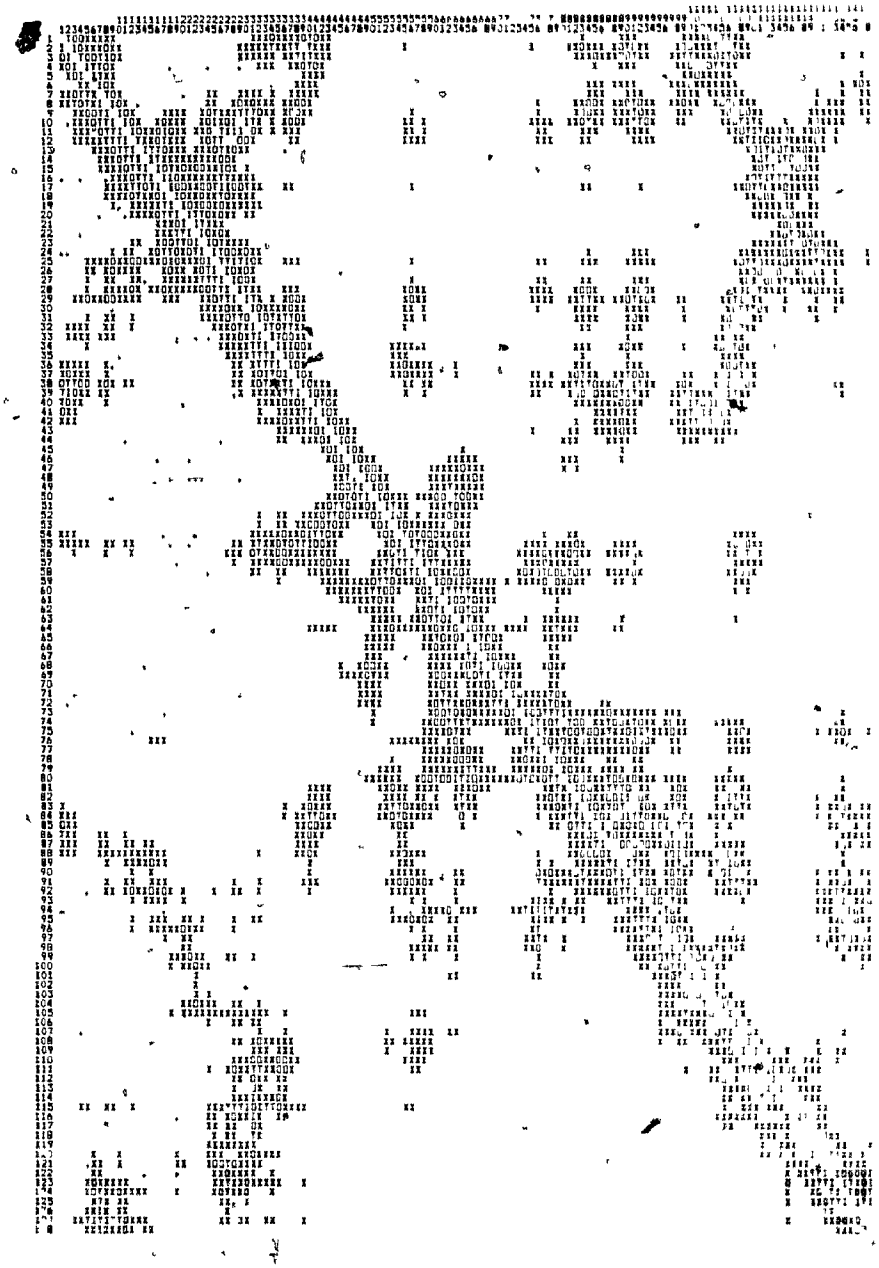
For each of the 4 displayed contact maps, the real structure is shown below the main diagonal of the matrix. The following representations are shown above the main diagonals in the contact maps: (i) the real structure, (ii) the initial configuration, (iii) the structure returned by the first call to Inner Loop, and (iv) the optimized structure. In the contact map, the distances $d_{i,i+j}$ between residues i and j (in Å) are letter coded as follows: ***blank*** = 0.0 to 3.0, I = 3.0 to 4.4, T = 4.4 to 6.0, O = 6.0 to 8.0, X = 8.0 to 12.5, and ***period*** = greater than 12.5.

Figure 11: Contact Maps for Lysozyme, the Four Disulfide Bonds Included. I. The Real Structure.



For each of the 4 displayed contact maps, the real structure is shown below the main diagonal of the matrix. The following representations are shown above the main diagonals in the contact maps: (i) the real structure, (ii) the initial configuration, (iii) the structure returned by the first call to Inner Loop, and (iv) the optimized structure In the contact map, the distances $d_{i,i+j}$ between residues i and j (in Å) are letter coded as follows: ***blank*** = 0.0 to 3.0, I = 3.0 to 4.4, T = 4.4 to 6.0, O = 6.0 to 8.0, X = 8.0 to 12.5, and ***period*** = greater than 12.5.

Figure 12: Contact Maps for Lysozyme, the Four Disulfide Bonds Included, II Initial Configuration



For each of the 4 displayed contact maps, the real structure is shown below the main diagonal of the matrix. The following representations are shown above the main diagonals in the contact maps: (i) the real structure, (ii) the initial configuration, (iii) the structure returned by the first call to Inner Loop, and (iv) the optimized structure. In the contact map, the distances $d_{i,i+j}$ between residues i and j (in Å) are letter coded as follows: *blank* = 0.0 to 3.0, I = 3.0 to 4.4, T = 4.4 to 6.0, O = 6.0 to 8.0, X = 8.0 to 12.5, and *period* = greater than 12.5.

Figure 13: Contact Maps for Lysozyme, the Four Disulfide Bonds Included. III. First Outer Loop.

For each of the 4 displayed contact maps, the real structure is shown below the main diagonal of the matrix. The following representations are shown above the main diagonals in the contact maps: (i) the real structure, (ii) the initial configuration, (iii) the structure returned by the first call to Inner Loop, and (iv) the optimized structure. In the contact map, the distances $d_{i,i+j}$ between residues i and j (in Å) are letter coded as follows: *blank* = 0.0 to 3.0, I = 3.0 to 4.4, T = 4.4 to 6.0, O = 6.0 to 8.0, X = 8.0 to 12.5, and *period* = greater than 12.5.

Figure 14: Contact Maps for Lysozyme, the Four Disulfide Bonds Included. IV. Optimized Structure.

that most of the secondary structures contain few residues, and in these local patterns the two structures are dissimilar. Most of the helical structures of the real structure are absent in the optimized structure, and extra close contacts are seen for residues 7–19 versus 20–32. The optimized structure thus has correctly predicted the global structure of lysozyme as measured by the RMS_y difference, but has errored in prediction of many of the local substructures as displayed in the contact map.

Distance	Real		Optimized		
	Mean	S.D.	Mean	S.D.	
$d_{i,i+1}$	4.32	1.47	3.80	0.01	
$d_{i,i+2}$	6.28	1.42	6.05	0.19	
$d_{i,i+3}$	7.35	2.00	8.08	1.18	
$d_{i,i+4}$	8.51	2.48	9.76	1.81	
Distance	Real		Optimized		Mean Difference
	Mean	S.D.	Mean	S.D	
$d_{i,cp}(i \in J_1)$	10.10	3.34	9.09	2.78	2.95
$d_{i,cp}(i \in J_2)$	15.12	4.11	14.99	3.12	2.61
$d_{i,cp}(i \in J_3)$	12.63	3.44	11.66	3.82	2.70
$d_{i,cp}$(all)	13.09	4.01	12.39	3.96	2.74

Statistical comparison of the optimized structure for lysozyme (4 disulfide bond constraints included) with the structure from X-ray diffraction results. The X-ray diffraction structure is considered "real". The amino acids are divided into three classes J_1 (hydrophobic), J_2 (hydrophilic) and J_3 (ambivalent). Let $d_{i,cp}$ represent the distance of residue i from the centroidal point of the molecule.

Table 12: Numerical Results for Lysozyme: Near Neighbour and Centroidal Point Distance Statistics.

The optimization statistics and RMS_y comparisons between the optimized structure and the real structure for lysozyme are found in Table 7. Further comparisons are found in Table 12. From the near neighbour $d_{i,i+j}$ comparisons in Table 12, it is seen

that the X-ray diffraction data employed is slightly in error, with a mean value of 4.32 Å for the first neighbour distance instead of the expected 3.80 Å. Also the convergence of the optimized structure is shown to be incomplete from the standard deviation results of 0.01 Å for $d_{i,i+1}$ and 0.19 Å for $d_{i,i+2}$, as compared to the corresponding values of 0.003 Å for $d_{i,i+1}$ and 0.02 for $d_{i,i+2}$ in the optimized structure for BPTI (Table 8). The centroidal point distances for the residues of the optimized structure are very close to their counterparts in the X-ray diffraction structure, differing by an average of only 2.74 Å.

An optimization of the lysozyme molecule was also carried out using only the first 128 residues, omitting the *C*-terminal Leu residue. This optimization resulted in a final tertiary structure that was remarkably dissimilar from the one discussed above. In fact, it showed an incorrect supersecondary structure in the neighbourhood of the *C*-terminus, with a final RMS_y result greater than 8 Å. This shows that the model can be sensitive to small changes in the primary structure, as is observed with *in vivo* proteins. Therefore, it is indicated that the model can correctly predict structural modifications resulting from primary structure insertions, deletions and substitutions, although this is a matter for further study.

The X-ray diffraction coordinates available for the present study were found to be incomplete in that the only 53 of the 54 residue C_α-atom coordinates were available for rubredoxin (*Clostridium pasteurianum*, 2 Å resolution, unrefined) and only 128 of the 129 C_α-atom coordinates were included for lysozyme (hen egg-white lysozyme, 2.5 Å resolution). Coordinates for the *C*-terminal residues were omitted in each case. Correspondingly, the RMS_y difference calculations for these two proteins were performed via the omission of the *C*-terminal residue for each protein. The effect of this

modification on the calculated results is expected to be negligible. The optimized structures of both these proteins correctly show the C-terminus behaving as a hydrophobic residue as judged by centroidal point distance, and also correctly predict the proximities of the C- and N-terminal residues in each case.

7.2 Repeated Optimization of BPTI from Different Initial Configurations.

Repeated runs from different initial configurations indicate whether or not the final structures are insensitive to the initial values. Also this exercise may be the best way to approach the global minimum for any such nonconvex problem.

Repeated optimizations from different random point initial conditions were performed for the globular protein BPTI, with inclusion of the three disulfide bond constraints in each case. All of these tests used identical parameter values and termination criteria for each Inner Loop.

The termination criteria used were as follows. For the Inner Loops that exclusively employed steepest descent iterations (Inner Loops $i = 1$, 2 and 3), the Inner Loop was terminated when the gradient norms at two successive iterations were calculated to be less than a specified value:

$$\|\nabla f(x^k)\| + \|\nabla f(x^{k-1})\| < \epsilon_i. \tag{20}$$

The values for the tolerances ϵ_i were given to be $\epsilon_1 = 10.0$, $\epsilon_2 = 10.0$, and $\epsilon_3 = 1.0$. Typically, the norm of the gradient for an initial configuration would be $\|\nabla f\| \approx 10^8$. The introduction of minimum and maximum far neighbour constraints at the second Inner Loop typically resulted in the initial gradient norm of the second Inner Loop having a value of $\|\nabla f\| \approx 2500$, and thus the stopping tolerance ϵ_2 also represents a

considerable reduction. At the start of the third Inner Loop, the gradient norm would generally have a value of $\|\nabla f\| \approx 20$.

For the first Inner Loop only, it was additionally required that the value of the function was to be reduced by a suitable amount. This constraint on the penalty function value was empirically chosen to be $p(x, 1.0) < 450$, where the initial configuration would typically correspond to a value of $p(x, 1.0) \approx 10^5$.

For the final Inner Loop, in which the truncated-Newton algorithm was performed, the termination criterion was as follows:

$$\|\nabla f(x^k)\| < \epsilon_4 = 10^{-3}. \tag{21}$$

Other stopping criteria in the form of time limits for the completion of each Inner Loop were available. However, these were found to be unnecessary in the optimizations performed. The results of these repeated optimizations are given in the following series of tables and figures. Initial configuration A is the same as that of Table 6, but the optimization results vary slightly because slightly different termination criteria were used for the two cases at each Inner Loop.

Bovine Pancreatic Trypsin Inhibitor (58 amino acids)
(3 disulfide bonds included)

Steepest Descent	Negative Curvature	Newton	CPU (sec)	$RMS_y(4)$	RMS_y	ΔRMS_y
788	0	0	271	1.85	4.78	6.78
91	0	0	151	1.65	4.57	1.01
116	0	0	158	1.62	4.52	0.41
0	11	8	171	1.63	4.21	1.30
			751			

Table 13: Numerical Results: BPTI, from Initial Configuration A.

Bovine Pancreatic Trypsin Inhibitor (58 amino acids)
(3 disulfide bonds included)

Steepest Descent	Negative Curvature	Newton	CPU (sec)	$RMS_y(4)$	RMS_y	ΔRMS_y
628	0	0	220	2.17	5.22	7.42
103	0	0	164	2.05	5.15	1.19
139	0	0	187	1.99	5.12	0.39
0	22	82	2245	1.90	4.84	1.46
			2816			

Table 14: Numerical Results: BPTI, from Initial Configuration B.

Bovine Pancreatic Trypsin Inhibitor (58 amino acids)
(3 disulfide bonds included)

Steepest Descent	Negative Curvature	Newton	CPU (sec)	$RMS_y(4)$	RMS_y	ΔRMS_y
678	0	0	236	1.80	4.92	7.52
174	0	0	250	1.69	4.78	1.51
95	0	0	127	1.69	4.80	0.25
0	9	12	283	1.74	4.95	0.63
			896			

Table 15: Numerical Results: BPTI, from Initial Configuration C.

Bovine Pancreatic Trypsin Inhibitor (58 amino acids)
(3 disulfide bonds included)

Steepest Descent	Negative Curvature	Newton	CPU (sec)	$RMS_y(4)$	RMS_y	ΔRMS_y
673	0	0	234	2.07	4.47	7.37
103	0	0	145	1.73	4.13	0.97
131	0	0	172	1.66	4.04	0.38
0	29	51	1534	1.73	3.94	1.17
			2085			

Table 16: Numerical Results: BPTI, from Initial Configuration D.

Bovine Pancreatic Trypsin Inhibitor (58 amino acids)
(3 disulfide bonds included)

Steepest Descent	Negative Curvature	Newton	CPU (sec)	$RMS_y(4)$	RMS_y	ΔRMS_y
652	0	0	228	2.24	4.93	7.08
152	0	0	218	2.07	4.78	1.11
150	0	0	213	1.89	4.67	0.63
0	55	57	2403	1.85	4.29	1.97
			3122			

Table 17: Numerical Results: BPTI, from Initial Configuration E.

Bovine Pancreatic Trypsin Inhibitor (58 amino acids)
(3 disulfide bonds included)

	A	B	C	D	E	Real
A	0	3.67	4.48	3.85	4.49	4.21
B		0	4.58	3.80	3.68	4.84
C			0	4.08	3.80	4.95
D				0	3.55	3.94
E					0	4.29
Real						0

Table 18: Comparison of Optimized RMS_y Structures for BPTI.

Bovine Pancreatic Trypsin Inhibitor (58 amino acids)
(3 disulfide bonds included)

	A	B	C	D	E	Real
A	0	1.19	1.14	0.99	1.33	1.63
B		0	0.98	1.10	1.26	1.90
C			0	0.96	1.05	1.74
D				0	1.08	1.73
E					0	1.85
Real						0

Table 19: Comparison of Optimized $RMS_y(4)$ Structures for BPTI.

Bovine Pancreatic Trypsin Inhibitor (58 amino acids)
(3 disulfide bonds included)

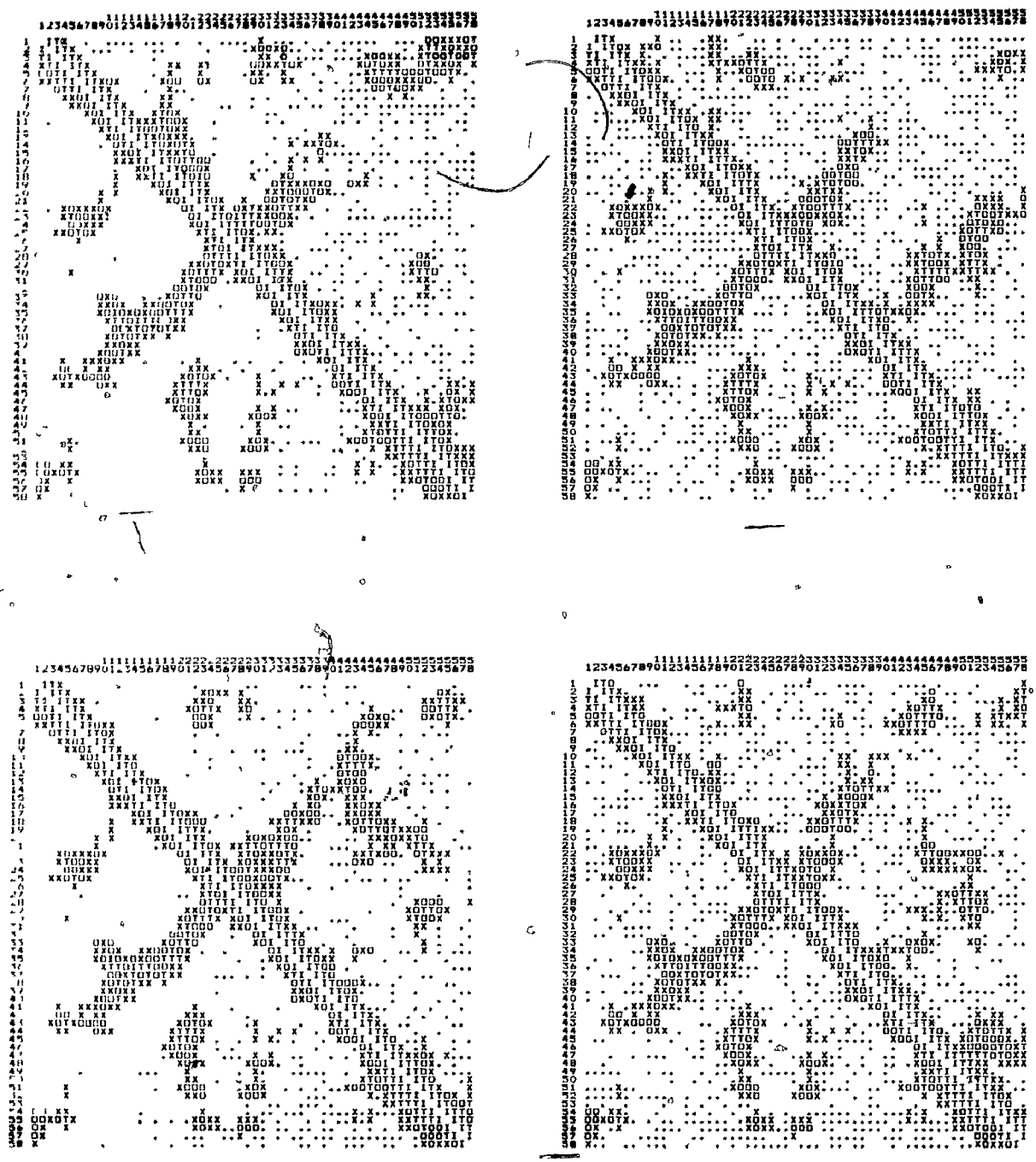
	RMS_y of Optimized Structure vs:		$RMS_y(4)$ of Optimized Structure vs:	
	Real	Others	Real	Others
A	4.21	4.12 ± 0.43	1.63	1.16 ± 0.14
B	4.84	3.93 ± 0.44	1.90	1.13 ± 0.12
C	4.95	4.23 ± 0.36	1.74	1.03 ± 0.08
D	3.94	3.82 ± 0.22	1.73	1.03 ± 0.07
E	4.29	3.88 ± 0.42	1.85	1.18 ± 0.14
ABCDE (ave.)	4.45 ± 0.43	4.00 ± 0.38	1.77 ± 0.11	1.11 ± 0.12

Table 20: Further Comparison of RMS_y and $RMS_y(4)$ Final Structures for BPTI.

The calculated statistics of Tables 18 and 20 indicate that the five optimized structures show an average RMS_y error of 4.45 Å when they are separately compared to the "real" X-ray diffraction structure. When the five optimized structures are compared with each other pairwise, the average RMS_y difference between optimized structures was found to be 4.00 Å. This shows that the structures are converging within a rather small neighbourhood, and that this neighbourhood is close to the configuration of the real structure. Also, from Tables 19 and 20 it is found that the structures are all locally similar from the evidence of their near neighbour $RMS_y(4)$ distance statistics.

Note that for this model, small RMS_y differences between structures also imply structural differences that are small with respect to the usual Cartesian coordinates.

Contact maps for the optimized structures of BPTI from the new initial configurations B, C, D and E are given in Figure 15. It is seen that these optimized structures as a group predict the secondary structures very well, including the long antiparallel β-structure dominating the off-diagonal of the contact maps and the short α-helix near the C-terminal end.



For each of the 4 displayed contact maps, the real structure is shown below the main diagonal of the matrix. Representations for the optimized structures of BPTI, with the three disulfide constraints included in each case, are shown above the main diagonals in the contact maps. The structures are respectively optimized from different random initial configurations: (i) configuration B, (ii) configuration C, (iii) configuration D and (iv) configuration E. In the contact map, the distances $d_{i,i+j}$ between residues i and j (in Å) are letter coded as follows: ***blank*** = 0.0 to 3.0, I = 3.0 to 4.4, T = 4.4 to 6.0, O = 6.0 to 8.0, X = 8.0 to 10.0, and ***period*** = greater than 10.0.

Figure 15: Contact Maps for BPTI. Optimized Structures from Dissimilar Initial Configurations.

It is believed that the *RMS* errors between the optimized and the real structures will improve as better and more accurate parameters are determined either semi-empirically or theoretically. The most obvious improvement would be to replace the second neighbour mean value constraint with a set of far neighbour chemically-derived constraints. These chemically-derived constraints may need to be tailored to the individual protein being folded. It may alternatively be possible to derive universal empirical far neighbour constraints from repeated chemical studies. However, the model can be used in its present form to calculate tertiary structures of globular proteins with a medium level (3 - 6 Å differences in RMS_y) of structural detail, that can subsequently be refined by other algorithms such as those focusing on secondary structure or free-energy minimization.

7.3 Comparison to Previous Distance Constraint Models.

7.3.1 X-ray Diffraction Technique.

The tertiary structures of over 200 crystallized globular proteins are now available from the laboratory technique of X-ray diffraction, although many of these proteins are homologous in structure.

X-ray diffraction works in two stages. In the first stage, the object under examination scatters the X-rays unevenly in all directions, forming a diffraction pattern. In the second stage, this recorded diffraction pattern is mathematically reconstituted into the image. It is necessary to apply X-rays (or electrons or neutrons) to molecular studies instead of visible light because the radiation must possess a wavelength small enough (1 - 2 Å) to produce the required resolution for an object this small.

In protein studies, individual molecules do not provide the necessary contrast for

a discernible diffraction pattern. Therefore, a large number of the molecules ($\approx 10^{15}$) are consolidated into a crystal, which serves to amplify the diffraction pattern. Further, in order to make possible the second stage of reconstructing the original protein, isomorphous replacement of some elements in the crystal lattice by heavy atoms must be performed. The crystallization of protein molecules is not a complete process; proteins rarely form crystals that are regular enough to diffract to a resolution better than 1.5 Å.

Reconstruction of the detailed structure of the protein from the diffraction pattern is achieved by Fourier transform analysis. The clarity of the reconstructed image, the "electron density map", depends upon the accuracy and resolution of the data, and the degree of order of the protein crystal. Individual atoms can only be detected if the resolution is better than 1.5 Å; with resolution better than 3.0 Å, peptide groups and the general shape of sidechains can usually be distinguished; with lower resolution, only gross features such as regular secondary structures can be identified. Portions of the protein that are mobile or can adopt several alternative conformations may be unidentifiable from the electron density map.

The primary structure of the protein must be known *a priori* in order to interpret the electron density map. The actual interpretation in terms of atom locations is rather subjective, and proceeds from an idealized model of the protein, containing standard bond lengths and angles. The mathematical refinement stage consists of finding a best fit between the idealized model and the experimental diffraction data, again using Fourier difference maps. This analysis attempts to extend the final resolution of the structure beyond the initial experimental resolution of the diffraction pattern.

The crystallization and diffraction stage of the X-ray diffraction procedure can be

very costly in terms of human time. It can literally require years of laboratory work to obtain the crystal structure of a single globular protein. Frequently, groups of proteins with similar primary sequences and functions are studied by X-ray diffraction researchers in order to expedite the procedure by employing similar methods on the homologous structures.

The refinement stage can be very costly in terms of computer time. For example, Deisenhofer and Steigemann [32] required over 60 hours of computer time on a Siemens 4004/150 computer to refine BPTI from 2.5 Å resolution to 1.5 Å, using a real-space model building procedure and five Fourier difference maps.

The X-ray diffraction technique has certain drawbacks. It models the crystal structure of a molecule, which in some cases is significantly different from the native structure. Also, a bias is introduced into the overall analysis of globular proteins since only protein structures that make good crystals are resolved. The technique is also extremely expensive and time consuming. However, it is thought to produce a good average of a protein's *in vivo* structures in most cases, and presently produces a far better resolution of the tertiary structure of proteins than any other experimental or theoretical prediction method.

7.3.2 Goel, Yčas *et al.*

In the approach of Goel, Yčas *et al.* [14,45,46,98,128], the constituent residues of a protein are represented by the locations of their C_α-atoms connected by virtual bonds. Their approach attempts to satisfy a set of distance constraints identically by optimizing a weighted penalty function of the constraints. Various constraint combinations are presented to be satisfied exactly, with the constraints being either fixed distances

between points, minimum or maximum bounds on distances between points, or "set average" constraints. The set averages are weak constraints requiring a set of distances to attain a specified average value, with no restrictions on individual pairwise distances. The penalty function, to be minimized, is written as:

$$F = w_1F_1 + w_2F_2 + w_3F_3 + w_4F_4 + w_SF_S + w_{cp}(F_{cp1} + F_{cp2} + F_{cp3}), \tag{22}$$

where

$$F_1 = \sum_{i=1}^{n-1}(\bar{d}_1 - d_{i,i+1})^2 \tag{23}$$

$$F_2 = \left[\sum_{i=1}^{n-2}(\bar{d}_2 - d_{i,i+2})\right]^2 \tag{24}$$

$$F_3 = \left[\sum_{i=1}^{n-3}(\bar{d}_3 - d_{i,i+3})\right]^2 \tag{25}$$

$$F_4 = \left[\sum_{i=1}^{n-4}(\bar{d}_4 - d_{i,i+4})\right]^2 \tag{26}$$

$$F_S = \sum_{(l,k)\in S}(\bar{d}_S - d_{lk})^2 \tag{27}$$

$$F_{cp1} = \left[\sum_{i\in J_1}(\bar{d}_{J_1,cp} - d_{i,cp})\right]^2 \tag{28}$$

$$F_{cp2} = \left[\sum_{i\in J_2}(\bar{d}_{J_2,cp} - d_{i,cp})\right]^2 \tag{29}$$

$$F_{cp3} = \left[\sum_{i\in J_3}(\bar{d}_{J_3,cp} - d_{i,cp})\right]^2 \tag{30}$$

The mean values $\bar{d}_1$, $\bar{d}_2$, $\bar{d}_3$, $\bar{d}_4$, $\bar{d}_S$, $\bar{d}_{J_1,cp}$, $\bar{d}_{J_2,cp}$ and $\bar{d}_{J_3,cp}$ are employed as the parameters of the model. They are estimated semi-empirically, using statistical data from the atomic coordinates of a set of twenty-one proteins with structures known from X-ray diffraction studies, supplied by the Protein Data Bank at the Brookhaven National Laboratory [1]. The values $\bar{d}_{J_1,cp}$, $\bar{d}_{J_2,cp}$ and $\bar{d}_{J_3,cp}$ denote the average distances

from the centroidal point of the molecule for the sets of hydrophobic (J_1), hydrophilic (J_2) and ambivalent (J_3) residues, respectively. Otherwise, the notation is identical to that given in Chapter 9.1. The first neighbour (F_1) and disulfide bond (F_S) constraints restrict pairwise distances individually; all other constraints are of the "set average" type. For example, F_2 is a single constraint employed for second neighbour distances, and is minimized (to $F_2 = 0$) when the average distance for the set of all second neighbour residues $\{d_{i,i+2}\}$ is equal to $\bar{d}_2$, with no specific requirements on the individual $d_{i,i+2}$ pairwise distances.

Their model optionally includes set average constraints on the standard deviations of the near neighbour distances. This is done by adding extra terms $w_j G_j$ to F, where the G_j are of the form:

$$G_j = \left[\bar{s}_j - \left(\frac{\sum_{i=1}^{n-j} (\bar{d}_j - d_{i,i+j})^2}{n-j} \right)^{\frac{1}{2}} \right]^2 ; \; j = 2,3,4, \tag{31}$$

where $\bar{s}_j$ is the specified standard deviation for the jth neighbour C_α-atoms.

Also, the near neighbour set average constraints can be modified to include minimum and maximum bounds for the pairwise distances for each residue k, as follows:

$$F'_j = \left[U_j^L(\bar{d}_j - d_{k,k-j}) + U_j^R(\bar{d}_j - d_{k,k+j}) + \sum_{i \neq k, k \pm j} (\bar{d}_j - d_{i,i+j}) \right]^2 ; \tag{32}$$
$$k = 1, \ldots, n; \; j = 2,3,4,$$

where

$$U_j^L = \begin{cases} 1 & L_j \leq d_{k,k-j} \leq U_j \\ M > 1 & \text{otherwise} \end{cases}$$

and

$$U_j^R = \begin{cases} 1 & L_j \leq d_{k,k+j} \leq U_j \\ M > 1 & \text{otherwise.} \end{cases}$$

Here the parameters L_j and U_j are minimum and maximum empirical bounds for the near neighbour distances.

It is also shown [45,98] that the hydrophobicity class constraints may be replaced by an equivalent set of constraints expressing the hydrophobicity rule in terms of average pairwise distances between the residues of the various hydrophobicity classes, such as:

$$F_{J_1} = \left[\sum_{(i \neq j) \in J_1} (\bar{d}_{J_1 J_1} - d_{ij}) \right]^2 . \tag{33}$$

Here, $\bar{d}_{J_1 J_1}$ represents an empirically derived average distance between residues that are both of hydrophobicity class J_1. Similar constraints would be imposed for all pairwise permutations of the classes J_1, J_2 and J_3.

The paper of Cariani and Goel [14] is concerned with the additional information that can be gained from imposing secondary structure conditions on the model.

The various forms of the penalty function in the approach of Goel, Yčas *et al.* [14,45,46,98,128] are solved by minimization in the corresponding Cartesian coordinates. A sequential optimization is performed, either by a method of repeated random direction linesearches [12] or by a steepest descent algorithm (*cf.*, Chapter 10.3). The residues of a protein are selected one at a time in a randomly generated sequence. For each selected residue in turn, the three variables (x_k, y_k, z_k) are optimized under the function F, while keeping the variables for all the other residues fixed. One iteration consists of all residues being chosen for optimization exactly once. Iterations are then repeated as desired. The optimizer has the advantage that the dimensionality of each

subproblem is small and no derivative calculations are required. However, since the variables in the nonlinear penalty function F are not truly separable, the sequential optimization can lead to a poor convergence rate and local minima trapping or cycling.

Goel and his co-workers have used this method to generate predictive structures for several proteins. Their model is quite successful in predicting protein tertiary structures, using the various combinations of terms used as input in F. It seems, however, to be overly dependent on the choice of initial configuration, implying either an ill-constrained system or an ineffective optimizer routine.

Using the methods outlined above, Goel and Yčas [46] folded BPTI using five different optimizing sequences for the residues, using a large perturbation of the real BPTI structure for the starting configuration. They obtained RMS_y errors of 5.39, 5.16, 4.76, 5.45 and 5.79 Å from the X-ray diffraction coordinates for these five trials. These results gave an average of 5.31 ± 0.38 Å from the real structure, and differed by 4.70 ± 0.65 Å when compared with one another pairwise. They also [46] folded BPTI using an identical optimizing sequence of residues, but folded from three different semi-random perturbations of the real structure. For this case, the final configurations differed by 5.49 ± 0.15 Å from the real structure and by 5.93 ± 0.05 Å from each other. The proteins lysozyme and staphylococcal nuclease (146 residues, no disulfide bonds) were also folded in a similar manner, except that all the parameters used in this case were derived specifically from the protein to be folded. The final RMS_y differences from the real structures were 5.63 and 7.12 Å, respectively.

Goel *et al.* [45], using an underconstrained system of "set average" constraints, obtained RMS_y values of 6.46 Å for BPTI and 6.80 Å for parvalbumin (a globular protein containing 108 residues and no disulfide bonds) from a random chain input

configuration, and an RMS_y value of 6.25 Å for BPTI from a perturbed real structure initial configuration. They also obtained RMS_y values of 4.86 Å for BPTI and 5.21 Å for parvalbumin from a similar input by the use of an overconstrained system in which the residues were forced to satisfy distance constraints individually. In all cases their results were found to be dependent upon the choice of initial configuration, indicating the possibility of unsuitably constrained systems. During this study, they also analyzed the model by folding BPTI when given complete exact distance information, complete information with some errors, several forms of complete information as approximate values without errors, and incomplete exact distance information. All of these latter studies included distance information that would generally be inaccessible *a priori*.

7.3.3 Kuntz, Crippen *et al.*

The original model of Kuntz, Crippen, Kollman and Kimelman [61,63] is written as a penalty function consisting of constraints on the pairwise distances between residues. Each amino acid residue is represented by the first atom of its side-chain (the C_β-atom), except for Gly which is represented by its C_α-atom. The $3 \times n$ Cartesian coordinates of the C_β-atoms are chosen as the independent variables to be optimized for a protein of length n. The penalty function, called an "error function", is composed of five terms:

$$F = w_1 F_1 + w_2 F_2 + w_3 F_3 + w_4 F_4 + w_5 F_5, \tag{34}$$

where each of the terms represents a distance constraint set. The weights (w_i) are empirically chosen to reflect the relative importance of the various constraints:

$$w_1 = w_3 = w_4 = w_5 = 100, \; w_2 = 25. \tag{35}$$

The first term of the penalty function (34) is given to be as follows:

$$F_1 = \sum_{i=1}^{n-1} \mid {d_{i,i+1}}^2 - {L_1}^2 \mid, \tag{36}$$

where $d_{i,i+1}$ represents the distance between C_β-atoms of nearest neighbour residues and L_1 is a discontinuous step function:

$$L_1 = \begin{cases} 3.8 \text{ Å} & \text{if } d_{i,i+1} < 3.8 \text{ Å} \\ 4.9 \text{ Å} & \text{if } d_{i,i+1} \geq 4.9 \text{ Å} \\ d_{i,i+1} & \text{otherwise.} \end{cases} \tag{37}$$

This term F_1 restricts the C_β-atom first neighbour distances to lie between a minimum distance of 3.8 Å and a maximum distance of 4.9 Å. Kuntz *et al.* also state that the $d_{i,i+2}$ and $d_{i,i+3}$ distances are constrained to an allowable range between α-helical and extended chain conformations, but do not show these terms explicitly.

The second term of the penalty function represents the interactions between various classes of residues. Pairs of hydrophobic residues are constrained to a pairwise distance of less than 10 Å, as are polar-ionic residue interactions and pairs consisting of residues with appropriately charged ionic sidechains. Ionic-hydrophobic or polar-hydrophobic residue pairs are constrained to lie more than 15 Å from each other. The N- and C-terminal residues are classified as ionic. This penalty term is achieved as follows:

$$F_2 = \sum_{i=1}^{n-1} \sum_{j=1}^{n-i} C_{ij} ({d_{i,i+j}}^2 - {L_2}^2), \tag{38}$$

where C_{ij} is a weighting coefficient for the interaction between residues i and $i+j$. The coefficient C_{ij} has a possible range of -100 to $+25$, with values determined by the types of residues involved. It has positive values when residues i and $i+j$ are ionic-hydrophobic or polar-hydrophobic pairs, negative values for hydrophobic-hydrophobic, polar-ionic or ionic-ionic residue pairs, and is equal to zero otherwise. The step function

L_2 is given by

$$L_2 = \begin{cases} 10 \text{ Å} & \text{when } C_{ij} > 0 \text{ and } d_{i,i+j} > 10 \\ 15 \text{ Å} & \text{when } C_{ij} < 0 \text{ and } d_{i,i+j} < 15 \\ d_{i,i+j} & \text{otherwise.} \end{cases} \tag{39}$$

The third term constrains hydrophilic residues (defined as all residues except Met, Val, Leu, Ile, Phe, Trp and Tyr) to a specified minimum distance from the centroidal point of the molecule:

$$F_3 = \sum_{i \in \text{hydrophilic}} (d_{i,cp}^{*\,2} - d_{i,cp}^{\,2}), \tag{40}$$

for all hydrophilics such that $d_{i,cp} < d^*_{i,cp}$, and zero otherwise. The scalar $d^*_{i,cp}$ is set equal to 10 Å. This parameter, estimated empirically from examination of BPTI and rubredoxin residue distributions, determines the volume of the hydrophobic core.

The fourth term prevents the chain from self-intersecting:

$$F_4 = \sum_{i=1}^{n-2} \sum_{j=2}^{n-i} (V_{i,i+j}^{\,2} - d_{i,i+j}^{\,2}). \tag{41}$$

Here $V_{i,i+j}$ is the sum of the effective van der Waals radii of the sidechains of residues i and $i+j$. Whenever $d_{i,i+j} \geq V_{i,i+j}$, the value of F_4 is set to zero.

The fifth term is a constraint for the Cys residue pairs that are connected by disulfide bonds:

$$F_5 = \sum_{(l,k) \in S} (d_{lk}^{\,2} - L_5^{\,2}), \tag{42}$$

where $L_5 = 6$Å is the desired pairwise distance for C_β-atoms of disulfide bonded Cys residues. Again, this constraint is set to zero whenever $d_{lk} \leq L_5$.

The initial configurations used for this model are extended chains, in which each residue is positioned at 3.8 Å from the preceding residue, the steps being made alternately along the x, y and z axes. The penalty function is minimized in Cartesian coordinates by use of a steepest descent algorithm *without* linesearch. Only one step

is attempted along the direction of steepest descent per iteration, with the step size being determined by the relative success of the previous iteration.

Using this model, Kuntz *et al.* [63] computed optimized structures for BPTI and for rubredoxin (using 53 residues). They reported that their *best* results for BPTI had RMS_y errors of 4.70 to 5.0 Å. They obtained a structure for rubredoxin with an RMS_y error of 4.7 Å when the same weight and parameter values as those of BPTI were used. Their *best* RMS_y result for rubredoxin was reported to be a value of 3.99 Å. When the Fe-S interaction distances were not included in rubredoxin, the prediction accuracy dropped to an RMS_y error of approximately 6 Å.

They generated "a number" of structures for BPTI and rubredoxin. The resulting RMS_y errors from the X-ray diffraction structure were in the range of 4.7 to 6.5 Å for BPTI when the three disulfide bonds were given correct distances. The RMS_y errors for rubredoxin were found to be 4.0 to 6.0 Å, given correct distances between the four Cys residues involved in the Fe-S_4 complex.

In their more recent approaches, Kuntz *et al.* [30,50,51,52,62] first impose a set of distance constraints directly on the matrix consisting of all pairwise distances between the C_α-atoms of the residues, where specified entries in this distance matrix are limited to be within upper and lower bounds. Therefore, the model works directly with a geometry of pairwise distances.

Distance constraints between residue pairs are incorporated as entries in upper and lower bound distance matrices, denoted by U and L, respectively. First neighbour residue distances are given corresponding values of $u_{i,i+1} = 3.80$ Å and $l_{i,i+1} = 3.80$ Å in the first diagonals of U and L, respectively. Values for the elements in the second diagonals (second neighbour distances) are chosen to be $u_{i,i+2} = 7.30$ Å and

$l_{i,i+2} = 6.00$ Å. Disulfide bond locations are assumed to be known, with pairs of residues l and k connected by these bonds being given the upper bound restrictions of $u_{l,k} = 6.50$ Å. All other elements of U are then set to a reasonable absolute upper bound distance (given by 40.0 Å in [50] and 38.0 Å in [62]), and all other elements of L are set to an absolute lower bound distance (5.0 Å in [50] and 6.0 Å in [62]). The resulting distance matrices U and L are then smoothed by satisfying triangle and reverse triangle inequalities for all residue triplets (*cf.*, Chapter 4). The boundary matrices thus obtained are representations of a nonintersecting ideal chain containing acceptable virtual bond angles and disulfide bonding.

The remaining pairwise distances are further restricted by consideration of secondary structure algorithms for α-helices, β-strands or hairpin turn contacts, or by prediction of hydrophobic contacts. In Havel *et al.* [51], constraints consisting of a tetrangle inequality, and pentangle and hexangle equalities are included (*cf.*, Chapter 4). These arise from the distance geometry itself, and ensure that a given distance matrix will correspond to a three-dimensional Euclidean structure.

The model is easily solved with respect to the distance coordinate system by simply assigning values in the distance matrices. The difficulty arises when the nonlinear transformation is made from distance space, a space of higher than three dimensions in general, into R^3. Either the distance bounds must initially be chosen carefully to limit the optimized configurations to R^3 structures or a supplementary process must be devised to embed the distance matrix configurations into R^3.

There is no obvious way to perform this embedding process optimally (*cf.*, Chapter 4), and the system behaves essentially as an overconstrained one. The most difficult step of this approach is to decide in some rigorous fashion which distance constraints

to relax during the embedding process, whether the embedding occurs during or after the optimization step. In spite of these obstacles, the method shows very promising results in the prediction of tertiary structure, and is improving as the properties of the transformation become more familiar.

Upon repeatedly optimizing the tertiary structure of BPTI, Kuntz *et al.* [62] reported RMS_y values of 5.46 ± 0.28 Å, employing the method of steepest descent combined with the conjugate gradient method. These results corresponded to RMS_x results of 6.59 ± 0.60 Å. Using a Monte Carlo procedure, similar results of 5.48 ± 0.28 Å were obtained for RMS_y. However, it is mentioned that the *best* results were in the range 3.75 - 4.25 Å for RMS_y errors and 4.8 - 5.2 Å for RMS_x errors. In this study, not only were the structures repeatedly generated, but the RMS_x errors were calculated as well.

7.3.4 Wako and Scheraga.

The model of Wako and Scheraga [100,117,118,120] consists of the application of successive approximations proceeding from a short-range distance constraint algorithm to subsequent incorporation of medium- and long-range interactions, followed by energy minimization of the entire molecule.

For this model, the mean distances $\bar{d}_j$ for all residues separated in primary structure by j residues are determined, with weights w_j determined from their standard deviations. The values $\bar{d}_j$ and w_j are determined from the primary sequence by taking into account differences between short-, medium- and long-range effects (where the ranges are defined to be $j \leq 8$, $9 \leq j \leq 20$ and $j \geq 21$, respectively), and also by considering the neighbouring residue types.

For near neighbour distances, the values of $\overline{d}_j$ and w_j are adopted from the empirical analysis of known protein structures. The penalty function employs empirically determined mean values $\overline{d}_j$, weighted according to their standard deviations:

$$w_j = \frac{1}{s_j^2}. \tag{43}$$

The near neighbour distances also use secondary structure prediction algorithms.

For medium and far neighbour distances, $\overline{d}_j$ and w_j are mainly determined by hydrophobicity and hydrophilicity indices, using the scales of Meirovitch *et al.* [74,75,76]. These evaluations additionally employ empirical mean and standard deviation results, which are supplied in linear regression form.

Other factors, such as disulfide bonds or interactions between nonbonding Cys residue pairs or between specific Cys and aromatic residues also contribute. Exact distances for specified pairs of residues are also included if this data is considered to be obtainable from experimental techniques.

All of the above effects are incorporated by appropriate choices of $\overline{d}_j$ and w_j, which vary according to the effects being considered.

The protein is represented by the coordinate locations of its C_α-atoms, connected by virtual bonds. The coordinates are varied to minimize a function of the form:

$$F = \sum_{i=1}^{n-1} \sum_{j=1}^{n-i} w_j (\overline{d}_j - d_{i,i+j})^2. \tag{44}$$

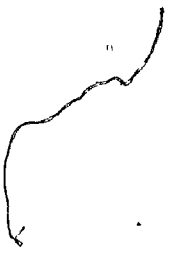
Possible chain self-intersection is not accounted for explicitly in the model. In actual practice, unfavorable far neighbour contacts were encountered when a two-dimensional representation for the protein was used [120], but no chain entanglement was encountered on folding in three dimensions [117,118].

For the optimization of the structure of BPTI, several choices for the initial configuration were used. The initial configurations generally were close to that of the real structure. For example, the optimization of one structure reduced the RMS_y deviation from 4.25 to 2.24 Å. For this particular case, the general model was supplemented with exact distance information on some residue pairs far apart (Arg1 - Lys15, 29.80 Å; Lys15 - Ala58, 35.12 Å) and some pairs close together (Leu6 - Ala25, 5.78 Å; Tyr10 - Asn43, 7.99 Å; Asn24 - Asn43, 9.26 Å) with respect to the tertiary structure.

Optimization of the penalty function was carried out in Cartesian space. First a gradient minimizing routine was used. This routine tended to become trapped in local minima as the structure became more compact. Therefore, the minimizing procedure was changed to a Monte Carlo method at the stage when the excluded volume effect became evident. The Monte Carlo method was local, in that it optimized the position of each residue sequentially along the chain, taking into account only the distances between near neighbour residues.

Wako and Scheraga [118] obtained tertiary structure results by folding BPTI from several initial conformations, using various constraint combinations. When only primary structure obtainable constraints plus the locations of the three disulfide bonds were used as constraints, the following RMS_y and RMS_x errors from the X-ray diffraction structure were obtained for the folded structures: $RMS_y = 4.83$ Å and $RMS_x = 7.77$ Å from an initial conformation with an error of $RMS_y = 5.90$ Å, $RMS_y = 4.30$ and $RMS_x = 5.98$ Å from an initial $RMS_y = 21.27$ Å, $RMS_y = 4.83$ and $RMS_x = 9.09$ Å from an initial $RMS_y = 8.10$ Å, $RMS_y = 4.43$ and $RMS_x = 9.84$ Å from an initial $RMS_y = 5.80$ Å and an $RMS_y = 4.10$ and $RMS_x = 5.88$ Å from an initial $RMS_y = 4.25$ Å. Aggregate statistics on these results would not be meaningful since

many of the trial results were folded from conformations similar to that of the real structure.



8 Discussion.

Distance constraint models are found to be valuable as reliable predictors of protein tertiary structure. The accuracy of resolution attainable by this type of model alone is indicated to be in the range of 3 - 5 Å error. This is generally considered to be of "medium" resolution, intermediate between experimental X-ray crystallography results and the results obtained from other types of theoretical models. All that is required for input into distance constraint models is the primary structure of a protein. However, the resolution can generally be improved by the inclusion of constraints representing disulfide bond locations or other specific information concerning distances between residues that are far apart in the primary structure.

Distance constraint models are a relatively quick and inexpensive method of obtaining medium resolution predictions for tertiary structures. For this reason, they may be implemented as first approximations to actual tertiary structures, which could be refined by laboratory techniques such as X-ray studies. Improvements in the resolution of theoretical predictor models may be attained by repeated cycling between distance constraint models, secondary structure predictor algorithms and methods of free-energy minimization.

8.1 Improvements for the Present Model.

Improvements for the model fall into two general categories: parameter resolution and computation.

The nonlinear optimization algorithm may possibly be improved by employing new optimization algorithms currently being developed which can utilize quasi-Newton methods (which usually perform better than Newton methods) for large-scale problems

such as these. Other algorithms can be explored which can better exploit the special structure of distance constraint models by operating in another Euclidean metric, such as those employing an L_1 or L_∞ norm (*cf.*, Coleman and Conn [24]). These types of algorithms are relatively new, and as yet have not been adequately developed to handle large scale systems.

The time required for execution of the algorithm can be lessened by use of initial configurations that include correct first neighbour distances and other constraints. Alternately, initial configurations can be quickly generated by use of low resolution algorithms such as those outlined in Section 8.2.

Marginal improvements to the parameter values will come as the database of known tertiary structures grows and is analyzed statistically. The parameters can certainly improve if new chemical or other techniques for obtaining short distance or long distance contacts in R^3 can be found for far neighbour residues in the molecule. Possible parameters of this type that may be accessible from empirical studies alone include interaction distances between Cys and aromatic residues [117,118], interaction distances between pairs of aromatic residues [13] or between pairs of nonbonding Cys residues [117,118], or *a priori* assignment of centroidal point distances to the hydrophobic residues [72,84].

The model in its present form can calculate tertiary structures for much larger molecules. On the other hand, it can also be used to predict structures for peptide hormones, neurotransmitters and other polypeptides that are fragments of protein precursor molecules. With some knowledge of specific intramolecular bonding, the model can be expanded to calculate structures of broken chain and multiple-strand proteins.

8.2 Alternative Algorithms for Solution.

There are several methods other than nonlinear optimization by which the tertiary structure prediction problem may be approached. The advantage of these methods is the speed in obtaining structures. The disadvantage is that the structures obtained are of low resolution (RMS_y errors in the vicinity of 6 Å) in comparison with distance geometry models (RMS_y errors approximately 4 Å).

Some possible approaches that sidestep the difficulties of nonlinear optimization include:

1. linearization of the objective function and all constraints with respect to the Cartesian coordinates;

2. reduction of the problem to a one-dimensional system by fixing two of the distance coordinate locations for each residue. For example, the first neighbour and second neighbour distances can be specified initially for each residue, making the radial distance from the centroidal point then calculable analytically from the hydrophobicity conditions;

3. optimization of the residues on a three-dimensional integer lattice ("packing the snake"). The remainder of the subsection is devoted to a brief outline of an implementation of this approach.

The "packing the snake" algorithm (M. Yčas, personal correspondence) exploits an important geometrical characteristic of tertiary structure, namely the high and effectively constant packing density of residues in a protein [94]. The high density of the packing of the residues varies little from protein to protein. Since this packing

density is difficult to achieve, it greatly limits the allowable conformations. The volume occupied by sidechains is also restrictive. It is specific for each sidechain and is observed to vary by only about 5%. Therefore, the tight folding of the protein under the excluded volume restrictions of the sidechains can be seen to be analogous to a snake that is tightly coiled.

Residues that are far apart with respect to primary sequence also tend to be far apart physically in R^3. This principle has been observed in all globular proteins containing clearly detectable domains [101]. Conversely expressed, the domain structure shows a high degree of "neighbourhood correlation", with the distance along the chain and three-dimensional distance exhibiting a positive correlation. This observed neighbourhood correlation is probably a consequence of the chain folding process. The correlation suggests that a folding chain is analogous to a string that is held at one end and allowed to fall down. The resulting coil is not random but shows neighbourhood correlation. It does not become entangled and it can be easily unravelled by picking up the end. This concept is confirmed by the absence of "knots" in all protein structures known thus far, the term knot being used in the everyday sense and not in the mathematical sense. This principle can also be utilized in the packing algorithm.

Initial attempts have been made by the author to implement this algorithm on a three-dimensional cubic lattice. The positions of the C_α-atoms of a protein are optimized on the vertices of the lattice. Each vertex is either occupied or unoccupied by a residue. First neighbour residues are denoted by occupied adjacent lattice vertices. For all other residue pairs, there is a volume exclusion rule prohibiting them from occupying identical or adjacent vertices. For example, this results in second neighbour residues that are either 2.0 C_α-atom units apart (probability = 0.20) or $\sqrt{2}$ units apart

(probability = 0.80). For residue pairs connected by disulfide bonds, the pairwise distance is restricted to a range of 1.0 to $\sqrt{3}$ units. The existence of disulfide bonds in a protein can facilitate the folding process by operating as nuclei for the overall structure. The centroidal point hydrophobicity conditions are included in the form of an objective function to be minimized, under the previously mentioned constraints.

This method was tested by predicting the tertiary structure for rubredoxin (54 residues, no disulfide bonds), using a modified tree-search integer programming optimizer. The optimized structure contained an RMS_y error of 5.37 Å from the X-ray diffraction structure, with $RMS_y(4) = 2.05$. The optimization process required less than 97 CPU seconds on a CYBER 170-730 mainframe computer. Overall, the mean difference in centroidal point distance between the real and optimized residues was only 2.14 ± 1.68 Å. The mean values for the first, second, third and fourth neighbour distances for the optimized structure were calculated to be 3.80 ± 0, 5.18 ± 0.89, 7.41 ± 1.10 and 8.61 ± 1.92 Å, respectively. These statistics compare well to those of the real structure, given in Table 9 of Chapter 7.1.

Present work includes modifying the above problem from one of integer programming to one of pattern generation. This change should result in a great saving of execution time for the algorithm in predicting structures of large proteins.

8.3 Implications for Future Study.

Distance constraint models certainly are not as yet optimal with respect to the constraint set imposed or to the mathematical form of the problem. All distance constraint models must conform to certain conditions relating to their formulation in distance geometry coordinates and to their inverse mapping into Cartesian coordinates, as discussed in

Chapter 4. In fact, one could envision a Hilbert space of distance constraint functions. The problem would then be to define a set of conditions in order to decide on a function that is "optimal". The work of Chapter 4 is a step toward categorizing the space of distance constraint functions.

Effects of substitutions, deletions or additions of residues to proteins can be easily evaluated by the model. In the present model for example, the protein lysozyme was folded with the *C*-terminal residue omitted from the chain (*cf.*, Chapter 7). It was found that this single deletion significantly affected the final structure obtained.

With improvement in the accuracy of theoretical models, the important relationship between the structure and function of proteins can be explored in general for the first time.

Theoretical aspects regarding evolution versus tertiary structures of proteins can be evaluated via the model. This is important for the taxonomy of proteins, in order to find evolutionary related proteins, which may perform far different functions at the present time. This research is also important theoretically, in the investigation of the relationship between the evolutionary rapidly substituting sections of a protein and their corresponding structures in DNA [6].

With a reliable distance constraint model now available for the prediction of tertiary structures using primary structures as the sole input, it is possible to generate an atlas of tertiary structures from the existent primary structure tables. There are thousands of proteins for which primary structures are known but tertiary structures are unknown. Once a catalogue of tertiary structures is generated, the general topologies of folded structures can be investigated. Also, the catalogue of folded structures may be used to explore the entire class of primary structures capable of folding into stable tertiary

structures, as opposed to the unstable and random coil sequences of polypeptides.

Perhaps the most important benefit from the cataloguing of protein tertiary structures concerns the inverse problem. This involves determining the set of primary structures that will fold into a given tertiary structure. As a corollary, it gives the possibility of inexpensively constructing an artificial protein with certain desired properties that are present in a natural protein. Once the classes of possible folding chains are known, the entire set of primary sequences that can fold to obtain a specified active site may be determined. This has obvious possibilities in the fields of agriculture, pharmacology and medicine.

9 Appendix: The Mathematical Model.

In this Appendix are given the details of the mathematical formulation of the distance geometry model for tertiary structure prediction. The mathematical representation comprises a set of restrictions on the Euclidean distances between the amino acid residues.

In Section 9.1, the mathematical notation for the parameters of the model is presented. Section 9.2 includes the distance coordinate description of the mathematical model and its conversion into a penalty function. The mathematical model is designed to specify the geometrical characteristics of globular proteins while remaining tractable. Optimization of the penalty function is performed in the space of Cartesian coordinates by techniques of nonlinear programming described in Chapter 10. The final sections of this Appendix are devoted to derivations of the numerical scalar values or expressions for the parameters used in the present model. A summary of the parameter values used in the model is given in Table 26 at the end of this Appendix.

9.1 Parameter Notation.

A protein will be represented by the locations of the C_α-atoms of its constituent residues. This simplified representation is one of the benefits of distance constraint models; these models can elicit the complex underlying energetic interactions of the atoms of the protein through relatively simple geometric characteristics. Since proteins are non-branching chain molecules, the residues can be numbered sequentially from 1 (the N-terminus) to n (the C-terminus).

Let

$S = \{ (l, k) \mid$ residue l and residue k are connected by a disulfide bond$\}$.

The following notation will be used in the description of the parameters of the model:

$\overline{d}_1$ = mean distance between residues i and $i+1$

$\overline{d}_2$ = mean distance between residues i and $i+2$

$\overline{d}_j$ = mean distance between residues i and $i+j$

$\overline{d}_S$ = mean distance between residue pairs forming a disulfide bond

L_j = minimum distance between jth neighbour residues for small j, (usually $j = \{3,4\}$)

U_j = maximum distance between jth neighbour residues for small j

L_N = minimum distance between jth neighbour residues for large j, (usually $j > 4$)

U_N = maximum distance between jth neighbour residues for large j.

In the model, the values of the parameters defined above will not be implemented as statistics corresponding to the specific protein being folded, but instead will be given as representative mean values and minimum and maximum bounds for the set of all globular proteins. Note that there are distinct values for the near neighbour minimum distances L_j and for the near neighbour maximum distances U_j for each different amount (j) of residue separation in primary sequence, whereas the far neighbour distance bounds L_N and U_N are independent of j.

For the C_α-atom of each residue, three variables x_i, y_i and z_i are introduced which represent its Cartesian coordinates in R^3. Let $d_{i,i+j}$ represent the Euclidean distance between residues i and $i+j$:

$$d_{i,i+j} = \left[(x_i - x_{i+j})^2 + (y_i - y_{i+j})^2 + (z_i - z_{i+j})^2\right]^{\frac{1}{2}}. \tag{45}$$

In the model, it is convenient for the centroidal point of the protein, defined by equation (1) of Chapter 3.3, to be situated at the origin with respect to a Cartesian coordinate frame. This is equivalent to placing a virtual residue at the centroidal point of the molecule, and assigning it the Cartesian coordinate representation of $(x_{cp}, y_{cp}, z_{cp}) = (0,0,0)$. Let $d_{i,cp}$ denote the distance between residue i and the centroidal point of the protein. It follows from (45) that the Euclidean distance between the centroidal point of the protein and any residue with coordinates (x_i, y_i, z_i) is given by:

$$d_{i,cp} = \left({x_i}^2 + {y_i}^2 + {z_i}^2\right)^{\frac{1}{2}}. \tag{46}$$

As explained in Chapter 3.3, the hydrophobicity rule describes the tendencies for each amino acid residue type to be situated in the interior or on the outer surface of a globular protein. The following notation will be used to designate the three hydrophobicity classes of the present model:

J_1 — hydrophobic — residue tends toward the centroidal point of the configuration,

J_2 — hydrophilic — residue tends toward the surface of the configuration,

J_3 — ambivalent — residue has no tendency.

For the model, the residues are divided into the three hydrophobicity classes as follows: hydrophobics {Val, Leu, Ile, Phe, Met}, hydrophilics {Arg, Asp, Glu, Gln, Gly, Lys,

Pro} and ambivalent {Ala, Asn, Cys, His, Ser, Thr, Trp, Tyr}. Chapter 3.3 explains the basis for categorizing the residue types into separate hydrophobicity classes and the reasons for selecting this particular classification for the present model.

In the model, an "ideal" position for all residues of the hydrophobic class is given to be at the centroidal point of the molecule. The parameter D is used to denote an ideal distance from the centroidal point for the individual hydrophilic residues. The value of D is derived in Chapter 9.3 by requiring the average distance from the centroidal point for all hydrophilic residues and hydrophobic residues of a protein to be equal to a specified semi-empirical value. The ideal positions from the centroidal point for the hydrophobic and hydrophilic residues are not strictly realized in the optimal structure of a protein folded by the model. These parameters are only used to evoke the empirical trends found for the locations of the residues.

9.2 The Nonlinear Programming Formulation.

A specific protein is required to conform to a set of average geometrical characteristics found from the class of all globular proteins. This is accomplished by constructing a mathematical model in the form of a nonlinear programming problem, as follows:

$$\text{Minimize } g = \sum_{i \in J_1} {d_{i,cp}}^2 + \sum_{i \in J_2} (D - d_{i,cp})^2 \tag{47}$$

(the hydrophobicity restrictions),

subject to:

$$\sum_{i=1}^{n} x_i = 0, \sum_{i=1}^{n} y_i = 0, \sum_{i=1}^{n} z_i = 0 \tag{48}$$

(centroidal point constraints, forcing the centroidal point to $(x_{cp}, y_{cp}, z_{cp}) = (0,0,0)$)

$$d_{i,i+j} = \bar{d}_j;\ i = 1, \ldots, n-j \tag{49}$$

(jth nearest neighbour constraints, usually $j = \{1, 2\}$)

$$d_{lk} = \overline{d}_S;\ (l,k) \in S \tag{50}$$

(disulfide bond constraints)

$$L_j \leq d_{i,i+j} \leq U_j;\ i = 1, \ldots, n-j \tag{51}$$

(minimum and maximum distance constraints for near neighbours: usually $j = \{3, 4\}$)

$$L_N \leq d_{i,i+j} \leq U_N;\ i = 1, \ldots, n-j \tag{52}$$

(minimum and maximum distance constraints for far neighbours: usually $j > 4$).

Thus, given that the constraints are to be satisfied exactly, the optimum value of a hydrophobicity measurement, the ***objective function***, is sought. The residues classified as hydrophobic are required by the objective function to tend toward the molecule's centroidal point and the residues classified as hydrophilic are required to tend toward the surface of a sphere of radius D centered at the centroidal point. The residues in the ambivalent category are considered to have no preference with respect to the inside or the outside of the molecule; therefore, they do not contribute a set of terms in the objective function. The sphere itself has no physical significance, as the radius (D) of the sphere is chosen simply to reflect the empirical packing density of the residues.

The hydrophobicity rule is presented as the objective function of the model and not as part of the constraint set because the residue types are empirically found to only exhibit tendencies for the inside hydrophobic or outside hydrophilic environments. The residue types are not observed to be situated at any specified distance from the molecule's center in general.

The centroidal point constraint is satisfied if and only if the centroidal point of the protein is situated at the origin with respect to the Cartesian coordinate representation of the residue locations. Since the model is eventually solved as a penalty function in Cartesian space, this constraint has the purpose of maintaining a sparse form for the Hessian matrix of second derivatives during the optimization process (Chapter 10).

The jth nearest neighbour constraints are satisfied if and only if for each pair of residues i and $i+j$, the distance between residue i and residue $i+j$ equals $\bar{d}_j$, the expected average distance between jth nearest neighbours. The minimum and maximum distance constraints for near neighbours are satisfied if and only if for each pair of residues i and $i+j$ separated by a specified number of positions (j) in primary sequence, the distance between residue i and residue $i+j$ is between the lower bound L_j and the upper bound U_j, the expected extrema for jth nearest neighbours. The minimum and maximum distance constraints for far neighbours are satisfied if and only if for each pair of residues i and $i+j$ for any "large" j (*i.e.*, $j > 4$) the distance between residue i and residue $i+j$ is between L_N and U_N, the absolute lower and upper bounds applicable to all far neighbour distances.

Although the nonlinear programming model as given by equations (47) - (52) is naturally expressed in terms of distance space coordinates $\{d_{i,i+j}\}$, it is more efficient to solve the model in the space of Cartesian coordinates (*cf.*, Chapter 4.1). Therefore, each distance $d_{i,i+j}$ in the model is transformed into its corresponding Cartesian coordinates by equation (45) before beginning any optimization steps.

The model is solved by transforming the constrained nonlinear optimization problem into a series of unconstrained problems via a penalty function approach. A

quadratic loss penalty function is used in the present model, and is given as follows:

$$\text{Minimize } p(x,\mu) = \mu g \tag{53}$$

$$+\left(\sum_{i=1}^{n} x_i\right)^2 + \left(\sum_{i=1}^{n} y_i\right)^2 + \left(\sum_{i=1}^{n} z_i\right)^2 \Bigg\{ \text{centroidal point penalties}$$

$$+\sum_{j=1}^{2}\sum_{i=1}^{n-j}(\bar{d}_j - d_{i,i+j})^2 \Bigg\{ j\text{th nearest neighbour penalties}$$

$$+\sum_{(l,k)\in S}(\bar{d}_S - d_{lk})^2 \Bigg\{ \text{disulfide bond penalties}$$

$$+\sum_{j=3}^{4}\sum_{i=1}^{n-j}[\min(0, U_j - d_{i,i+j})]^2 \Bigg\{ \text{near neighbour maximum distance penalties}$$

$$+\sum_{j>4}\sum_{i=1}^{n-j}[\min(0, U_N - d_{i,i+j})]^2 \Bigg\{ \text{far neighbour maximum distance penalties}$$

$$+\sum_{j=3}^{4}\sum_{i=1}^{n-j}[\min(0, d_{i,i+j} - L_j)]^2 \Bigg\{ \text{near neighbour minimum distance penalties}$$

$$+\sum_{j>4}\sum_{i=1}^{n-j}[\min(0, d_{i,i+j} - L_N)]^2 \Bigg\{ \text{far neighbour minimum distance penalties.}$$

As with most other penalty methods, the overall solution for the original unconstrained problem, given by equations (47) - (52), is found by alternately minimizing $p(x,\mu)$ for a fixed value of the scaling parameter μ and reducing μ. In practice about four calls to the Inner Loop algorithm are executed by Outer Loop to perform this repeated minimization (*cf.*, Chapter 10.7), reducing μ to a tenth of its former value each time. As $\mu \to 0$, the local minimizers of $p(x,\mu)$ that are sufficiently close to satisfying the constraints will approach local solutions of the constrained nonlinear programming problem.

A main reason for choosing this penalty function approach is that the Hessian (*i.e.*, the matrix of second order mixed partial derivatives) of $p(x,\mu)$ has a block structure

which is sparse. The sparsity of the Hessian is ensured in this case by requiring the Cartesian coordinate representation to have its origin situated at the centroidal point of the protein. Without sparsity in the Hessian, the practical usefulness of any proposed algorithm would be severely restricted because of the data storage limitations of computers.

Quadratic loss penalties are chosen for the present model, as opposed to quartic loss penalties used by some other distance constraint models [62,63,119]. Quartic loss penalties may be written in the following general form:

$$\sum\sum(\bar{d}_j^2 - d_{i,i+j}^2)^2. \tag{54}$$

Quadratic loss penalties are found to result in a smoother penalty function; quartic loss penalties cause the resulting penalty function to possess sharper contours and stronger local minima in the vicinity of a solution, which can hinder or trap an optimization procedure.

9.3 The Values of the Parameters.

The numerical values of the parameters used in the model are presented in this section. The reference sources or required derivations for the parameters are also shown here. For a general exposition of the parameters used in distance constraint modelling, the reader is referred to Chapter 3.

9.3.1 Near Neighbour Parameters.

Possible near neighbour distance parameters for use in distance constraint models are shown in Table 21, including the relevant parameters used in the present model.

In the present model, the near neighbour parameters include mean distances between first neighbour and second neighbour residues, and extrema bounds for third and fourth neighbour residues. Note that for the first neighbour distances, the mean, the maximum and the minimum are effectively equal. For second neighbours, the use of only maximum and minimum bounds may be a more logical choice for parameters, but the mean distance is used because it imposes a much stronger constraint on the degrees of freedom of the system, allowing the system to be "suitably constrained" (*cf.*, Chapter 4). This condition may be relaxed when additional strong conditions on the residues can be imposed, such as the acquisition of chemically-derived hydrophobicity conditions. Parameters for the mean third neighbour and fourth neighbour distances are not used. Instead, the distances between third and fourth neighbour residues are constrained to be between specified minimum and maximum bounds.

Distance	Mean Value	Standard Deviation	Minimum Bound	Maximum Bound
First neighbour	3.80[a]		3.80[a]	3.80[a]
Second neighbour	5.95[b]	0.63[b]	4.7[b]	7.1[c]
Third neighbour	7.24[b]	1.82[b]	4.5[b]	10.7[c]
Fourth neighbour	8.77[b]	2.44[b]	4.5[b]	13.9[c]

Statistical information for the distances (in Å) between near neighbour C_α-atoms. The sources used for the Table are: (*a*) Pauling *et al.* [81], (*b*) Goel and Yčas [46], (*c*) Chapter 11.4.

Table 21: Near Neighbour Parameters for Distance Constraint Models.

As indicated in Table 21, the first neighbour distance parameter originates from the studies of Linus Pauling and his group [25,81], who were the first to study the details of polypeptide structures.

The values for the near neighbour maximum bounds are obtained theoretically in Chapter 11. Starting with a set of standard bond angles and lengths [89] and restrictions on the bond rotations [90,91,92,99], strong upper bound estimates for these maximum distances are computed. These upper bounds are used as parameters in the present model. Estimates for near neighbour mean values and minimum bounds are also calculated theoretically in Chapter 11. However, the lower bound estimates calculated for the minimum distances were found to be too weak to be employed as parameters in the model. The theoretical mean value parameters for near neighbour distances were reasoned to be largely dependent upon the types and proportions of secondary structures present within the individual proteins. It was decided to defer the implementation of these mean value parameters until more reliable estimators are developed for the proportions or primary sequence locations of the secondary structures.

For the present model, mean value parameters are employed for the first neighbour and second neighbour distances. The numerical values for these parameters are obtained from the empirical results of Goel and Yčas [46], who calculated mean distances, minimum values and maximum values for the first to fourth neighbour C_α-atoms in a set of twenty globular proteins.

The model also employs empirical results for near neighbour minimum distances. Using the identical set of twenty proteins as that of Goel and Yčas [46], the minimum distance parameters were obtained by rounding from the set of near neighbour distances: the smallest 1% of the distances found between jth neighbours were discarded as possible measurement errors from the X-ray diffraction technique, and the next smallest value chosen to act as the parameter.

9.3.2 Far Neighbour Parameters.

Unfortunately, individual far neighbour parameters cannot be effectively included for each amount of residue separation in primary structure. The mean value for the set of all far neighbour residue pairs separated by some fixed number of residues shows a high degree of variability both within a single protein and between proteins. Therefore, mean value parameters for far neighbour distances will not be used in the present model. The only far neighbour parameters used in the model are absolute minimum and maximum bounds on far neighbour distances. These parameters are independent of the amount of separation of the residues with respect to primary structure.

The minimum value parameter is obtained through Havel *et al.* [50], who reasoned that 5 Å is a commonly observed minimum C_α-C_α distance in proteins and hence an effective excluded volume diameter for a protein residue. Far neighbour minimum value constraints are a necessity in distance constraint models in order to prevent self-intersection of the protein chain.

The parameter U_N was found semi-empirically during this research in order to find a value for the maximum distance for pairs of residues in a protein of length n. This maximum distance is a function of the size of the protein, and replaces the upper bound scalar estimates [50] or protein-specific values [46] of previous models.

The parameter for the maximum far neighbour distance U_N is now derived. Using all possible pairs of C_α-atoms in each protein, the maximum far neighbour distances between residues were found for the same set of twenty proteins as that of Goel and Yčas [46]. An analysis was performed on these maximum distances, where the linear least-squares regression equation and the linear correlation coefficient were calculated.

The results were as follows:

$$y = 9.91n^{\frac{1}{3}} - 8.75 \qquad (55)$$

(correlation: $r = 0.93$)

where y represents the maximum pairwise distance between C_α-atoms in a protein of length n.

This type of constraint may not be cost effective in distance constraint models that also include explicit hydrophobicity constraints, since there are of the order of n^2 pairs of far neighbour residues, and the hydrophobicity conditions (the objective function of the present model) will already force the tertiary structure to have a globular shape. At any rate, the constraint may easily be removed from the model by means of a simple binary flag in the computer implementation.

The parameters used in the far neighbour distance constraints of the present model are shown in Table 22:

Distance	Mean Value	Minimum Bound	Maximum Bound
Far neighbour	highly variable	5.00[a]	$(9.91n^{\frac{1}{3}} - 8.75)$[b]

Statistical information for the distances (in Å) between far neighbour C_α-atoms. Let n represent the number of residues in the protein. The value for the minimum distance bound is independent of n. The sources used for this Table are: (*a*) Havel *et al.* [50], (*b*) Chapter 9.3.

Table 22: Far Neighbour Parameters for Distance Constraint Models.

9.3.3 Hydrophobicity Parameters.

The hydrophobicity parameters used in the present model are given in Table 24. In the model, the ideal position for all hydrophobic residues is given to be the centroidal point. The parameter D, an ideal distance from the centroidal point for hydrophilic residues, has been calculated using semi-empirical results (Table 23) for the average distances of the residues from the centroidal point of the protein. The value of D, a function of the number of residues of the protein, is determined by requiring the average centroidal point distance for the hydrophilic and hydrophobic residues to equal a corresponding semi-empirical value.

Goel and Yčas [46] classified three hydrophobicity categories for the residues empirically (Table 3 of Chapter 3.3), according to their observed distances from the centroidal points of proteins. As explained in Chapter 3.3, this "geometrical hydrophobicity" classification was not explicitly calculated from physicochemical properties of the sidechains as were the classifications from several other sources [16,17,34,55,59,64,80], but best describes the empirical hydrophobicity characteristics of known protein tertiary structures. The average distance of each kind of amino acid from the centroidal point was calculated for twenty-one protein structures, and the residues were classified on the basis of differences in the average distances from the centroidal point and consistency of behaviour. Geometrically, Gly and Pro were found to behave as hydrophilic residues, due to their abundance in external turns. The residue His showed erratic behaviour with respect to the centroidal point, and so it was classified as ambivalent..

Goel and Yčas found a very high correlation between the cube roots of the number of residues in the proteins and the mean hydrophobicity class distance from the centroidal

points; therefore, they fitted linear regression equations by the method of least squares to express the relationship between these two variables. Their results are used in the calculation of the parameter D in the present model and are consequently shown here (Table 23).

Class	a	b	Correlation
Hydrophobic	-2.52	2.74	0.92
Hydrophilic	0.31	2.80	0.97
Ambivalent	-2.02	3.02	0.96
{Gly,Pro}	3.57	2.23	0.88
All residues	-0.89	2.80	0.98

Linear least squares regression equations for the various hydrophobicity classes of amino acids. Let $y = a + bn^{\frac{1}{3}}$, where y is the mean distance of the C_α-atoms of the class from the centroidal point in Å, and n is the number of residues in the protein. The amino acid types that constitute each of the three hydrophobicity classes are given in Table 3.

Table 23: Regression Statistics for Hydrophobicity Classes of the Amino Acids (from Goel and Yčas 1979).

The parameter D is estimated using the expression of Goel and Yčas (Table 23) for the mean distance of any residue from the centroidal point for a protein of length n:

$$y = 2.80n^{\frac{1}{3}} - 0.89 \tag{56}$$

(correlation: $r = 0.98$).

If it is assumed that hydrophobic residues (class J_1) have a tendency toward the centroidal point (*i.e.*, to $y = 0$) and the hydrophilic residues (class J_2) tend away from the centroidal point (*i.e.*, to $y = D$), then for the residues in classes J_1 and J_2 combined to satisfy their mean distance formula, it must be that:

$$D = \frac{n_1 + n_2}{n_2}(2.80n^{\frac{1}{3}} - 0.89), \tag{57}$$

where the class J_1 contains n_1 residues and class J_2 contains n_2 residues of the protein.

The value of the parameter D, then, is chosen to elicit the hydrophobicity rule of "hydrophilics out, hydrophobics in", while causing the protein to conform to a desired volume (generated semi-empirically). The close packing of residues is known to vary little from protein to protein [94]. This principle is envoked in deriving the equation (57) for D, and is reflected in the high correlation coefficient found for the relationship (56).

Class	Ideal Distance from Centroidal Point
Hydrophobic	0.00
Hydrophilic	$\frac{n_1+n_2}{n_2}(2.8n^{\frac{1}{3}} - 0.89)$
Ambivalent	No tendency

The model parameters for "ideal" distances (in Å) of the hydrophobicity classes from the centroidal point of the protein. These parameters will not be satisfied exactly upon optimization of the model algorithm, but instead represent tendencies for the various residues to be situated close to, or away from, the centroidal point. Let n_1 and n_2 represent the number of hydrophobic and hydrophilic residues, respectively, in a protein of total length n. The hydrophilic class parameter used in this Table is derived in this section, using the results from Goel and Yčas [46] shown in Table 23.

Table 24: Hydrophobicity Parameters for the Distance Constraint Model.

The hydrophobicity rule is presented as the objective function of the nonlinear programming formulation of the present model. This means that the residues of a protein will attempt to reach the centroidal point distances given by the parameters of Table 24 as closely as possible, such that the other model constraints are satisfied *exactly*. These "ideal" centroidal point distances are not actually realized in the final folded structure of a protein.

The objective function represents an idealization of the hydrophobicity rule, even to the extent of being chemically inaccurate. Chemically, the rule follows from the

polarity of the sidechains of the residues; nonpolar sidechains seek the hydrophobic interior of the globule (not the centroidal point) and the polar sidechains seek the aqueous environment at the protein-water interface (not the surface of a sphere of radius D). However, the objective function in its present form can mathematically reflect the empirically found hydrophobicity tendencies of the residues in terms of pairwise distances. This effects the hydrophobicity rule in a much simpler form than one which would be obtained by chemical considerations.

The artificial construction of the sphere of radius D centered at the centroidal point avoids the difficult problem of attempting to define a "surface" for a globular protein, an imaginary shell separating an "inside" from an "outside" of the molecule. Proteins really do not possess anything that resembles a surface. Methods for geometrically defining a surface for a protein in order to empirically assign residues to an inside or an outside of a protein of known structure [65,102,124] tend to be difficult and somewhat arbitrary.

As a final note, it has been empirically determined that the N- and C-terminal residues of globular proteins behave as hydrophilic, irrespective of the type of amino acid [63]. In respect for this property, the distance constraint models of Kuntz *et al.* [63], Goel *et al.* [45] and Sanati [98] reclassify the terminal residues of the chain (and sometimes their nearest neighbours) as hydrophilic, regardless of their original hydrophobicity classes. These reclassifications are not carried out in the present model. It was found during the course of experimenting with the model that better results in RMS error were obtained by not separately reclassifying these residues, and that these residues tended generally to situate on the outside of the folded protein in any event. This result is due in part to the hydrophobicity condition being implemented as the

objective function in the mathematical model. The observed centroidal point tendency for the *N*- and *C*-terminals appears to be an artifact of the folding geometry and not an actual alteration of the hydrophobicity characteristics of the residues.

9.3.4 Disulfide Bond Parameters.

Disulfide bonds are covalent cross links between pairs of Cys residues. They are technically not part of the primary structure, but can be readily found by chemical means.

Goel and Yčas [46] used a set of twenty disulfide bonds from six proteins in order to calculate empirical statistics for C_α-atoms of cysteine residues linked by disulfide bonds. Their results are shown in Table 25. Thornton [114] subsequently found similar statistics for a larger database of disulfide bonds. He found that, within the set of all proteins of known structure, there were fifty-five independent examples of disulfide bridges, from twenty-eight proteins. Omitting those disulfides from proteins with very similar structures, he found statistics for the thirty remaining disulfide bridges of known geometry. These statistics are also given in Table 25.

The present model will have the option to use a mean value distance parameter for C_α-atoms of disulfide-bonded Cys residues. The mean value of 5.69 found by Thornton [114] will be used for this purpose since it represents the stronger empirical finding. It should be noted, however, that the model is not particularly sensitive to the value of this parameter.

If the pairwise distance $d_{i,i+j}$ between any two specific residues i and $i+j$ of a protein can be determined with some accuracy by chemical or other means (*cf.*, Chapter 3.4), their tertiary separation can be easily included in the mathematical model

Reference	Mean S-S Value	Standard Deviation	Minimum Distance	Maximum Distance
Goel and Yčas [46]	5.46	0.80	3.87	6.62
Thornton [114]	5.69	0.76	4.6	7.4

Statistical information for the distances (in Å) between C_α-atoms of pairs of Cys residues participating in disulfide bonds. In the model of this thesis, the mean S-S value of 5.69 from Thornton [114] is employed.

Table 25: Disulfide Bond Parameters for Distance Constraint Models.

by substituting the value $d_{i,i+j}$ for the disulfide mean value $\overline{d}_S$ in a disulfide mean value constraint. The disulfide bond constraints in the model may then be envisaged as the general store of extra-primary information.

9.3.5 Summary of Parameter Values.

The values of the parameters used in this model are given in Table 26. When the value for each parameter is decided upon, the question arises whether to attempt a theoretical derivation from first principles or to estimate the value statistically from a database of the known tertiary structures of proteins. It is argued in Chapter 11 that theoretical calculations are valuable in estimating upper bound distance parameters, whereas it is most reasonable to estimate mean value and lower bound distance parameters semi-empirically. This method has been followed, and the calculated theoretical results from Chapter 11 are used along with the empirical findings of Pauling *et al.* [81], Havel *et al.* [50], Goel and Yčas [46] and Thornton [114].

Mean value parameters:	$\overline{d}_1 = 3.80$	$\overline{d}_2 = 5.95$	$\overline{d}_S = 3.69$
Lower bound parameters:	$L_3 = 4.5$	$L_4 = 4.5$	$L_N = 5.0$
Upper bound parameters:	$U_3 = 10.7$	$U_4 = 13.9$	$U_N = 9.91n^{\frac{1}{3}} - 8.75$
Centroidal point distance parameter:	$D = \frac{n_1+n_2}{n_2}(2.80n^{\frac{1}{3}} - 0.89)$		

The values of the parameters used in the mathematical model. The value for $\overline{d}_1$ is obtained from the polypeptide structure determination of Pauling *et al.* [81]. The values for $\overline{d}_2$, L_3 and L_4 are taken from the semi-empirical results of Goel and Yčas [46]. The value for L_N is obtained from Havel *et al.* [50]. The value for $\overline{d}_S$ is from Thornton [114]. The values for U_3 and U_4 were calculated theoretically by the author from standard chemical bonding considerations (Chapter 11), using normal Ramachandran limits on the dihedral angles (ψ, ϕ). The values for U_N and D are derived by the author in Chapter 9.3, using semi-empirical methods. The notations n_1 and n_2 represent the number of hydrophobic and hydrophilic residues, respectively, in the protein of length n.

Table 26: Parameter Values for the Present Model.

10 Appendix: The Algorithms.

The optimization method employed in this research was specifically designed for this problem by the author in collaboration with P.F. O'Neill of the Department of Mathematics, Statistics and Computing Science at Dalhousie University (Foster and O'Neill [41], unpublished). The optimization method minimizes the penalty function $p(x, \mu)$ given by equation (53) of Chapter 9.2. The method is derived from two of the fundamental algorithms for minimizing a continuously differentiable function: *Newton's method* and *steepest descent.* The reader is referred to a standard text such as Gill and Murray [43] or Fletcher [39] for a more detailed discussion of these techniques.

It was found during the course of the research that the large-scale nature of the present problem limited the effectiveness of the current general purpose algorithms available for nonlinear optimization. If second order information is to be utilized, as in the case of quasi-Newton algorithms, the computer storage of the matrix of second derivatives quickly becomes a difficult issue. If the second order information is not exploited, as in the steepest descent algorithm or Monte Carlo methods, the convergence rate will be slow or nonexistent and execution time can become prohibitive to a solution. Exact penalty function methods (*cf.,* Coleman and Conn [24]), although not requiring second derivative calculations, generally entail a great deal of computer space for storage of the active constraint locations and gradients. Therefore, an algorithm was developed in order to combine efficient second order convergence properties with low storage requirements.

10.1 Notation.

In the notation used to describe the algorithms, iteration counters appear as superscripts. Thus, x^k is the vector x at iteration k and not the scalar x to the power k. In each case, the algorithm is given to minimize $f(x)$ where $x \in R^n$ and $f(x) \in R^1$. The notation $\nabla f(x^k)$ refers to the vector of partial derivatives of f evaluated at x^k; $H(x^k)$ refers to the Hessian matrix of second order mixed partial derivatives of f evaluated at x^k.

Let n refer to the number of residues in a protein to be folded. Then the number of variables to be optimized is $3 \times n$ when the protein is represented by the Cartesian coordinates of its C_α-atoms.

10.2 Newton's Method.

Newton's method is based on a quadratic model. A Taylor series expansion of $f(x)$ about x^k, truncated after the quadratic term, is used to approximate $f(x)$. In this way, local second order information from the Hessian matrix (H) of second partial derivatives can be utilized. The first and second partial derivatives for the terms comprising the penalty function of the model can quite easily be calculated explicitly and therefore, exact formulae are used for all required derivatives in the program.

Algorithm **NEWTON**:

step (1). Input x^0. Set $k \leftarrow 0$.

step (2). Solve $H(x^k)d = -\nabla f(x^k)$ for d.

step (3). Set $x^{k+1} \leftarrow x^k + d$.

Set $k \leftarrow k + 1$.

step (4). If $\nabla f(x^k) = 0$, stop;

otherwise, go to step (2).

(In practice, terminate when $\|\nabla f(x^k)\| < \epsilon$).

The system of linear equations at step (2) is referred to as the Newton equations The search vector d is called the Newton direction or Newton step. In the present case, the $3 \times n$ vector x^k is the representation of the Cartesian coordinate locations for the C_α-atoms of a protein at iteration k.

The advantage of Newton's method is a quadratic rate of convergence in the neighbourhood of a strong local minimizer. There are several disadvantages, however:

1. the sequence of iterates may not converge;
2. the algorithm is undefined if $H(x^k)$ is singular;
3. the vector d may not be a descent direction; hence, the iterates may converge to a maximum or to a saddle point and not to a minimum;
4. a $3n \times 3n$ system of linear equations must be solved at each iteration.

10.3 Steepest Descent Method.

Algorithm **STEEPEST DESCENT:**

step (1). Input x^0. Set $k \leftarrow 0$.

step (2). Set $d \leftarrow -\nabla f(x^k)$.

step (3). Find λ such that $f(x^k + \lambda d) << f(x^k)$.

step (4). Set $x^{k+1} \leftarrow x^k + \lambda d$.

Set $k \leftarrow k + 1$.

step (5). If $\nabla f(x^k) = 0$, stop;

otherwise, go to step (2).

(In practice, terminate when $\|\nabla f(x^k)\| < \epsilon$).

In the present model, a cubic linesearch algorithm is implemented to calculate the steplength λ at each iteration. To guarantee that $f(x^k + \lambda d) << f(x^k)$, the Armijo-Goldstein conditions [47] are applied:

Find λ such that $f(x^k + \lambda d) \leq f(x^k) + \alpha\lambda\nabla f(x^k)^T d$ and $\nabla f(x^k + \lambda d)^T d \geq \beta\nabla f(x^k)^T d$, where $\alpha \in (0, \frac{1}{2})$ and $\beta \in (\alpha, 1)$.

The steepest descent method has several advantages:

1. the sequence of iterates always converges;

2. the algorithm is always defined if f is continuously differentiable;

3. the vector d is always a descent direction;

4. there are relatively few arithmetic operations per iteration.

The disadvantage of the steepest descent algorithm is that it has only a linear rate of convergence in the neighbourhood of a solution. In fact, this method usually shows oscillatory behaviour in the vicinity of a solution, and round-off effects can cause termination before the solution is reached.

Thus, if x^k is close to a minimum, Newton's method will exhibit a much superior rate of convergence than the steepest descent method. If x^k is not close to a minimum, Newton's method may not offer a stronger convergence and will be much more expensive to calculate per iteration. The obvious strategy is to use steepest descent as long as x^k appears not to be in the vicinity of a minimum (*i.e.*, if $\|\nabla f(x^k)\|$ is "large");

otherwise, use Newton's method. This strategy can be refined, however, by using an approximation of the Newton direction and allowing the accuracy of the approximation to increase as a minimum is approached, thus decreasing the average number of calculations per iteration. Furthermore, it is desirable to compute the approximate Newton direction in such a way that the sparsity of H can be fully exploited.

10.4 Truncated-Newton Method.

The "refined strategy" given above has been called a "truncated-Newton method" by Dembo and Steihaug [33], who discuss its properties in detail. The essential reasoning behind the truncated-Newton algorithm is that the complete solution of the Newton equations at each iteration is expensive to compute and is not expressly required when far from a solution. In large-scale problems such as this, the Newton equations must be solved by means of an iterative method due to computer storage limitations In this event, there is a trade-off between the amount of accuracy with which the Newton equations are solved and the execution time used to compute a search direction. In the truncated-Newton method, imprecise solutions are found for the Newton equations using an iterative method, in order to find an acceptable approximation for the Newton direction. The accuracy of these solutions is gradually increased as the algorithm approaches an overall solution.

The truncated-Newton strategy is implemented by a modification of the algorithm NEWTON at step (2). The second step of algorithm NEWTON is replaced by:

step (2). Apply an iterative method to solve the system of linear equations

$$H(x^k)d = -\nabla f(x^k).$$

If $\|H(x^k)d^i + \nabla f(x^k)\| \leq \gamma^k \|\nabla f(x^k)\|$ at iteration i of the iterative method for

some constant $\gamma^k > 0$, terminate the iterations and set $d \leftarrow d^i$.

If $\nabla f(x^k)^T d < 0$, accept d as a useable approximation of the Newton direction; otherwise set $d \leftarrow -\nabla f(x^k)$ and perform a steepest descent iteration.

The value of the scalar γ^k in the above is chosen to be

$$\gamma^k = \min\left(\frac{1}{k}, \|\nabla f(x^k)\|\right).$$

When far from a solution, $\|\nabla f(x^k)\|$ is large and hence it is inexpensive to compute an acceptable approximate direction d. However, as a minimum is approached, the sequence $\{\gamma^k\}$ forces the approximate solution of the Newton equations to become increasingly accurate.

There are additional safeguards that are observed in the truncated-Newton algorithm. If at iteration i of the iterative method used to solve the Newton equations there is encountered a direction of negative curvature in the update for d^i (*i.e.*, $d^{i^T} H(x^k) d^i < 0$), then the iterations terminate. The existence of such a direction implies that f is not convex in the neighbourhood of x^k and therefore, the Newton direction may not point toward a local minimum. In this case, the current estimate d^i is a useable descent direction [33], although it does not constitute a proper Newton step and the linesearch must be implemented to estimate a proper step size. In order to guard against nonconvergence of the iterative scheme, a maximum bound is placed on the number of iterations allowed.

10.5 Conjugate Gradient Method

The essence of the approximation technique for finding the Newton direction is to use an iterative scheme for solving the Newton equations and to terminate the iterations

when the trial solution is either a sufficiently good approximation or reaches a direction of negative curvature. In the present approach, a *conjugate gradient* algorithm has been implemented as the iterative scheme for solving the set of linear equations. The conjugate gradient method was chosen because of its robustness, low storage requirements and desirable convergence properties. Also, it is well-suited for use with the truncated-Newton method because it minimizes the quadratic approximation of $f(x)$ over the subspace spanned by the directions that are generated. The reader is referred to Gill and Murray [42] for more details. Note that the conjugate gradient method was employed only for solving the linear Newton equations $H(x^k)d = -\nabla f(x^k)$, and a conjugate gradient algorithm was not used to solve the general nonlinear programming problem.

Let $H(x^k)$, $\nabla f(x^k)$ and γ^k be denoted by H, ∇f and γ, respectively. For each k, the maximum number of iterations (denoted by i) permitted in the conjugate gradient algorithm is denoted by *maxit*.

Algorithm **CNJGRD** (conjugate gradient method).

step (1). Set $d^0 \leftarrow 0$, $r^0 \leftarrow -\nabla f$, $p^0 \leftarrow r^0$, $\delta^0 \leftarrow {r^0}^T r^0$, $i \leftarrow 0$.

step (2). Set $q^i \leftarrow Hp^i$.

If ${p^i}^T q^i \leq \epsilon \delta^i$, set $d \leftarrow \begin{cases} p^0 & \text{if } i = 0 \\ d^i & \text{otherwise} \end{cases}$ and stop, (p^i is a direction of negative curvature);

otherwise, go to step (3).

step (3). Set $\alpha^i \leftarrow ({r^i}^T r^i)/({p^i}^T q^i)$, $d^{i+1} \leftarrow d^i + \alpha^i p^i$, $r^{i+1} \leftarrow r^i - \alpha^i p^i$.

If $\|r^{i+1}\| \leq \gamma \|\nabla f\|$, set $d \leftarrow d^{i+1}$ and stop, (d is an approximation of the Newton

direction);

otherwise, go to step (4).

step (4). Set $\beta^i \leftarrow (r^{i+1^T} r^{i+1})/(r^{i^T} r^i)$, $p^{i+1} \leftarrow r^{i+1} + \beta^i p^i$, $\delta^{i+1} \leftarrow (r^{i+1^T} r^{i+1}) + \beta^i \beta^i \delta^i$.

Set $i \leftarrow i + 1$.

If $i \leq maxit$, go to step (2);

otherwise, stop.

A sufficient condition for convergence of the algorithm CNJGRD is that H is positive definite. However, the algorithm may converge even if this condition is not satisfied. To prevent infinite cycling in case of nonconvergence, no more than *maxit* iterations are allowed to be performed. The truncated-Newton strategy rarely permits the number of iterations in CNJGRD to reach $maxit = 3 \times n$ in practice.

10.6 Solving the Nonlinear Programming Problem.

The nonlinear programming problem is solved by converting the objective function and constraints into the penalty function $p(x, \mu)$, which allows the constrained problem to be solved by the use of an unconstrained algorithm. The overall solution is obtained by performing several iterations wherein first $p(x, \mu)$ is minimized for a fixed value of μ and then μ is reduced in value. If the starting point (x^0) is sufficiently close to a constrained local minimum of the nonlinear programming problem, then as $\mu \rightarrow 0$, the sequence of points thereby generated will converge to a constrained local minimum.

The truncated-Newton procedure will always converge to a local minimum of the penalty function. However, unless x^0 is sufficiently close to the set of points which satisfy the constraints of the nonlinear programming model, it may not converge to a

point which satisfies the constraints.

10.7 Outer Loop Algorithm.

An algorithm called Outer Loop is employed to control the reduction of μ and to test for convergence. Outer Loop calls another algorithm Inner Loop that minimizes $p(x, \mu)$ for a fixed value of μ.

Let m be the maximum number of iterations of Outer Loop; typically $m = 4$ or $m = 5$ Let $\epsilon_1, \epsilon_2, \ldots, \epsilon_m$ be vectors containing stopping tolerances for each successive call to Inner Loop. Let x^0 be the starting point (*i.e.,* the initial configuration). Let δ be a tolerance for ∇f which controls when Newton steps are to be attempted. Let *newt* be a binary flag which allows Newton steps to be performed only on the final call to Inner Loop. Let j be the iteration counter for the number of Inner Loops.

Algorithm **OUTER LOOP**:

step (1). Input $m, \epsilon_1, \epsilon_2, \ldots, \epsilon_m, x^0, \delta$.

Set $j \leftarrow 1$, $\mu \leftarrow 1$, *newt* $\leftarrow$*false.*

step (2). If $j = m$, set *newt* $\leftarrow$*true.*

step (3). Call Inner Loop with $\epsilon = \epsilon_j$.

step (4). Set $x^{j+1} \leftarrow x^*$, $j \leftarrow j + 1$.

step (5). If $j > m$, stop;

otherwise, set $\mu \leftarrow \mu/10$ and go to step (2).

Due to the efficiency of the steepest descent algorithm away from the neighbourhood of the optimum point, the truncated-Newton method is used only for the final iteration

of the Outer Loop algorithm in the current version of the model. In all preceding Outer Loop iterations, steepest descent is employed exclusively. Thus for all Outer Loops except the last, the optimization is continued only until a neighbourhood of a $p(x, \mu)$ minimum is reached. It was found that further reduction of $p(x, \mu)$ in these early outer loops does not give a significant difference in the coordinates of x^*, and is not worth the expense. The final Outer Loop iteration is continued until the truncated-Newton method converges to a strong minimum.

10.8 Inner Loop Algorithm.

For a fixed value of μ, let $f(x) = p(x, \mu)$. Algorithm Inner Loop solves the problem of minimizing $f(x)$, using a combination of truncated-Newton and steepest descent methods. The test $\|\nabla f\| < \delta$ is used to determine when a neighbourhood of a local minimum is reached, (*i.e.*, when it is appropriate to attempt truncated-Newton steps). In practice, however, for all the optimizations whose results are reported in Chapter 7, the truncated-Newton steps are attempted only in the final iteration of the Outer Loop algorithm. That is, *newt* = *false* until the initiation of the final Outer Loop iteration, when *newt* = *true*.

The algorithm Inner Loop is given by means of a flow chart, labelled Figure 16.

10.9 Compact Storage of the Hessian.

The design of the penalty function is such that the Hessian is sparse. This is of practical necessity since the full storage and manipulation of the $3n \times 3n$ matrix even for an average-sized protein would not be possible on most mainframe computers.

Due to the simple form of the penalty function, it is possible to calculate explicitly all the terms for the Hessian matrix. Because the centroidal point of the configuration

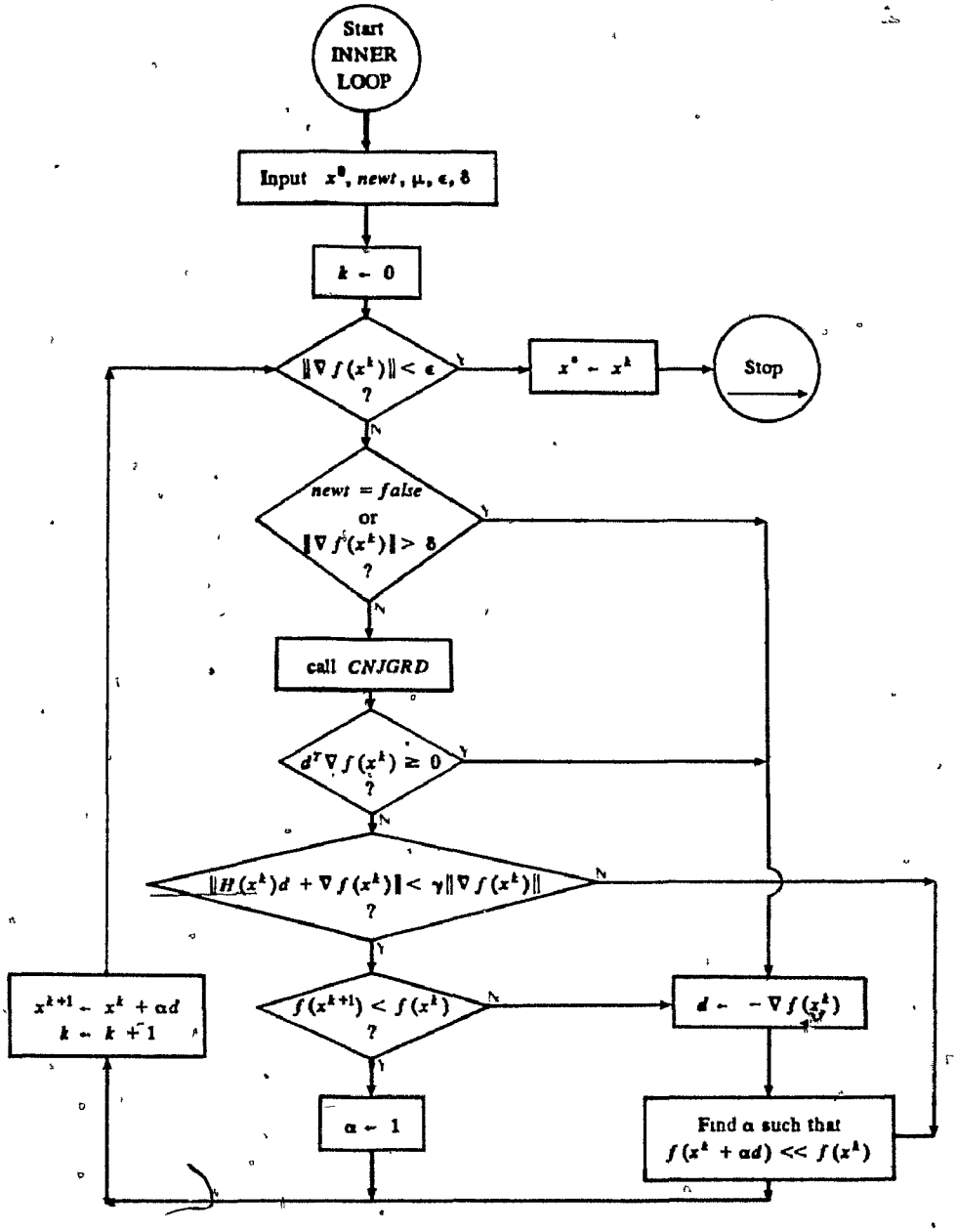


Figure 16: Flowchart of the Inner Loop Algorithm.

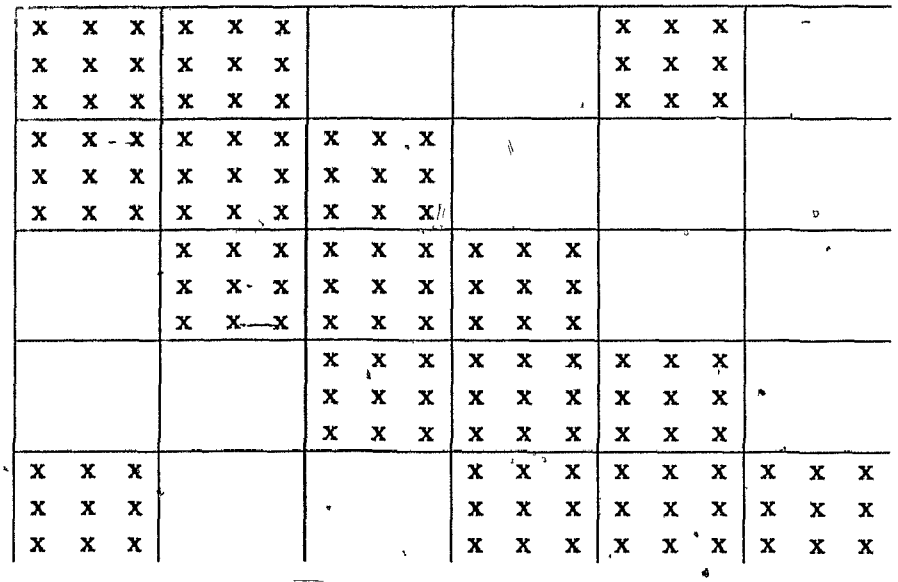


Figure 17: The Nonzero Hessian Matrix Elements.

was required to lie at the origin in R^3, the hydrophobicity conditions give rise to nonzero Hessian elements only in a pattern of 3×3 blocks along the main diagonal. The constraint that forces the centroidal point to lie at the origin results in nonzero, but constant, terms for every third diagonal of the Hessian, and zeroes elsewhere; however, none of these elements need to be stored explicitly. (The nonzero entries are the $\partial x_j \partial x_k$, $\partial y_j \partial y_k$, $\partial z_j \partial z_k$ terms, and the zero entries are the cross terms.) Each of the near neighbour constraints, disulfide bond constraints, or unsatisfied maximum or minimum distance constraints contributes a term to a 3×3 block on the main diagonal as well as giving an extra 3×3 block of nonzero elements on either side on the main diagonal. In the diagram of Figure 17, each "x" represents an element of the Hessian which may be nonzero, for the case where the jth nearest neighbour constraint set includes only the first neighbour distances ($j = 1$) and a disulfide bond exists between residues 1 and 5.

The first band of blocks above the main diagonal are for first neighbour penalties;

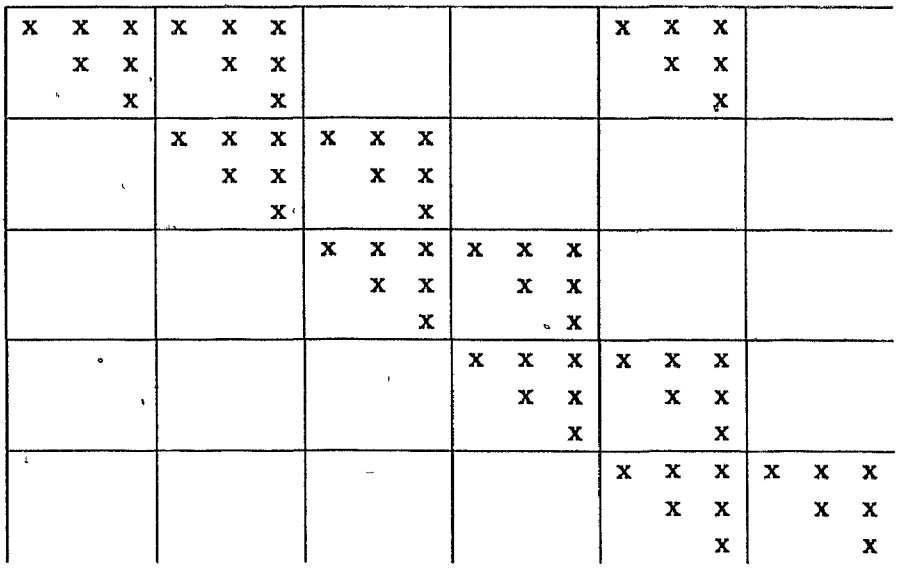


Figure 18: Required Hessian Matrix Elements, Symmetry Included.

the second band of blocks above the main diagonal are for second neighbour penalties, and so on. If a pair of residues (l,k) are connected by a disulfide bond, then the blocks in positions (l,k) and (k,l) are nonzero. If maximum or minimum distance constraint for $d_{i,i+j}$ is violated then the blocks in positions $(i,i+j)$ and $(j+i,i)$ are nonzero.

Because each 3×3 block is symmetric, and the Hessian itself is symmetric, only the nonzero elements shown in the diagram labelled Figure 18 need to be stored.

Thus the Hessian can be stored compactly during the optimization process, but its elements are easily accessed by a suitable row and column indexing scheme for the blocks.

The set of blocks along the main diagonal of the Hessian are actually stored in an array H_D of n blocks with six elements in each block. The nonzero elements of the off-diagonal blocks are stored in an array H_F containing a total of n^* blocks with six elements in each block, where n^* varies with the number of violated constraints. Experience with the algorithm indicates that n^* is typically of the order of n^* =

$0.005(n^2) + 2n$. This means that the Hessian may be compactly stored in only about $6n + 6n^* = 0.03(n^2) + 18n$ storage locations, compared to the $9(n^2)$ locations that would have been required to store the entire Hessian.

Let the indexing of the full Hessian matrix be referred to as $G(I, J)$, where I is the index of the row, and J is the index of the column. To access an element that would lie in a block along the main diagonal, the correspondences for the six distinct elements of the block are as follows: $H_D(1, I) = G(3I-2, 3I-2)$, $H_D(2, I) = G(3I-1, 3I-1)$, $H_D(3, I) = G(3I, 3I)$, $H_D(4, I) = G(3I-2, 3I-1)$, $H_D(5, I) = G(3I-1, 3I)$ and $H_D(6, I) = G(3I-2, 3I)$. To access the nonzero elements of the off-diagonal blocks, an indexing pointer scheme is used. Let $irow(J)$ and $icol(J)$ refer to the block row number and block column number, respectively, of the block $H_F(I, J)$ which contains nonzero elements. For example, consider the case when $G(10, 20)$ contains a nonzero element, such as $G(10, 20) = 3$. This means that the block consisting of $\{G(10, 19)$, $G(10, 20)$, $G(10, 21)$, $G(11, 19)$, $G(11, 20)$, $G(11, 21)$, $G(12, 19)$, $G(12, 20)$, $G(12, 21)\}$ would be denoted a nonzero block. In this case, for some J in the set $J = 1, \ldots, n^*$, the values of the index counters would be $irow(J) = 4$ (where the "4" refers to a block containing rows 10-12) and $icol(J) = 7$ (where the "7" refers to a block containing columns 19-21). Also, the fourth element of the six elements comprising the Jth block of H_F would be equal to 3, or $H_F(4, J) = 3$, since $H_F(4, J)$ would correspond to the element $G(10, 20)$.

Even though the disulfide constraints could sometimes overwrite existing nonzero blocks from near neighbour or minimum or maximum constraints, they are always given separate storage locations in order to not waste execution time searching through H_F each time a disulfide constraint is calculated.

It is possible to devise indexing schemes that are even more compact. One such indexing scheme would be to use a total of n^* row pointers $irow(J)$ as before, but to employ another pointer index $irown(K)$ in place of the set of n^* column pointers $icol(J)$. The pointer $irown(K)$ would contain the number of nonzero blocks in the Kth row. This would result in a vector $irown$ requiring only n elements. The savings of this indexing scheme are not enough to warrant the extra execution time used to store H_F and to access the elements $H_F(6, n^*)$ once they are stored.

In the event of an unusual initial conformation, such as the residues forming a straight line, the first Outer Loop iteration could still encounter Hessian storage problems due to violation of a large number of maximum far distance constraints. The minimum and maximum distance constraints for far neighbours are purposely omitted during the first Outer Loop for this reason, and also because this set of constraints is large (of the order n^2) and costly to evaluate. The maximum far neighbour constraints are relatively unimportant in any event, since the hydrophobicity conditions by themselves effectively shape the residues into its compact globular shape After the first Outer Loop, however, the full set of constraints is included. In the latter stages of the optimization, inclusion of the minimum far neighbour constraints is necessary to ensure that the chain is self-avoiding.

In conclusion, the design of the mathematical model and algorithms are such that the tertiary structure can be efficiently predicted for any single-strand globular protein of natural strand length, from virtually any initial configuration of its points.

11 Appendix: On Theoretical Near Neighbour Distance Parameters.

The purpose of this chapter is to investigate the geometry of near neighbour residues theoretically, and to use the results of this investigation to generate parameters for near neighbour constraints in distance constraint models. The fundamental resource used for the theoretical calculations is a set of chemically derived bond angles and bond lengths for the polypeptide backbone. These chemical data can be considered effectively fixed, or constant. They have been derived from the crystal study of small polypeptides [25,73,81,89], in which resolutions as high as 0.1 Å can be attained. The resulting "standard" bond angles and lengths from these studies are shown in Figure 3 of Chapter 1.

The calculation of the distance between the C_α-atoms of adjacent residues in the polypeptide chain from a set of standard bond lengths and angles is straightforward. Due to the planar nature of the peptide bond, this first neighbour distance $d_{i,i+1}$ can be found by elementary trigonometry. The value is found to be a constant: $d_{i,i+1} = 3.80$Å. Let this distance be denoted by d_1.

The situation is more complicated for a system of three residues. The general backbone configuration for three residues is not planar, nor does it correspond to a constant $d_{i,i+2}$ distance. However, relationships can be found by methods of planar trigonometry that result in an expression for $d_{i,i+2}$ as a function of the peptide bond lengths and angles. The result obtained is identical to the equation (80) derived later in this Appendix, and is found to vary with the nonconstant dihedral angles (ψ, ϕ). These "Ramachandran" angles (ψ, ϕ), discussed in Section 11.1 following, correspond

to angles of rotation about the single bonds in the polypeptide backbone.

It is impractical to attempt the derivation of higher neighbour distances, such as $d_{i,i+3}$ or $d_{i,i+4}$, directly from the bonding lengths and angles. Therefore, the following sections are devoted to their derivation by utilizing a virtual bond description of the polypeptide chain.

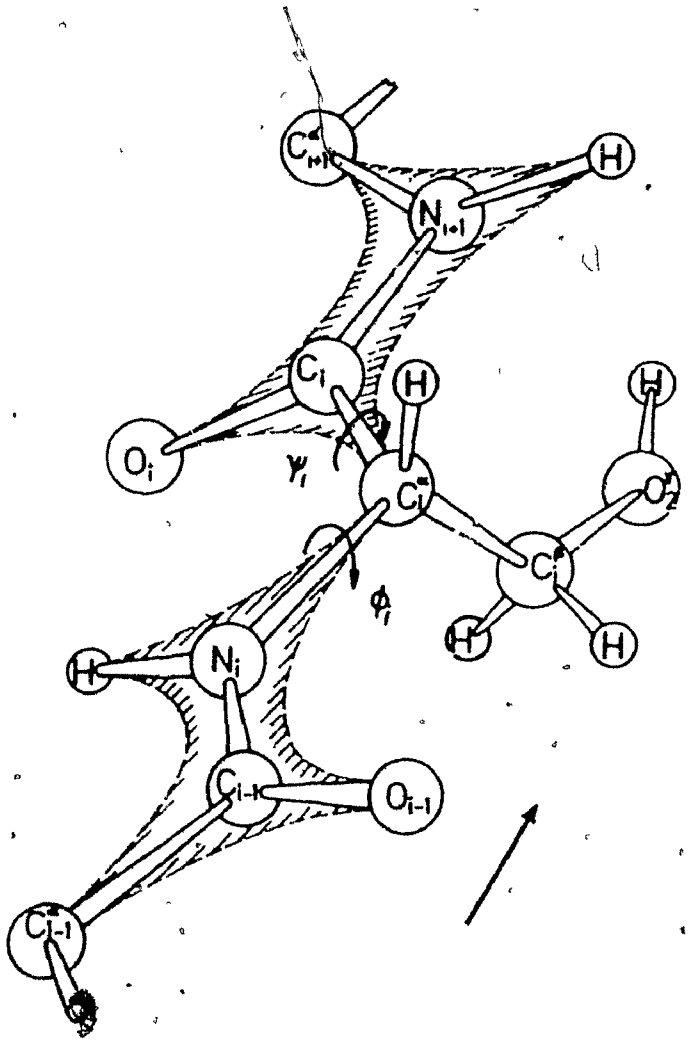
11.1 Ramachandran Angle and Virtual Bond Descriptions of a Polypeptide.

A characteristic feature of a polypeptide chain is that each peptide group is a rigid planar unit: the $C - N$ link is a partial double-bond, which allows for no freedom of rotation. In contrast, the links $C_\alpha - C$ and $C_\alpha - N$ on either side of the peptide unit are pure single bonds, and these bonds allow for a large amount of rotational freedom.

With rigid bond angles and bond lengths and a planar peptide bond, the polypeptide chain has only two degrees of freedom for each residue. These are described by the dihedral angles ψ and ϕ at the C_α-atoms, as shown in Figure 19.

The angle ψ represents the amount of rotation about the axis of the single bond $C_\alpha - C$; similarly, the angle ϕ gives the rotation about the $C_\alpha - N$ axis. The direction of rotation for both ψ and ϕ are defined to be positive when the C-terminal side of the specified bond is rotated in a clockwise direction as viewed from the atom on the N-terminal side of the bond. The zero positions for both ψ and ϕ occur when the two peptide planes joined at the C_α-atom are coplanar and *trans*. Thus the (ψ, ϕ) angles of the dipeptide configuration depicted in Figure 19 are $\psi_i = 180°$ and $\phi_i = -180°$. The angles ψ and ϕ may assume any values from $-180°$ to $+180°$.

A list of the (ψ, ϕ) values for all residues of a single strand protein will completely define the tertiary chain path.



The dihedral angle ψ_i represents the amount of rotation about the axis of the single bond $C^{\alpha}_i - C_i$; the angle ϕ_i represents the rotation about the $C^{\alpha}_i - N_i$ axis. The relative orientation of the two peptide planes in the figure correspond to (ψ, ϕ) angles of $\psi_i = 180°$ and $\phi_i = 180°$.

Figure 19: Definition of Ramachandran Bond Angles for a Polypeptide Chain (from Schulz and Schirmer 1979).

Due to electrostatic interactions and the steric hindrance caused by the bulky peptide unit, not all possible combinations of the rotational angles (ψ, ϕ) are realizable in a protein. G.N. Ramachandran and his group at Madras have studied in detail the restrictions on the ranges of the (ψ, ϕ) angles [90,91,92,99]. For this reason, the dihedral angles (ψ, ϕ) are sometimes referred to as "Ramachandran" angles.

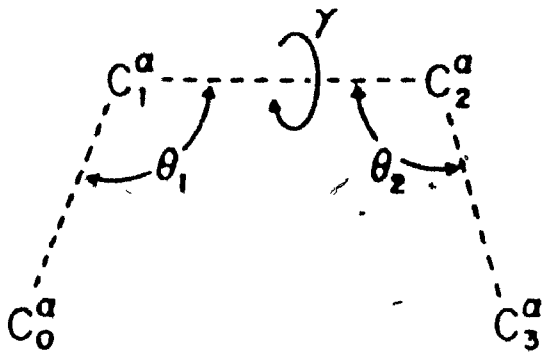
The backbone of a dipeptide is usually described by its four dihedral angles (ψ_1, ϕ_1) and (ψ_2, ϕ_2). This would constitute the complete Ramachandran angle description of a dipeptide configuration. However, the vast majority of peptide groups are planar. Whenever a peptide group is planar, the distance from C_i^α to C_{i+1}^α will be independent of the dihedral angles. This permits an alternate and simpler *virtual bond* description of a dipeptide (Figure 20), wherein the backbone configuration can be specified by only three variables:

1. the angle between the virtual bonds $C_0^\alpha - C_1^\alpha$ and $C_1^\alpha - C_2^\alpha$, designated as θ_1;
2. the angle between the virtual bonds $C_1^\alpha - C_2^\alpha$ and $C_2^\alpha - C_3^\alpha$, designated as θ_2;
3. the dihedral angle around the $C_1^\alpha - C_2^\alpha$ bond, designated as γ.

The virtual bond angle γ is defined to be $\gamma = 0°$ when C_0^α is *cis* to C_3^α, as in Figure 20, and γ may take on any value from $-180°$ to $+180°$. As shown in Figure 20, the clockwise rotation of $C_2^\alpha - C_3^\alpha$ when looking from C_1^α to C_2^α gives a positive change in γ.

The angle θ is restricted, due to the geometry of the peptide chain, to lie within a range of values:

$$\theta_{\min} \leq \theta \leq \theta_{\max}.$$



Definitions of the angles θ_1 and θ_2 between the virtual bonds and the dihedral angle γ for a dipeptide.

Figure 20: Definition of Virtual Bond Angles for a Dipeptide (from Nishikawa *et al.* 1974).

The range of possible values for this virtual bond angle θ can be found empirically in much the same fashion as the bounds for the dihedral angles (ψ, ϕ), or it may be calculated directly from known (ψ, ϕ) bounds once a mapping from (ψ, ϕ) coordinates to (θ, γ) coordinates can be determined.

Both the Ramachandran angle (ψ, ϕ) description and the virtual bond angle (θ, γ) description of a polypeptide will be used for the theoretical research on near neighbour distances in the following sections. The use of virtual bonds facilitates the calculations involved, whereas the known range restrictions on the more traditional Ramachandran angles are used to obtain numerical results.

11.2 Theoretical Calculation of Near Neighbour Distances as Functions of Virtual Bond Angles.

In this section, the near neighbour distances will be calculated as functions of the virtual bond angles (θ, γ). In Section 11.3, the relationship between the virtual bond angles and the Ramachandran angles (ψ, ϕ) will be utilized to formulate the near neighbour distances as functions of the Ramachandran angles.

The pairwise distances $d_{i,i+2}$, $d_{i,i+3}$ and $d_{i,i+4}$ will first be found as functions of the virtual bond angles (θ, γ), which is quite easily accomplished due to the way that this coordinate system simplifies the geometry of the polypeptide. Then the work of Nishikawa *et al.* [79] will be followed in order to generate a mapping between the two sets of coordinates (θ, γ) and (ψ, ϕ). By this approach, it will be possible to find distributions, as well as theoretical mean values and minimum and maximum distances, for $d_{i,i+2}$, $d_{i,i+3}$ and $d_{i,i+4}$, when given various limits and probability distribution sets for the Ramachandran angles (ψ, ϕ) as input.

11.2.1 Finding $d_{i,i+2} = f(\theta)$.

From the chemistry of the peptide bond [73,81,89], it may be assumed that $d_{i,i+1} = d_1$ is a constant and that the peptide groups under consideration are planar.

It is seen from Figure 21 that the distance between C_1^α and C_3^α is a function only of d_1 and the interior virtual bond angle θ. This is defined as the second neighbour distance, designated as $d_{i,i+2}$. The geometry forms an isosceles triangle, which gives the relationship:

$$d_{i,i+2}{}^2 = 2{d_1}^2 (1 - \cos\theta). \tag{58}$$

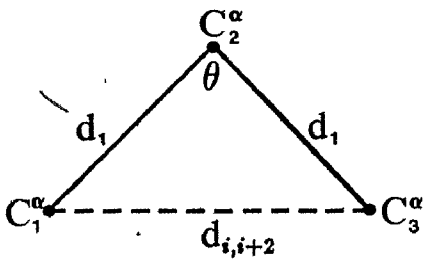


Figure 21: Second Neighbour Distance: Virtual Bond Residues.

From this equation, the following useful inverse relations are calculated:

$$\cos\theta = \frac{2{d_1}^2 - {d_{i,i+2}}^2}{2{d_1}^2} \tag{59}$$

$$\sin\theta = \frac{d_{i,i+2}}{2{d_1}^2}\sqrt{4{d_1}^2 - {d_{i,i+2}}^2}. \tag{60}$$

11.2.2 Finding $d_{i,i+3} = f(\gamma)$.

Consider a dipeptide configuration in a fixed Cartesian coordinate system in R^3, such that all first neighbour distances are of a length d_1. Let:

$$\begin{aligned} C_2^\alpha &= (0,0,0) \\ C_3^\alpha &= (0,-d_1,0), \end{aligned} \tag{61}$$

as in Figure 22.

Translations in the coordinate planes and rotations about the coordinate axes will now be used to determine the coordinate locations of C_1^α and C_4^α. The possible rotations about the three coordinate axes are given by:

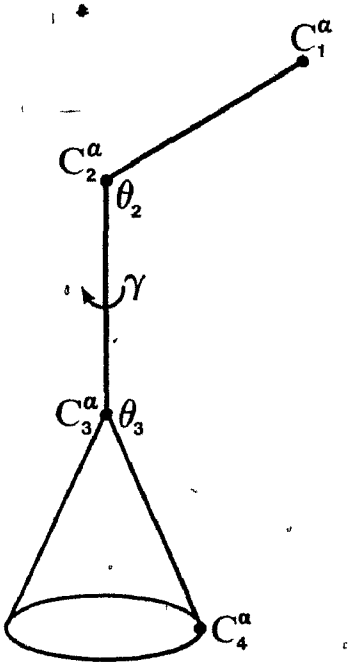


Figure 22: Third Neighbour Distance: Virtual Bond Residues.

$$T_\sigma^x = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\sigma & -\sin\sigma \\ 0 & \sin\sigma & \cos\sigma \end{bmatrix} \tag{62}$$

$$T_\sigma^y = \begin{bmatrix} \cos\sigma & 0 & \sin\sigma \\ 0 & 1 & 0 \\ -\sin\sigma & 0 & \cos\sigma \end{bmatrix}. \tag{63}$$

$$T_\sigma^z = \begin{bmatrix} \cos\sigma & \sin\sigma & 0 \\ -\sin\sigma & \cos\sigma & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{64}$$

Equations (62), (63) and (64) represent the clockwise rotation of an angle σ about the x-axis, y-axis and z-axis, respectively. With reference to Figure 22, it can be seen that:

$$C_1^\alpha = T_{-\theta_2}^z C_3^\alpha = (d_1 \sin\theta_2, -d_1\cos\theta_2, 0) \tag{65}$$

and

$$\begin{aligned} C_4^\alpha &= C_3^\alpha + T_{-\gamma}^y T_{\theta_3}^z (C_2^\alpha - C_3^\alpha) \\ &= (d_1 \sin\theta_3 \cos\gamma, -d_1 + d_1\cos\theta_3, d_1 \sin\theta_3 \sin\gamma). \end{aligned} \tag{66}$$

Equations (65) and (66) may be used to calculate an expression for

$$d_{i,i+3}{}^2 = \|C_4^\alpha - C_1^\alpha\|^2 \tag{67}$$

in terms of θ_2, θ_3 and γ:

$$d_{i,i+3}{}^2 = d_1{}^2[3 + 2\cos\theta_2\cos\theta_3 - 2\cos\theta_2 - 2\cos\theta_3 - 2\sin\theta_2\sin\theta_3\cos\gamma]. \tag{68}$$

Using equations (59) and (60), this expression (68) may be simplified to an equation stating the third neighbour distance $d_{i,i+3}$ as a function of only γ and the interior

pairwise distances:

$$d_{i,i+3}{}^2 = d_1{}^2 + \frac{d_{13}d_{24}}{2d_1{}^2}\left(d_{13}d_{24} - \sqrt{4d_1{}^2 - d_{13}{}^2}\sqrt{4d_1{}^2 - d_{24}{}^2}\cos\gamma\right). \qquad (69)$$

As with equations (59) and (60), the inverse relationship for equation (69) may be found, by solving for γ as a function of the $d_{i,i+j}$:

$$\cos\gamma = \frac{d_{13}{}^2 d_{24}{}^2 - 2d_1{}^2\left(d_{i,i+3}{}^2 - d_1{}^2\right)}{d_{13}d_{24}\sqrt{4d_1{}^2 - d_{13}{}^2}\sqrt{4d_1{}^2 - d_{24}{}^2}}. \qquad (70)$$

11.2.3 Finding $d_{i,i+4} = f(d_{i,i+j})$.

The calculation of the fourth neighbour distance $d_{i,i+4}$ is performed similiarly to the calculation of the third neighbour distance $d_{i,i+3}$. That is, a tripeptide configuration is fixed within an (x, y, z) orthogonal coordinate system in R^3, and the distance $d_{i,i+4}$ is calculated using translations along the coordinate axes and rotations about the axes.

Let the point of departure for the $d_{i,i+4}$ calculation be the dipeptide configuration representation of Figure 22. Consider the conversion of this configuration into a tripeptide by the addition of another residue, represented by a C_0^α atom attached to C_1^α by a virtual bond as in Figure 23.

The value of $d_{04} = d_{i,i+4}$ will be a function of the variables $\{\theta_1, \theta_2, \theta_3, \gamma_1, \gamma_2\}$. Alternatively, by employing equations (59), (60) and (70), it may be expressed as a function of the interior small-neighbour distances $\{d_1, d_{02}, d_{13}, d_{24}, d_{03}, d_{14}\}$, where $d_{i,i+1} = d_1$ is a constant for all i.

The fourth neighbour distance $d_{i,i+4}$ is calculated from

$$d_{i,i+4}{}^2 = \|C_4^\alpha - C_0^\alpha\|^2. \qquad (71)$$

Therefore, the coordinate locations of C_4^α and C_0^α must be found. The coordinates of C_4^α are given in equation (66), replacing γ by γ_2. The coordinates for C_0^α are calculated

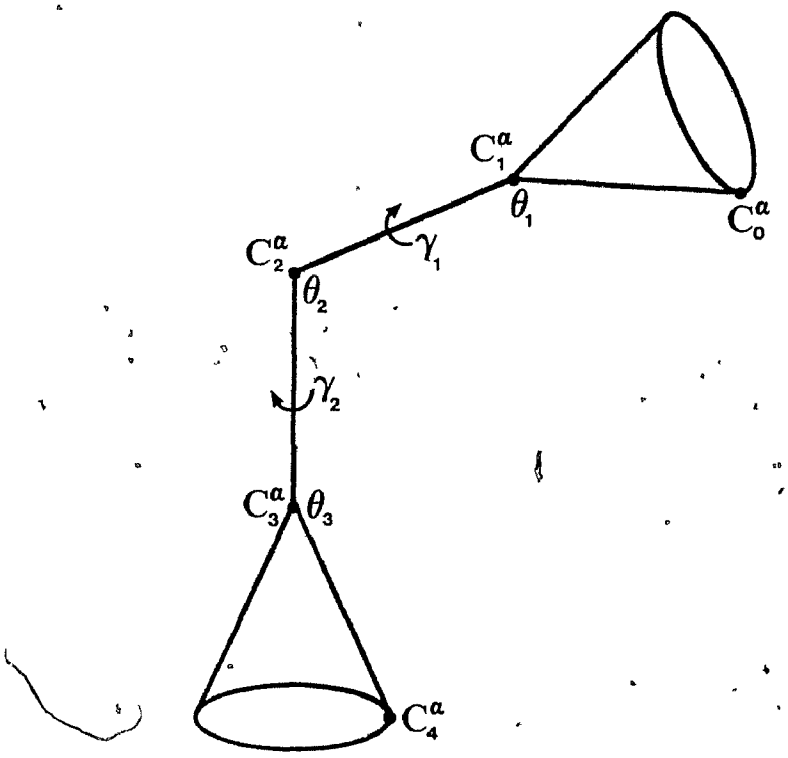


Figure 23: Fourth Neighbour Distance: Virtual Bond Residues.

as follows, using the coordinates of C_3^α given in (61) and the coordinates of C_1^α from (65):

$$C_0^\alpha = C_1^\alpha - T^z_{-\theta_2} T^y_{\gamma_1} T^z_{-\theta_1} C_3^\alpha, \tag{72}$$

yielding:

$$C_0^\alpha = \begin{bmatrix} +d_1 \sin\theta_2 - d_1 \cos\theta_1 \sin\theta_2 - d_1 \sin\theta_1 \cos\theta_2 \cos\gamma_1 \\ -d_1 \cos\theta_2 + d_1 \cos\theta_1 \cos\theta_2 - d_1 \sin\theta_1 \sin\theta_2 \cos\gamma_1 \\ d_1 \sin\theta_1 \sin\gamma_1 \end{bmatrix} \tag{73}$$

Now an expression for $d_{i,i+4}$ may be found, by substituting equations (66) and (73) into equation (71):

$$\begin{aligned} {d_{i,i+4}}^2 \;=\; & 2{d_1}^2[2 - \cos\theta_1 - \cos\theta_2 - \cos\theta_3 + \\ & \cos\theta_1\cos\theta_2 + \cos\theta_2\cos\theta_3 - \cos\theta_1\cos\theta_2\cos\theta_3 + \\ & \sin\theta_1\sin\theta_2\cos\gamma_1(\cos\theta_3 - 1) + \sin\theta_2\sin\theta_3\cos\gamma_2(\cos\theta_1 - 1) + \\ & \sin\theta_1\sin\theta_3(\cos\theta_2\cos\gamma_1\cos\gamma_2 - \sin\gamma_1\sin\gamma_2)]. \end{aligned} \tag{74}$$

Using the identities (59), (60) and (70), $d_{i,i+4}$ may be expressed as a function solely of its interior $d_{i,i+j}$. This expression is cumbersome, however, and therefore will not be shown.

11.3 Theoretical Calculation of Near Neighbour Distances as Functions of the Ramachandran Angles.

In this section, the work of Nishikawa, Momany and Scheraga [79] is followed essentially. They have derived the formulae giving the virtual bond angles (θ, γ) as a function of the Ramachandran angles (ψ, ϕ). Their work is carried further in this thesis, by the calculation of the near neighbour distances $d_{i,i+2}$, $d_{i,i+3}$ and $d_{i,i+4}$ as functions of (ψ, ϕ), using representative values for the fixed bond lengths and angles of the dipeptide units.

Nishikawa *et al.* found that the angle θ depends upon only the two angles (ψ, ϕ) of the particular residue in question; however, the angle γ was found to possess a more complicated relationship, being a function of all four dihedral angles $\{\psi_1,\phi_1,\psi_2,\phi_2\}$ of a dipeptide. In the relationships that are calculated, they include the following angles, which are essentially fixed within a normal dipeptide:

$$\begin{aligned} \alpha &= N_1 - C_2^\alpha - C_2 \\ \xi &= C_1^\alpha - C_2^\alpha - N_1 \\ \eta &= C_3^\alpha - C_2^\alpha - C_2. \end{aligned} \tag{75}$$

The calculated standard values for the angles α, ξ and η as derived from chemical studies [25,73,81,89] are shown in Table 27. The set of values given by Ramachandran *et al.* [89] is considered a refinement of the work of Corey and Pauling [25] and Marsh and Donohue [73]; therefore, it will be used for the numerical calculations in the present study. The results of Corey and Pauling are included in Table 27 because they are still employed in much of the current literature.

angle	Corey and Pauling	Ramachandran *et al.*
$\alpha = N_1 - C_2^\alpha - C_2$	110°	111.6°
$\xi = C_1^\alpha - C_2^\alpha - N_1$	13.2°	14.7°
$\eta = C_3^\alpha - C_2^\alpha - C_2$	22.2°	21.0°

The "standard" values for the peptide bond angles as calculated from the data of two reference sources (Corey and Pauling [25] and Ramachandran *et al.* [89]) are given. Note that these angles are not truly constant over the set of all polypeptides, but can vary in response to local environment. The definitions for the angles α, ξ and η are given in equations (75).

Table 27: Standard Values for the Peptide Bond Angles.

In order to simplify notation, the following are defined initially:

$$\begin{aligned} a &= \cos\eta\cos\alpha + \sin\eta\sin\alpha\cos\psi \\ b &= \cos\eta\sin\alpha - \sin\eta\cos\alpha\cos\psi \\ d &= \cos\xi\cos\alpha + \sin\xi\sin\alpha\cos\phi \\ e &= \cos\xi\sin\alpha - \sin\xi\cos\alpha\cos\phi. \end{aligned} \tag{76}$$

Nishikawa *et al.* use the defined values of a and b in their expression calculated for the value of $\theta = f(\psi, \phi)$:

$$\cos\theta = a\cos\xi + b\sin\xi\cos\phi - \sin\xi\sin\eta\sin\phi\sin\psi. \tag{77}$$

In equation (77), only positive values of θ are considered. From this, it can be seen that θ can be given uniquely within the range $0° < \theta < 180°$ for any value of the set (ψ, ϕ). The maximum and minimum values of θ are obtained by substituting $(\psi, \phi) = (180°, 180°)$ and $(\psi, \phi) = (0°, 0°)$, respectively, into (77):

$$\begin{aligned} \theta_{\max} &= \alpha + (\xi + \eta) \\ \theta_{\min} &= \alpha - (\xi + \eta). \end{aligned} \tag{78}$$

Substituting the standard values for the angles α, ξ and η into (77) via the definitions of a and b gives the value of $\theta = f(\psi, \phi)$:

$$\cos\theta = -0.33 + 0.32\cos\psi + 0.22\cos\phi + 0.03\cos\psi\cos\phi - 0.09\sin\psi\sin\phi. \tag{79}$$

Now the work of Nishikawa *et al.* can be carried further for the present study. The direct relationship between $d_{i,i+2}$ and (ψ, ϕ) can be calculated, using (79) and equation (58) from the previous section:

$$d_{i,i+2}{}^2 = 38.48 - 9.31\cos\psi - 6.36\cos\phi - 0.97\cos\psi\cos\phi + 2.63\sin\psi\sin\phi. \tag{80}$$

As usual, the distance d_1 is treated as a constant $d_1 = 3.80$. It is noted that this equation (80) is identical to the expression for d_1 that can be found using trigonometric methods as discussed at the outset of this Appendix, in which the dipeptide angle measurements are employed directly.

In Nishikawa *et al.*, the virtual bond angle γ is calculated as a function of the Ramachandran angles by first defining auxiliary variables λ_1 and λ_2 as the dihedral angles for rotation of the planar peptide groups about the virtual bonds $C_1^\alpha - C_2^\alpha$ and $C_2^\alpha - C_3^\alpha$, respectively, with respect to the plane formed by the points C_1^α, C_2^α and C_3^α. The angle γ, the angle λ_2 of the first single-residue unit and the angle λ_1 of the second single-residue unit are all defined about the same virtual bond $C_2^\alpha - C_3^\alpha$. Hence, they are found to be related as follows:

$$\gamma = (\lambda_2)_{1st} + (\lambda_1)_{2nd} + 180^\circ, \tag{81}$$

where the constant 180° arises from the definition of the zero positions of $(\lambda_2)_{1st}$ and $(\lambda_1)_{2nd}$. The values for the auxiliary angles λ_1 and λ_2 are found to be

$$\tan\lambda_1 = \frac{-b\sin\phi - \sin\eta\cos\phi\sin\psi}{a\sin\xi - b\cos\xi\cos\phi + \cos\xi\sin\eta\sin\phi\sin\psi} \tag{82}$$

$$\tan\lambda_2 = \frac{-e\sin\psi - \sin\xi\cos\psi\sin\phi}{d\sin\eta - e\cos\eta\cos\psi + \cos\eta\sin\xi\sin\phi\sin\psi} \tag{83}$$

where a, b, d and e are defined in equations (76).

The standard values for the fixed angles from Ramachandran *et al.* [89], given in Table 27, can be substituted into equations (82) and (83), via the relations given in (76). The following are obtained:

$$\tan\lambda_1 = \frac{0.87\sin\phi + 0.13\cos\psi\sin\phi + 0.36\cos\phi\sin\psi}{0.09 - 0.08\cos\psi + 0.84\cos\phi + 0.13\cos\psi\cos\phi - 0.35\sin\phi\sin\psi} \tag{84}$$

$$\tan\lambda_2 = \frac{0.90\sin\psi + 0.09\cos\phi\cos\psi + 0.25\cos\psi\sin\phi}{0.13 - 0.08\cos\phi + 0.84\cos\psi + 0.09\cos\phi\cos\psi - 0.24\sin\phi\sin\psi}. \quad (85)$$

Equation (84) has two solutions for λ_1, but one of the solutions can be eliminated because it gives a negative value for θ when applied to an equation relating ϕ, λ_1 and θ. Similarly, equation (85) has two solutions for λ_2, but one such solution gives a negative value for θ when applied to an equation relating ϕ, λ_2 and θ.

The equations relating $\{\phi,\lambda_1,\theta\}$ and relating $\{\phi,\lambda_2,\theta\}$ are not calculated in Nishikawa *et al.*; and are therefore calculated here.

The equation relating ϕ, λ_1 and θ is obtained by eliminating λ_2 and ψ from the following identity, given by Nishikawa *et al.*:

$$T^x_{\lambda_1}T^z_{(\pi-\theta)}T^x_{\lambda_2} = T^z_{\xi}T^x_{\phi}T^z_{(\pi-\alpha)}T^x_{\psi}T^z_{\eta}. \quad (86)$$

This equation is now rearranged, and multiplied on both sides by the row vector $v = (1,0,0)$ and the column vector $u = v^T$, in order to eliminate ψ and λ_2 from the right-hand side, yielding

$$\begin{aligned}\cos\eta &= \cos\theta(\cos\alpha\cos\xi + \sin\alpha\cos\phi\sin\xi)\\ &+ \sin\theta\cos\lambda_1(\sin\alpha\cos\phi\cos\xi - \cos\alpha\sin\xi)\\ &+ \sin\alpha\sin\phi\sin\theta\sin\lambda_1.\end{aligned} \quad (87)$$

Similarly, the equation relating ϕ, λ_2, and θ is found by rearranging and then multiplying through by a row vector and a column vector:

$$T^x_{\lambda_1}T^z_{(\pi-\theta)}T^x_{\lambda_2}T^z_{-\eta}T^x_{-\psi} = T^z_{\xi}T^x_{\phi}T^z_{(\pi-\alpha)}. \quad (88)$$

Multiplying through by $v = (1,0,0)$ and $u = v^T$ and rearranging gives:

$$\cos\xi\cos\alpha - \cos\eta\cos\theta + \sin\xi\sin\alpha\cos\phi + \sin\eta\sin\theta\cos\lambda_2 = 0. \quad (89)$$

When values for the fixed angles are substituted into these identity equations for λ_1 and λ_2, the results are, respectively:

$$\cos\theta(0.24\cos\phi - 0.36) + \sin\theta\cos\lambda_1(0.09 + 0.90\cos\phi) + 0.93\sin\phi\sin\theta\sin\lambda_1 = 0.93 \quad (90)$$

and

$$-0.93\cos\theta + 0.24\cos\phi + 0.36\sin\theta\cos\lambda_2 = 0.36. \quad (91)$$

To summarize the calculation of γ as a function of Ramachandran angles (ψ, ϕ), it was shown (equation (81)) that γ was given by:

$$\gamma = (\lambda_2)_{1st} + (\lambda_1)_{2nd} + 180°,$$

where expressions for the auxiliary variables λ_1 and λ_2 are found in equations (84) and (85). Now, since equations (84) and (85) afford two solutions for λ_1 and λ_2, it must be decided which solution is correct. This is done by first noting the value of θ, given angles ψ and ϕ, from equation (79):

$$\cos\theta = -0.33 + 0.32\cos\psi + 0.22\cos\phi + 0.03\cos\psi\cos\phi - 0.09\sin\psi\sin\phi.$$

Given this solution for θ, each of the two values of λ_1 are substituted into equation (90). The correct candidate will give an identity in equation (90). Similarly, the two values of λ_2 are substituted into equation (91) to find the true value.

Once the value for γ is obtained, the distance $d_{i,i+3}$ can be found from equation (69) of the previous section:

$$d_{i,i+3}{}^2 = d_1{}^2 + \frac{d_{13}d_{24}}{2d_1{}^2}(d_{13}d_{24} - \sqrt{4d_1{}^2 - d_{13}{}^2}\sqrt{4d_1{}^2 - d_{24}{}^2}\cos\gamma).$$

In this equation, $d_1 = 3.80$ is assumed, d_{13} and d_{24} are the second neighbour distances which can be calculated by equation (80), and the value of $(\cos\gamma)$ can be calculated using equation (81):

$$\begin{aligned} \cos\gamma &= \cos[(\lambda_2)_{1st} + (\lambda_1)_{2nd} + 180^\circ] \\ &= -\cos[(\lambda_2)_{1st} + (\lambda_1)_{2nd}]. \end{aligned} \tag{92}$$

Whereas the calculation of the $d_{i,i+3}$ distance from the corresponding Ramachandran angles requires a good deal of effort, the derivation of the distance $d_{i,i+4}$ is comparatively simple. The major work involved is the calculation of γ_1 and γ_2, and these are found in the course of evaluating $d_{i,i+3}$. The $d_{i,i+4}$ distance can subsequently be found by direct application of equation (74). Alternatively, it can be calculated solely as a function of the interior $d_{i,i+2}$ and $d_{i,i+3}$ distances, without reference to either the virtual bond angles or the Ramachandran angles.

11.4 Numerical Results: Theoretical Near Neighbour Distances.

Using the theoretical results from the previous section, a FORTRAN computer implementation for the calculation of $d_{i,i+2}$, $d_{i,i+3}$ and $d_{i,i+4}$ distances from Ramachandran angle sets $\{\psi_1, \phi_1, \psi_2, \phi_2, \psi_3, \phi_3\}$ has been made. From this program, numerical results were obtained regarding maximum, minimum and mean value statistics, as well as distributions for $d_{i,i+2}$, $d_{i,i+3}$ and $d_{i,i+4}$ distances from various input information regarding Ramachandran angle bounds and secondary structure proportions. The results are expounded in this section, and the possibility of deriving theoretical parameters for use in distance constraint models is discussed. Theoretical maximum bounds for near neighbour distances are obtained for use in the present model.

11.4.1 Minimum and Maximum Distance Parameters.

First, let the near neighbour distance bounds be found with no restriction on the (ψ, ϕ) angles (that is, $-180° \leq \psi_i \leq +180°$ and $-180° \leq \phi_i \leq +180°$). The extreme values for the second neighbour distances can be found from equations (58) and (78), using the set of standard angles given in Table 27. The calculated extreme values are found to be $\theta_{\min} = 75.9°$ and $\theta_{\max} = 147.3°$, with corresponding distances of $min(d_{i,i+2}) = 4.67$ and $max(d_{i,i+2}) = 7.29$. The minimum and maximum distance bounds, plus the mean values of the distributions assuming equal probabilities for each angle, were calculated by the program for distances $d_{i,i+2}$, $d_{i,i+3}$ and $d_{i,i+4}$, and are shown in Table 28.

Distance	Value	(ψ_i, ϕ_i)
$\min(d_{i,i+2})$	4.67	(0,0)
$\max(d_{i,i+2})$	7.29	(180,180)
$\overline{d}_2$	6.13 ± 0.67	
$\min(d_{i,i+3})$	3.72	(0,0,0,180)
$\max(d_{i,i+3})$	10.96	(180,180,180,180)
$\overline{d}_3$	7.95 ± 1.42	
$\min(d_{i,i+4})$	1.53	(0,0,180,180,0,0)
$\max(d_{i,i+4})$	14.54	(180,180,180,180,180,180)
$\overline{d}_4$	9.39 ± 2.23	

Near neighbour distance statistics and their corresponding Ramachandran angles (ψ, ϕ). The possible Ramachandran angles are unrestricted. The mean values and standard deviations were determined by assuming equal probabilities for each pair of angles (ψ, ϕ).

Table 28: Theoretical Near Neighbour Distance Statistics.

The distributions for the $d_{i,i+2}$, $d_{i,i+3}$ and $d_{i,i+4}$ distances were also found for this case of equiprobable (ψ, ϕ) angles with no restrictions on the (ψ, ϕ) angle pairs. Graphs of these distributions are shown in Figures 24 - 26, along with the distributions ob-

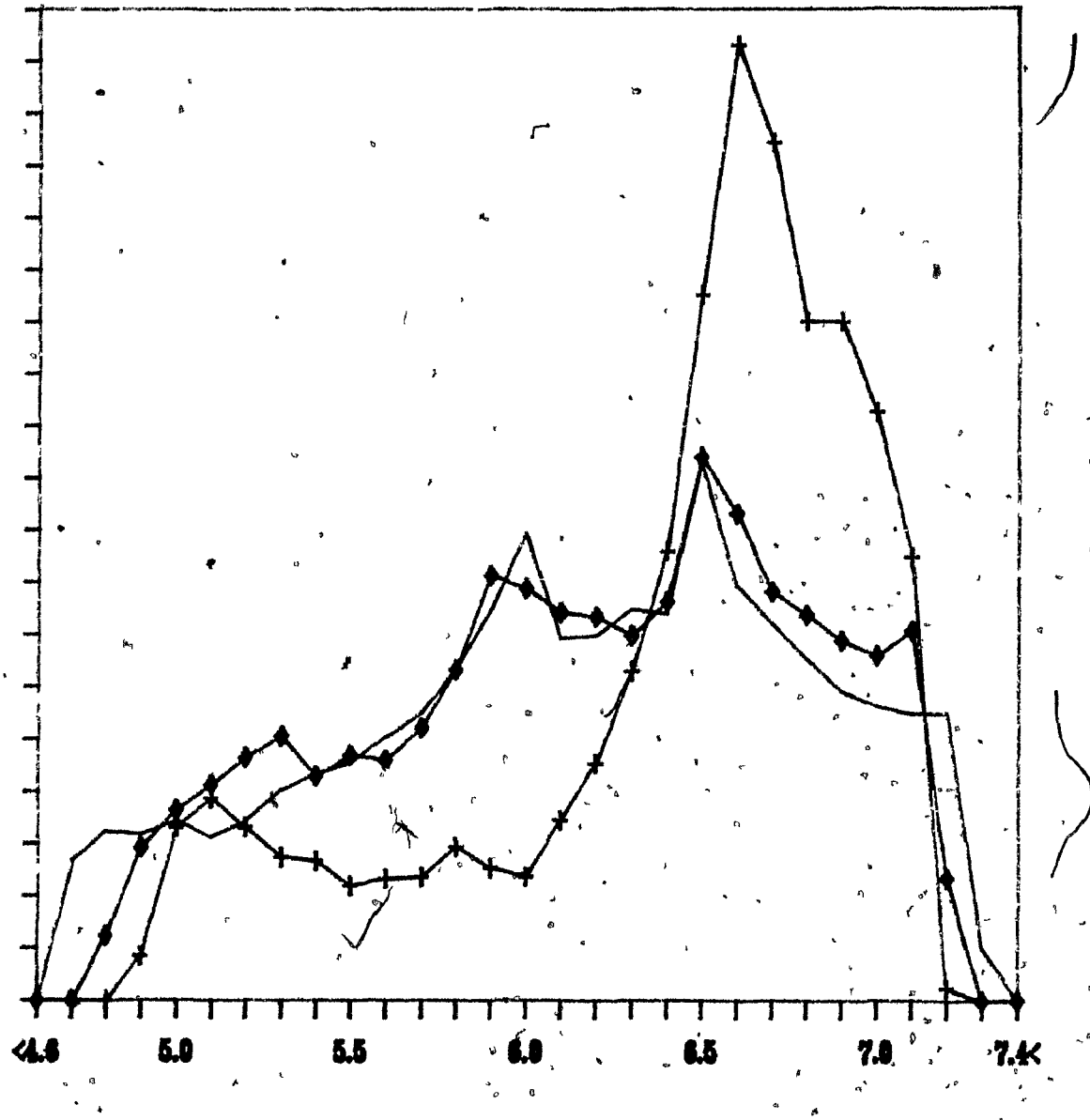
tained when the (ψ, ϕ) angles were restricted to what are termed normal limits and extreme outer limits (Table 30). To produce these distributions, sets of random (ψ, ϕ) angles were repeatedly generated from a uniform distribution of all allowable angles. For this study, 1500 separate random angle sets were generated for each example of Ramachandran limits.

The results shown in Table 28 contain several notable features. One is that all the extrema values correspond to planar configurations of the peptide groups. This result could have been easily predicted for the case of $d_{i,i+2}$, where the equations are found to have local extrema only at $(\psi, \phi) = (0°, 0°)$, $(0°, 180°)$, $(180°, 0°)$ and $(180°, 180°)$ by requiring the first derivatives to simultaneously equal zero. The equations for $d_{i,i+3}$ and $d_{i,i+4}$ in terms of the (ψ, ϕ) angles, however, are found to contain many local extrema, and it is not obvious that the resulting global extrema would be planar configurations.

j	$\min(d_{i,i+j})$	$\min(d_{i,i+j}) - \min(d_{i,i+j-1})$	$\max(d_{i,i+j})$	$\max(d_{i,i+j}) - \max(d_{i,i+j-1})$
1	3.80		3.80	
2	4.67	+0.87	7.29	+3.49
3	3.72	−0.95	10.96	+3.67
4	1.53	−2.19	14.54	+3.58

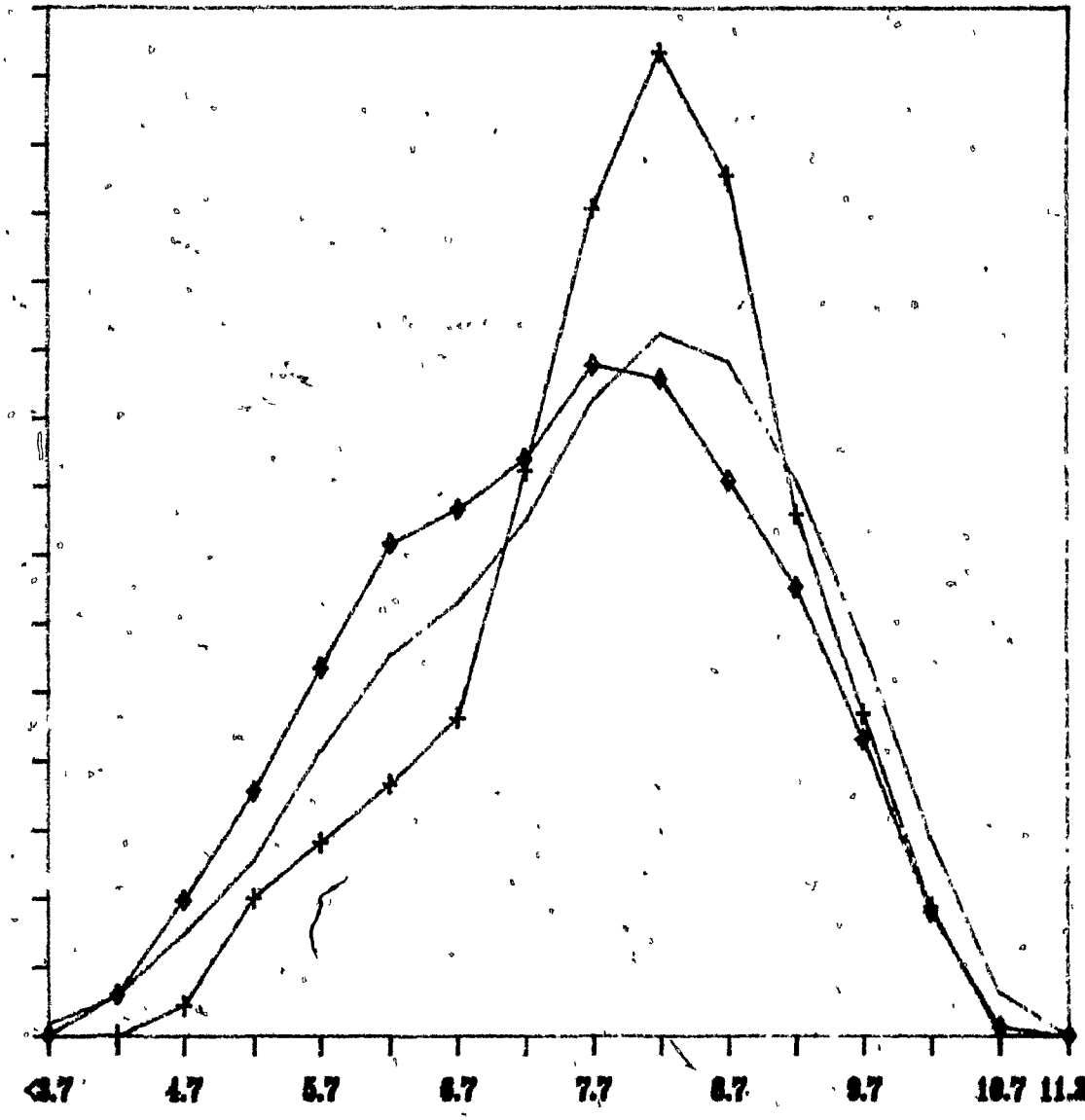
Table 29: Differences in the Near Neighbour Distance Bounds: (ψ, ϕ) Angles Unrestricted.

Other features of the results can be discussed with reference to Table 29. Since no restrictions have been put on the (ψ, ϕ) angles, the $max(d_{i,i+j})$ values are almost as large as $(d_{i,i+j-1} + d_1)$ for every $j = 2, 3, 4$. This result, in which the maximum values for $d_{i,i+j}$ are only slightly smaller than $j \times d_1$, does not correspond to the results of empirical measurements. The anomaly springs from allowing the (ψ, ϕ) angles to be



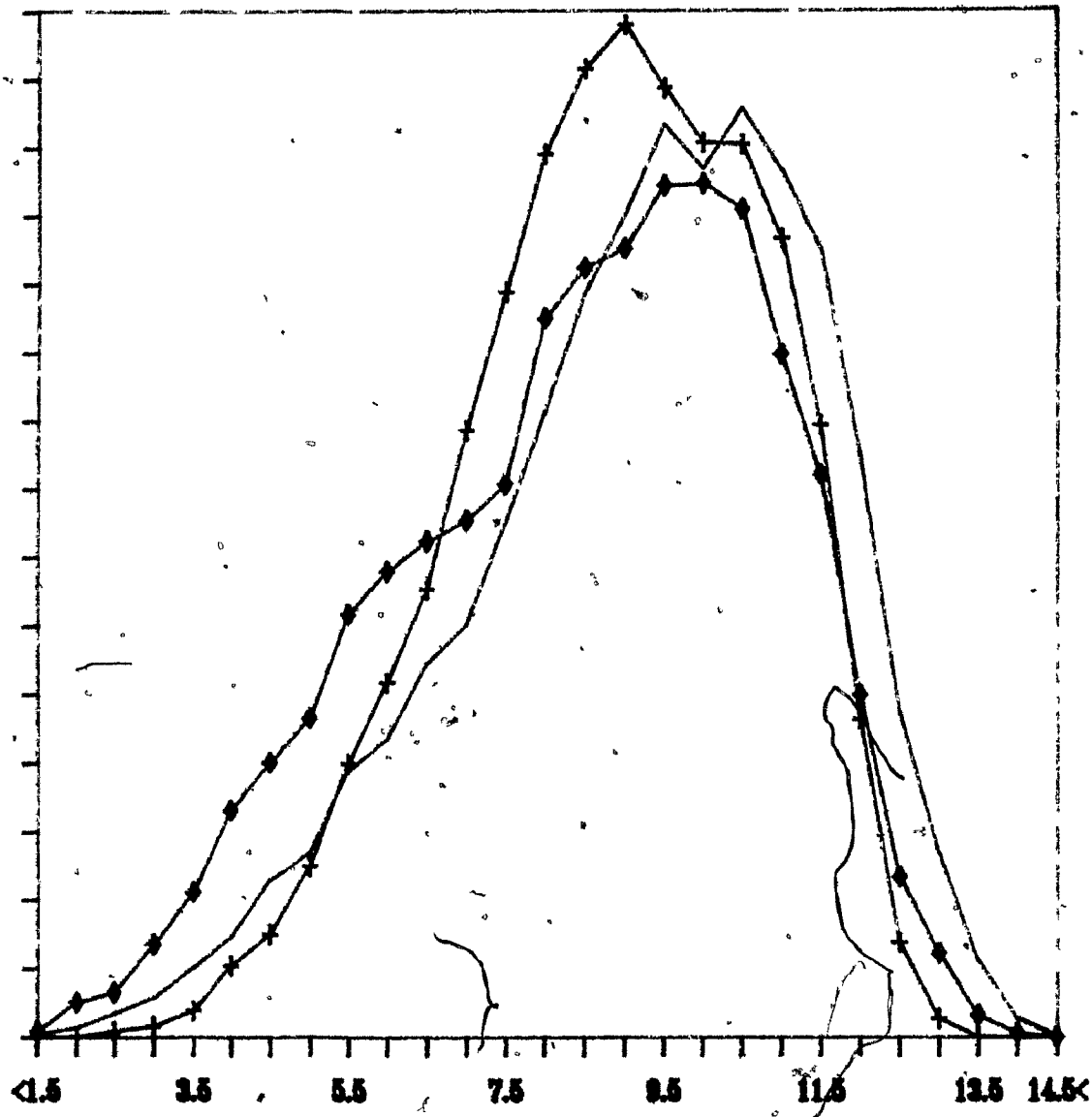
The theoretical probability distributions for second neighbour distances (in Å) are found by generating 1500 separate random (ψ, ϕ) angle sets, assuming equal probability for each pair of angles (ψ, ϕ) within the allowable angle ranges. The ranges used are: (—) no restriction on the angles, (+) normal Ramachandran limits, and (◇) extreme outer limits.

Figure 24: Theoretical Distributions of $d_{i,i+2}$ Distances.



The theoretical probability distributions for third neighbour distances (in Å) are found by generating 1500 separate random (ψ, ϕ) angle sets, assuming equal probability for each pair of angles (ψ, ϕ) within the allowable angle ranges. The ranges used are: (—) no restriction, (+) normal limits, and (◇) extreme outer limits.

Figure 25: Theoretical Distributions of $d_{i,i+3}$ Distances.



The theoretical probability distributions for fourth neighbour distances (in Å) are found by generating 1500 separate random (ψ, ϕ) angle sets, assuming equal probability for each pair of angles (ψ, ϕ) within the allowable angle ranges. The ranges used are: (—) no restriction, (+) normal limits, and ($\diamond$) extreme outer limits.

Figure 26: Theoretical Distributions of $d_{i,i+4}$ Distances.

unrestricted. The residues of actual polypeptides are found to have largely restricted possible (ψ, ϕ) angles ranges, due to steric hindrance and other noncovalent factors [34,90,91,99]. The restriction of (ψ, ϕ) angles to so-called Ramachandran limits is thus necessary in these numerical calculations.

The third and fourth neighbour $min(d_{i,i+3})$ values from Table 29 are smaller than those found empirically. This finding also stems from the lack of restrictions on the (ψ, ϕ) angles. However, the values cannot be rectified in this case by requiring the angles to be within specified Ramachandran limits. Here the phenomenon of steric hindrance also plays a part.

The overall effect of steric hindrance cannot easily be incorporated into the near neighbour distance equations. Whereas steric hindrance will restrict the value ranges for the two angles (ψ_i, ϕ_i) for each residue, the aggregate restriction on the ranges for the four angles $(\psi_i, \phi_i, \psi_{i+1}, \phi_{i+1})$ of adjacent residues will be greater than that of the two residues considered separately. The values calculated using Ramachandran limits will give lower bounds for the minimum near neighbour distance estimates and upper bounds for the maximum near neighbour distance estimates. The calculated estimates for the minimum distances are found to be of minimal value in deriving distance constraint parameters. For this reason the empirical values, not the theoretical values, will be used to derive parameters for $min(d_{i,i+3})$ and $min(d_{i,i+4})$.

A danger of utilizing empirical results for point values (such as maxima or minima), as opposed to aggregated values (such as means or standard deviations) is that point values are highly susceptible to measurement or calculation error, whereas the derivation of aggregated statistics tends to "average out" this type of error. The values used for $min(d_{i,i+3})$ and $min(d_{i,i+4})$ in the present model are found by observation of

the near neighbour distances for a set of twenty proteins with known tertiary structures (the set of proteins from Goel and Yčas [46]), discarding the smallest 1% of the distances as possible measurement errors, and using the next-smallest values as the parameters.

Normal limits:	$\psi = \{90° - 180°, 300° - 320°\}$	$\phi = \{20° - 130°\}$
Extreme outer limits:	$\psi = \{30° - 190°, 290° - 330°\}$	$\phi = \{0° - 140°, 220° - 240°\}$

Table 30: Limits on the Ramachandran Angles (ψ, ϕ).

Distances	No Restriction	Normal (ψ, ϕ)	Extreme (ψ, ϕ)	Empirical
$\min(d_{i,i+2})$	4.67	4.89	4.79	4.65
$\max(d_{i,i+2})$	7.29	7.16	7.20	7.77
$\overline{d}_2$	6.13 ± 0.67	6.42 ± 0.57	6.15 ± 0.63	5.95 ± 0.63
$\min(d_{i,i+3})$	3.72	4.6	4.0	4.33
$\max(d_{i,i+3})$	10.96	10.71	10.82	10.88
$\overline{d}_3$	7.95 ± 1.42	8.10 ± 1.16	7.66 ± 1.40	7.24 ± 1.82
$\min(d_{i,i+4})$	1.53	2.3	1.6	4.39
$\max(d_{i,i+4})$	14.54	13.95	14.37	13.84
$\overline{d}_4$	9.39 ± 2.23	9.01 ± 1.89	8.63 ± 2.37	8.77 ± 2.44

Near neighbour distance statistics with various inputs for the allowable Ramachandran angle sets (ψ, ϕ). Pairwise distance values are given for: (i) no restriction on the (ψ, ϕ) angles, (ii) the angles (ψ, ϕ) restricted to their "normal" values, and (iii) the (ψ, ϕ) restricted to their "extreme outer limit" values. These are compared to the empirical values found by Goel and Yčas [46].

Table 31: Theoretical Near Neighbour Distance Statistics for the Standard Limits on (ψ, ϕ) Angle Sets.

More accurate theoretical predictions for the maximum bounds on $d_{i,i+2}$, $d_{i,i+3}$ and $d_{i,i+4}$ can be found by restricting the Ramachandran angles (ψ, ϕ) to fall within definite limits [92]. These are referred to as the normal and the extreme Ramachan-

Extreme Ramachandran angles:

	Distance	(ψ_i, ϕ_i)
$\max(d_{i,i+3})$	10.82	(189,140,160,140)
$\max(d_{i,i+4})$	14.37	(178,140,159,140,190,220)

Normal Ramachandran angles:

	Distance	(ψ_i, ϕ_i)
$\max(d_{i,i+3})$	10.71	(180,130,151,130)
$\max(d_{i,i+4})$	13.95	(90,130,180,130,164,130)

Table 32: Ramachandran Angle Sets Corresponding to the Theoretical Near Neighbour Distance Maxima.

dran limits (Table 30), and were determined by Ramachandran and his group using graphical methods. The "normal" Ramachandran angle (ψ, ϕ) range corresponds to generally accepted minimum contact distances for short-range non-bonded atoms; the "extreme outer limit" (ψ, ϕ) range corresponds to configurations where the minimum short-contact distances are estimated to be at their absolute minimum values.

For each of the two Ramachandran limit ranges, distance statistics and distributions were calculated for near neighbour residues. The distributions of the second, third and fourth neighbour distances are shown in Figures 24 - 26. In generating these distributions, each angle within the specified Ramachandran limits is given an equal probability of being chosen. Table 31 shows the values obtained for the various near neighbour distance statistics, and compares these theoretical values to the empirical results of Goel and Yčas [46]. Table 32 includes the Ramachandran angle sets that were found to correspond to the $max(d_{i,i+j})$ values.

The near neighbour distance parameters $max(d_{i,i+j})$ are to be estimated theoretically. For the present model, it was decided to use the set of results calculated from the normal Ramachandran angles. Specifically, the model utilizes the following third

and fourth neighbour maximum bounds:

$$\max(d_{i,i+3}) = 10.7$$

$$\max(d_{i,i+4}) = 13.9$$

It would be unrealistic to proceed as if the peptides can assume any (ψ, ϕ) orientation. Whereas it may be possible for a single residue to attain extreme outer limit (ψ, ϕ) values outside the normal range (this is certainly possible for Gly residues), it must be reiterated that the calculated values represent upper bounds for the true values. In general, the dihedral angle ranges will actually be *more* restrictive than the normal Ramachandran limits when a complete triplet or quadruplet of residues is considered as a unit. This is the reason for choosing the bounds obtained from the normal range as model parameters.

11.4.2 Mean Value Parameters: Secondary Structure Distances.

Mean value parameters for near neighbour distances can only be calculated theoretically with any accuracy if some estimate of the secondary structures (*cf.*, Chapter 1.2.9) of the molecule can be determined. Since the majority of the residues in globular proteins are involved in secondary structures which contain distinctive dihedral (ψ, ϕ) angles and hence distinctive near neighbour distances, it would not be valid to calculate theoretical near neighbour mean distances under the assumption of equal probabilities for the (ψ, ϕ) angles, regardless of the (ψ, ϕ) angle limits that are specified. The probability distribution for the (ψ, ϕ) angles must instead depend upon the proportion of each type of secondary structure that exists within the protein to be folded. Therefore, the calculation of near neighbour mean value parameters requires secondary structural

information of two types: the proportion of residues that participate in each type of secondary structure in a given protein and the theoretical near neighbour distances for the residues involved in these structures.

If, in addition to the secondary structure proportions, the actual primary sequence locations of the helical structures could be accurately estimated *a priori* for a protein, the distance constraint model could employ a set of theoretical mean value parameters, where each parameter would be specific to a secondary structure type.

The theoretical calculation of near neighbour distances for the residues within regular secondary structures is easily accomplished by methods of elementary trigonometry. The usual secondary structures, such as the α-helix and β-strand, are helices. These are periodically repeating hydrogen-bonded structures, such that the near neighbour $d_{i,i+j}$ distances are constant for each i. For example, within a regular α-helix, all second neighbour distances are expected to be equal. This implies that with perfect helices the mean value parameters for near neighbour distances have a special significance: the variance of the observed distances will be zero. That is, all near neighbour distances $d_{i,i+j}$ will match the mean value, as is generally the case for first neighbour distances d_1.

Consider a general simple helix containing m residues per turn. Let p denote the pitch, or the depth of one full turn of helix. A general equation expressing the pairwise distance $d_{i,i+j}$ between any two residues can be calculated in terms of the parameters m and p, and can be written as follows:

$$d_{i,i+j}{}^2 = d_1{}^2 \frac{\sin^2\left(\frac{j\pi}{m}\right)}{\sin^2\left(\frac{\pi}{m}\right)} + \frac{p^2}{m^2}\left(j^2 - \frac{\sin^2\left(\frac{j\pi}{m}\right)}{\sin^2\left(\frac{\pi}{m}\right)}\right). \qquad (93)$$

Note that the calculated pairwise distances in this case depend upon the primary

sequence separation j of the two residues, but do not depend upon their specific positions i within the helix. From the general equation (93), particular expressions for the pairwise distances $d_{i,i+2}$, $d_{i,i+3}$ and $d_{i,i+4}$ can be calculated:

$$d_{i,i+2}{}^2 = 4\left[d_1{}^2\cos^2\left(\frac{\pi}{m}\right) + \frac{p^2}{m^2}\sin^2\left(\frac{\pi}{m}\right)\right] \tag{94}$$

$$d_{i,i+3}{}^2 = d_1{}^2\left[4\cos^2\left(\frac{\pi}{m}\right) - 1\right]^2 + 8\frac{p^2}{m^2}\sin^2\left(\frac{\pi}{m}\right)\left[2\cos^2\left(\frac{\pi}{m}\right) + 1\right] \tag{95}$$

$$d_{i,i+4}{}^2 = 16\left[d_1{}^2\cos^2\left(\frac{\pi}{m}\right)\cos^2\left(\frac{2\pi}{m}\right) + \frac{p^2}{m^2}\left(1 - \cos^2\left(\frac{\pi}{m}\right)\cos^2\left(\frac{2\pi}{m}\right)\right)\right]. \tag{96}$$

The α-helix ideally will have 3.61 residues per turn and a pitch of $p = 5.41$ [92]. However, perfect α-helices are rarely seen in globular proteins. Most helices are found to be twisted or tilted from the vertical, not fully hydrogen-bonded, or otherwise distorted. As an example, statistics for the secondary structures of bovine pancreatic trypsin inhibitor are shown in Table 33. These statistics show a larger variability than would be expected within the classes $d_{i,i+2}$, $d_{i,i+3}$ and $d_{i,i+4}$. Also their mean values do not conform exactly to the theoretically expected averages calculated below (Table 34). These statistical results are typical of the failure of *in vivo* secondary structures to form completely. Therefore, although secondary structures yield particularly simple results for near neighbour distances theoretically, the values obtained are not readily applicable in practice.

For the fully extended chain structure, the β-strand, the axial distance between adjacent residues is estimated from theoretical hydrogen bonding studies to ideally have a value of 3.5 Å, with a helical radius of 1.0 Å. This structure yields a pitch of $p = 6.95$, with $m = 2.00$ residues per turn [92].

Table 34 gives a set of pairwise distances of the form $d_{i,i+j}$ for near neighbour residues within regular secondary structures: α-helix, β-strand and 3_{10}-helix. The

	α-helix (residues 47 - 56)	antiparallel β-strands (residues 16 - 25 vs 28 - 36)
$\overline{d}_2$	5.56 ± 0.37	6.61 ± 0.42
$(\min(d_{i,i+2}), \max(d_{i,i+2}))$	$(5.15, 6.09)$	$(5.83, 7.13)$
$\overline{d}_3$	5.26 ± 0.56	9.78 ± 0.45
$(\min(d_{i,i+3}), \max(d_{i,i+3}))$	$(4.58, 6.04)$	$(9.12, 10.66)$
$\overline{d}_4$	5.79 ± 0.70	12.84 ± 0.62
$(\min(d_{i,i+4}), \max(d_{i,i+4}))$	$(4.55, 6.56)$	$(11.36, 13.85)$

Near neighbour distance statistics within the secondary structures found in BPTI. Shown are the mean, standard deviation, minimum value and maximum value for each of the second, third and fourth neighbour distances for the residues involved in the secondary structures of BPTI. Residue locations for the secondary structures are from Deisenhofer and Steigemann [32].

Table 33: Secondary Structure Examples: BPTI.

values are calculated from equations (94), (95) and (96)

	α-helix	β-strand	3_{10}-helix
m	3.61	2.00	3.0
p	5.41	6.95	6.0
$d_{i,i+2}$	5.41	6.95	5.14
$d_{i,i+3}$	5.06	10.54	6.00
$d_{i,i+4}$	6.18	13.90	8.63

Table 34: Theoretical Near Neighbour Distances: Secondary Structures.

At the present time, it is difficult to predict *a priori* the proportions of secondary structures that will be contained in a particular protein, unless its evolutionary family is known. It is has been estimated from empirical studies that globular proteins contain on average nearly 40% α-helix residues, 15% β-strand residues and 25% of the residues in reverse turns [20]. Unfortunately, these proportions vary greatly from protein to

protein, and are not correlated with the size of a protein. Also, many examples of secondary structure within proteins are considerably distorted, and it is often difficult to determine whether or not a particular residue should be included as an element of a discerned helix [56].

Due to the difficulty in predicting the proportions of the various secondary structures in a protein to be folded, mean value parameters for near neighbour distances are not employed in the present model. Due to the difficulty in predicting the locations of the secondary structures and the variability in near neighbour distances observed within the helices of globular proteins, predictions for the locations of protein secondary structures are also not used in the model.

Statistical secondary structure prediction techniques (*cf.*, Chou and Fasman [20]) will undoubtedly become more reliable in the future. In this event, general distance constraints for near neighbour mean values or specific distance constraints for the mean values of residues believed to be involved in secondary structures could be accommodated by the model. The general mean value parameters could be found by incorporating the proportions of each type of secondary structure into a probability distribution of (ψ, ϕ) angles. The mean distance corresponding to this probability distribution would then be calculated by the method of the previous section. Specific mean value parameters used for each separate type of secondary structure are much more easily implemented; they would simply be the values given in Table 34. That is, each pair of residues that are predicted to exist within a given secondary structure could be constrained to lie at an exact pairwise distance, given by the parameters in Table 34.

References

[1] *Protein Data Bank.* Department of Chemistry, Brookhaven National Laboratory, Associated Universities Inc., Upton L. I., New York 11973.

[2] D. Amir and E. Haas. Determination of intramolecular distance distributions in a globular protein by nonradiative excitation energy transfer measurements. *Biopolymers*, 25:235–240, 1986.

[3] C.B. Anfinsen, E. Haber, M. Sela, and F.H. White Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proceedings of the National Academy of Sciences USA*, 47:1309–1314, 1961.

[4] C.B. Anfinsen and H.A. Scheraga. Experimental and theoretical aspects of protein folding. *Advancements in Protein Chemistry*, 29:205–300, 1975.

[5] T. Ashida, I. Tanaka, T. Yamane, and M. Kakudo. .. In R. Srinivasan, editor, *Biomolecular Structure, Conformation, Function and Evolution.*, pages 607–620, Pergamon, New York, 1980.

[6] M. Bajaj and T. Blundell. Evolution and the tertiary structure of proteins. *Annual Review of Biophysics and Bioengineering*, 13:453–492, 1984.

[7] M. Billeter, W. Braun, and K. Wüthrich. Sequential resonance assignments in protein ^1H nuclear magnetic resonance spectra: Computation of sterically allowed proton-proton distances and statistical analysis of proton-proton distances in single crystal protein conformations. *Journal of Molecular Biology*, 155:321–346, 1982.

[8] C.C.F. Blake, D.F. Koenig, G.A. Mair, A.C.T. North; D.C. Phillips, and V.R. Sarma. Structure of hen egg-white lysozyme. *Nature*, 206:757–763, 1965.

[9] L.M. Blumenthal. *Theory and Applications of Distance Geometry.* Chelsea, New York, 1970.

[10] W. Braun, C. Bosch, L.R. Brown, N. Gō, and K. Wüthrich. Combined use of proton-proton Overhauser enhancements and a distance geometry algorithm for determination of polypeptide conformations. Application to micelle-bound glucagon. *Biochimica et Biophysica Acta*, 667:377–396, 1981.

[11] J.M. Breechem and L. Brand. Time-resolved fluorescence of proteins. *Annual Review of Biochemistry*, 54:43–71, 1985.

[12] H. Bremermann. A method of unconstrained global optimization. *Mathematical Biosciences*, 9:1–15, 1970.

[13] S.K. Burley and G.A. Petsko. Aromatic-aromatic interaction: A mechanism of protein structure stabilization. *Science*, 229:23–28, 1985.

[14] P. Cariani and N.S. Goel. On the computation of the tertiary structure of globular proteins. IV. Use of secondary structure information. *Bulletin of Mathematical Biology*, 47:367–407, 1985.

[15] J.C. Cavadore. *Polycondensation d'α-amino acides en milieu aqueux.* PhD thesis, Université des Sciences et Techniques du Languedoc, Montpellier, 1971.

[16] M. Charton. Protein folding and the genetic code: An alternative quantitative model. *Journal of Theoretical Biology*, 91:115–123, 1981.

[17] M. Charton and B.I. Charton. The structural dependence of amino acid hydrophobicity parameters. *Journal of Theoretical Biology*, 99:629–644, 1982

[18] C. Chothia. The nature of the accessible and buried surfaces in proteins. *Journal of Molecular Biology*, 105:1–14, 1976.

[19] P.Y. Chou and G.D. Fasman. Empirical predictions of protein conformation. *Annual Review of Biochemistry*, 47:251–276, 1978.

[20] P.Y. Chou and G.D. Fasman. Prediction of protein conformation. *Biochemistry*, 13:222–245, 1974.

[21] P.Y. Chou and G.D. Fasman. Prediction of the secondary structure of proteins from their amino acid sequence. *Advancements in Enzymology*, 47:45–148, 1978.

[22] F.E. Cohen and M.J.E. Sternberg. On the prediction of protein structure: The significance of the root-mean-square deviation. *Journal of Molecular Biology*, 138:321–333, 1980.

[23] F.E. Cohen and M.J.E. Sternberg. On the use of chemically derived distance constraints in the prediction of protein structure with myoglobin as an example. *Journal of Molecular Biology*, 137:9–22, 1980.

[24] T.F. Coleman and A.R. Conn. Nonlinear programming via an exact penalty function: Global analysis. *Mathematical Programming*, 24:137–161, 1982.

[25] R.B. Corey and L. Pauling. .. *Proceedings of the Royal Society of London, Series B. Biological Sciences*, 141:10–20, 1953.

[26] A.J. Corrigan and P.C. Huang. A BASIC microcomputer program for plotting the secondary structure of proteins. *Computer Programs in Biomedicine*, 15:163–168, 1982.

[27] T.E. Creighton. Experimental studies of protein folding and unfolding. *Progress in Biophysics and Molecular Biology*, 33:231–297, 1978.

[28] T.E. Creighton. *Proteins: Structures and molecular principles.* W.H. Freeman and Company, New York, 1983.

[29] G.M. Crippen. Global optimization and polypeptide conformation. *Journal of Computational Physics*, 18:224–231, 1975.

[30] G.M. Crippen. A novel approach to the calculation of conformation: Distance geometry. *Journal of Computational Physics*, 24:96–107, 1977.

[31] J.F. Danielli. Protein films at the oil-water interface. *Cold Spring Harbour Symposia on Quantitative Biology*, 6:190–195, 1938.

[32] J. Deisenhofer and W. Steigemann. Crystallographic refinement of the structure of bovine pancreatic trypsin inhibitor at 1.5 Å resolution. *Acta Crystallographa*, B31:238–250, 1975.

[33] R.S. Dembo and T. Steihaug. Truncated-Newton algorithms for large-scale unconstrained optimization. *Mathematical Programming*, 26:190–212, 1983.

[34] R.E. Dickerson and I. Geis. *The Structure and Action of Proteins.* Benjamin/Cummings Publishing Company, Menlo Park, California, 1969.

[35] J. Donohue. Hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences USA*, 39:470–478, 1953.

[36] R.F. Doolittle. Proteins. *Scientific American*, 253:88–99, 1985.

[37] A.V. Finkelstein. Theory of protein molecule self-organization. III. A calculating method for the probabilities of the secondary structure formation in an unfolded polypeptide chain. *Biopolymers*, 16:525–529, 1977.

[38] A.V. Finkelstein and O.B. Ptitsyn. Theory of protein molecule self-organization. I. Thermodynamic parameters of local secondary structures in unfolded protein chain. *Biopolymers*, 16:469–495, 1977.

[39] R. Fletcher. *Practical Methods of Optimization*. Volume 1, John Wiley & Sons, Ltd., New York, 1980.

[40] E.A.D. Foster. A distance constraint model for the prediction of protein tertiary structure. *The Canadian Journal of Fisheries and Aquatic Sciences*, 43:1035–1044, 1986.

[41] E.A.D. Foster and P.F. O'Neill. Modelling the tertiary structure of globular proteins using mathematical programming. To be submitted.

[42] P.E. Gill and W. Murray. Methods for large-scale linearly constrained problems. In P.E. Gill and W. Murray, editors, *Numerical Methods for Constrained Optimization*, pages 93–148, Academic Press, New York, 1974.

[43] P.E. Gill and W. Murray. *Numerical Methods for Constrained Optimization*. Academic Press, 1974.

[44] D. Givol, F. DeLorenzo, R.F. Goldberger, and C.B. Anfinsen. Disulfide interchange and the three-dimensional structure of proteins. *Proceedings of the National Academy of Sciences USA*, 53:676–684, 1965.

[45] N.S. Goel, B. Rouyanian, and M. Sanati. On the computation of the tertiary structure of globular proteins. III. Inter-residue distances and computed structures. *Journal of Theoretical Biology*, 99:705–757, 1982.

[46] N.S. Goel and M. Yčas. On the computation of the tertiary structure of globular proteins. II. *Journal of Theoretical Biology*, 77:253–305, 1979

[47] A.A. Goldstein. On steepest descent. *SIAM Journal on Control and Optimization*, 3:147–151, 1965.

[48] B. Gutte and R.B. Merrifield. The total synthesis of an enzyme with ribonuclease A activity. *Journal of the American Chemical Society*, 91:501–506, 1969.

[49] E Haber and C.B. Anfinsen. Side-chain interactions governing the pairing of half-cystine residues in ribonuclease. *Journal of Biological Chemistry*, 237:1839–1844, 1962.

[50] T.F. Havel, G.M. Crippen, and I.D. Kuntz. Effects of distance constraints on macromolecular conformation. II. Simulation of experimental results and theoretical predictions. *Biopolymers*, 18:73–81, 1979.

[51] T.F. Havel, I.D. Kuntz, and G.M. Crippen. The theory and practice of distance geometry. *Bulletin of Mathematical Biology*, 45:665–720, 1983.

[52] T.F. Havel and K. Wüthrich. A distance geometry program for determining the

structures of small proteins and other macromolecules for nuclear magnetic resonance measurements of intermolecular ^1H - ^1H proximities in solutions. *Bulletin of Mathematical Biology*, 46:673–698, 1984.

[53] J.R. Herriott, K.D. Watenpaugh, L.C. Sieker, and L.H. Jensen. Sequence of rubredoxin by X-ray diffraction. *Journal of Molecular Biology*, 80:423–432, 1973.

[54] R. Hirschmann, R.F. Nutt, D.F. Veber, R.A. Vitali, S.L. Varga, T.A. Jacob, F.W. Holly, and R.G. Denkewalter. Studies on the total synthesis of an enzyme. V. The preparation of enzymically active material. *Journal of the American Chemical Society*, 91:507–508, 1969.

[55] D.D. Jones. Amino acid properties and side-chain orientation in proteins: A cross-correlation approach. *Journal of Theoretical Biology*, 50:167–183, 1975.

[56] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2638, 1983.

[57] M. Karplus and J.A. McCammon. The dynamics of proteins. *Scientific American*, 254:42–51, 1986.

[58] W. Kauzmann. Some factors in the interpretation of protein denaturation. *Advances in Protein Chemistry*, 14:1–63, 1959.

[59] A.A. Kossiakoff. Protein dynamics investigated by the neutron diffraction-hydrogen exchange technique. *Nature*, 296:713–721, 1982.

[60] D. Kotelchuck and H.A. Scheraga. The influence of short-range interactions on

protein conformation. II. A model for predicting the α-helical regions of proteins. *Proceedings of the National Academy of Science USA*, 62:14–21, 1969.

[61] I. Kuntz. An approach to the tertiary structure of globular proteins. *Journal of the American Chemical Society*, 97:4362–4366, 1975.

[62] I. Kuntz, G. Crippen, and P. Kollman. Applications of distance geometry to protein tertiary structure calculations. *Biopolymers*, 18:939–957, 1979.

[63] I. Kuntz, G. Crippen, P. Kollman, and D. Kimmelman. Calculation of protein tertiary structure. *Journal of Molecular Biology*, 106:983–994, 1976.

[64] E.Q. Lawson, A.J. Sadler, D. Harmatz, D.T. Brandau, R. Micanovic, R.D. MacElroy, and C.R. Middaugh. A simple experimental model for hydrophobic interactions in proteins. *Journal of Biological Chemistry*, 259:2910–2912, 1984.

[65] B. Lee and F.M. Richards. The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*, 55:379–382, 1971.

[66] M. Levitt. Conformational preferences of amino acids in globular proteins. *Biochemistry*, 17:4277–4285, 1978.

[67] M. Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology*, 104:59–107, 1976.

[68] M. Levitt and A. Warshel. Computer-simulation of protein folding. *Nature*, 253:694–698, 1975.

[69] V.I. Lim. Algorithms for prediction of α-helical and β-structural regions in globular proteins. *Journal of Molecular Biology*, 88:873–894, 1974.

[70] A.H. Louie and R.L. Somorjai. Differential geometry of proteins: A structural and dynamical representation of patterns. *Journal of Theoretical Biology*, 98:189–209, 1982.

[71] A.H. Louie and R.L. Somorjai. Differential geometry of proteins: Helical approximations. *Journal of Molecular Biology*, 168:143–162, 1983.

[72] P. Manavalan and P.K. Ponnuswamy. A study of the preferred environment of amino acid residues in globular proteins. *Archives of Biochemistry and Biophysics*, 184:476–487, 1977.

[73] R.E. Marsh and J. Donohue. Crystal structure studies of amino acids and peptides. *Advances in Protein Chemistry*, 22:235–256, 1967.

[74] H. Meirovitch, S. Rackovsky, and H.A. Scheraga. Empirical studies of hydrophobicity. I. Effects of protein size on the hydrophobic behavior of amino acids. *Macromolecules*, 13:1398–1405, 1979.

[75] H. Meirovitch and H.A. Scheraga. Empirical studies of hydrophobicity. II. Distribution of the hydrophobic, hydrophilic, neutral and ambivalent amino acids in the interior and exterior layers of native proteins. *Macromolecules*, 13:1406–1414, 1980.

[76] H. Meirovitch and H.A. Scheraga. Empirical studies of hydrophobicity. III. Radial distribution of clusters of hydrophobic and hydrophilic amino acids. *Macromolecules*, 14:340–345, 1981.

[77] J. Moult, A. Yonath, W. Traub, A. Smilansky, A. Podjarny, D. Rabinovich, and A. Saya. The structure of triclinic lysozyme at 2.5 Å resolution. *Journal of*

Molecular Biology, 100:179–195, 1976.

[78] G. Nemethy and H.A. Scheraga. Protein folding. *Quarterly Review of Biophysics*, 10:239–352, 1977.

[79] K. Nishikawa, F.A. Momany, and H.A. Scheraga. Low-energy structures of two dipeptides and their relationship to bend configurations. *Macromolecules*, 6:797–806, 1974.

[80] Y. Nozaki and C. Tanford. The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. *Journal of Biological Chemistry*, 246:2211–2217, 1971.

[81] L. Pauling and R.B. Corey. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proceedings of the National Academy of Sciences USA*, 37:235–241, 1951.

[82] L. Pauling and R.B. Corey. Configurations of polypeptide chains with favored orientations around single bounds: Two new pleated sheets. *Proceedings of the National Academy of Science USA*, 37:729–740, 1951.

[83] L. Pauling, R.B. Corey, and H.R. Branson. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences USA*, 37:205–211, 1951.

[84] M. Prabhakaran and P.K. Ponnuswamy. Spatial assignment of amino acid residues in globular proteins: An approach from information theory. *Journal of Theoretical Biology*, 87:623–637, 1980.

[85] O.B. Ptitsyn and A.A. Rashyn. A model of myoglobin self-organization. *Biophysical Chemistry*, 3:1–20, 1975.

[86] S. Rackovsky and H.A. Scheraga. Differential geometry and polymer configuration. I. Comparison of protein conformations. *Macromolecules*, 11:1168–1174, 1978.

[87] S. Rackovsky and H.A. Scheraga. Differential geometry and polymer configuration. II. Development of a conformational distance function. *Macromolecules*, 13:1440–1443, 1980.

[88] S. Rackovsky and H.A. Scheraga. Differential geometry and polymer configuration. III. Single-site and near-neighbour distributions and nucleation of protein folding. *Macromolecules*, 14:1250–1269, 1981.

[89] G.N. Ramachandran, A.S. Kolaskar, C. Ramakrishnan, and V. Sasisekharan. The mean geometry of the peptide unit from crystal structure data. *Biochimica et Biophysica Acta*, 359:298–302, 1974.

[90] G.N. Ramachandran, C. Ramakrishnan, and V. Sasisekharan. Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, 7:95–99, 1963.

[91] G.N. Ramachandran and V. Sasisekharan. Conformation of polypeptides and proteins. *Advancements in Protein Chemistry*, 23:283–437, 1968.

[92] C. Ramakrishnan and G.N. Ramachandran. Stereochemical criteria for polypeptide and protein chain conformations. II. Allowed conformations for a pair of peptide units. *Biophysical Journal*, 5:909–933, 1965.

[93] S.J. Remington and B.W. Matthews. A systematic approach to the comparison of protein structures. *Journal of Molecular Biology*, 140:77–99, 1980.

[94] F.M. Richards. The interpretation of protein structures: Total volume, group volume distributions and packing density. *Journal of Molecular Biology*, 82:1 14, 1974.

[95] R. Rosen. On control and optimal control in biodynamic systems. *Bulletin of Mathematical Biology*, 42.889–897, 1980.

[96] R. Rosen. Protein folding: A prototype for control of complex systems. *International Journal of Systems Science*, 11:527–540, 1980.

[97] M.G. Rossmann and P. Argos. Protein folding. *Annual Review of Biochemistry*, 50:497–532, 1981.

[98] M. Sanati. *On the computation of the tertiary structure of globular proteins.* PhD thesis, State University of New York at Binghamton, 1980.

[99] V. Sasisekharan. Stereochemical criteria for polypeptide and protein structures In N. Ramanathan, editor, *Collagen*, John Wiley & Sons, Ltd., New York, 1960.

[100] H.A. Scheraga. Recent progress in the theoretical treatment of protein folding *Biopolymers*, 22:1–14, 1983.

[101] G. E. Schulz and R. H. Schirmer. *Principles of Protein Structure.* Springer Verlag, New York, 1979.

[102] A. Shrake and J.A. Rupley. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Journal of Molecular Biology*, 79:351 371, 1973.

[103] M.J. Sippl and H.A. Scheraga. Solution of the embedding problem and decomposition of symmetric matrices. *Proceedings of the National Academy of Sciences USA*, 82:2197–2201, 1985.

[104] J. Skolnick. Theory of the helix-coil transition in single-chain polypeptides with interhelical contacts. The broken α-helical hairpin model. *Macromolecules*, 18:1073–1083, 1985.

[105] F.B. Straub. SH groups and SS bridges in the structure of enzymes. In *Proceedings of the 7th International Conference on Biochemistry*, pages 42–50, 1967.

[106] L. Stryer. *Biochemistry*. W.H. Freeman and Company, San Francisco, 1981.

[107] S. Tanaka and H.A. Scheraga. Statistical mechanical treatment of protein conformation. I. Conformational properties of amino acids in proteins. *Macromolecules*, 9:142–159, 1976.

[108] S. Tanaka and H.A. Scheraga. Statistical mechanical treatment of protein conformation. II. A three-state model for specific sequence copolymers of amino acids. *Macromolecules*, 9:159–167, 1976.

[109] S. Tanaka and H.A. Scheraga. Statistical mechanical treatment of protein conformation. III. Prediction of protein conformation based on a three-state model. *Macromolecules*, 9:167–182, 1976.

[110] S. Tanaka and H.A. Scheraga. Statistical mechanical treatment of protein conformation. IV. A four-state model for specific-sequence copolymers of amino acids. *Macromolecules*, 9:812–833, 1976.

[111] S. Tanaka and H.A. Scheraga. Statistical mechanical treatment of protein conformation. V. A multi-state model for specific sequence copolymers of amino acids. *Macromolecules*, 10:9–20, 1977.

[112] S. Tanaka and H.A. Scheraga. Statistical mechanical treatment of protein conformation. VI. Elimination of empirical rules for prediction by use of a high-order probability. Correlation between amino acid sequences and conformations for homologous neurotoxin proteins. *Macromolecules*, 10:305–316, 1977.

[113] K.A. Thomas and A.N. Schechter. Protein folding. In R.F. Goldberger, editor, *Biological Regulation and Development*, pages 43–100, Plenum Press, New York, 1980.

[114] J. M. Thornton. Disulphide bridges in globular proteins. *Journal of Molecular Biology*, 151:261–287, 1981.

[115] C.M. Venkatachalam. Stereochemical criteria for polypeptides and proteins. V Conformation of a system of three linked peptide chains. *Biopolymers*, 6:1425 1436, 1968.

[116] G. Wagner and K. Wüthrich. Sequential resonance assignments in protein ^1H nuclear magnetic resonance spectra: Basic pancreatic trypsin inhibitor. *Journal of Molecular Biology*, 155:347–366, 1982.

[117] H. Wako and H.A. Scheraga. Distance-constraint approach to protein folding. I. Statistical analysis of protein conformations in terms of distances between residues. *Journal of Protein Chemistry*, 1:5–45, 1982.

[118] H. Wako and H.A. Scheraga. Distance-constraint approach to protein folding. II. Prediction of three-dimensional structure of bovine pancreatic trypsin inhibitor. *Journal of Protein Chemistry*, 1:85–117, 1982.

[119] H. Wako and H.A. Scheraga. On the use of distance constraints to fold a protein. *Macromolecules*, 14:961–969, 1981.

[120] H. Wako and H.A. Scheraga. Visualization of the nature of protein folding by a study of a distance constraint approach in two-dimensional models. *Biopolymers*, 21:611–632, 1982.

[121] K.D. Watenpaugh, L.C. Sieker, and L.H. Jensen. Crystallographic refinement of rubredoxin at 1.2 Å resolution. *Journal of Molecular Biology*, 138:615–633, 1980.

[122] J.D. Watson. *Molecular Biology of the Gene.* Benjamin, New York, 1965.

[123] G.R. Welch, B. Somogyi, and S. Damjanovich. The role of protein fluctuations in enzyme action: A review. *Progress in Biophysics and Molecular Biology*, 39:109–146, 1982.

[124] D.H. Wertz and H.A. Scheraga. Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. *Macromolecules*, 11:9–15, 1978.

[125] F. Wold. Bifunctional reagents. In *Methods in Enzymology Volume 25: Enzyme Structure, Part B.*, chapter 57, Academic Press, New York, 1972.

[126] F. Wold. *In vivo* chemical modification of proteins (post-translational modifications). *Annual Review of Biochemistry*, 50:783–814, 1981.

[127] K. Wüthrich, G. Wider, G. Wagner, and W. Braun. Sequential resonance assignments as a basis for determination of spatial protein structures by high resolution proton nuclear magnetic resonance. *Journal of Molecular Biology*, 155:311–319, 1982.

[128] M. Yčas, N.S. Goel, and J.W. Jacobsen. On the computation of the tertiary structure of globular proteins. *Journal of Theoretical Biology*, 72:443–457, 1978.

[129] S.S. Zimmerman, M.S. Pottle, G. Nemethy, and H.A. Scheraga. Conformational analysis of the twenty naturally occurring amino acids residues using ECEPP. *Macromolecules*, 10:1–9, 1977.