UNIVERSITY
of
GLASGOW

Biochemistry
Ph.D. Thesis

Hydrogen bonding and the stability of the polypeptide backbone

Peter Hugh Maccallum

Submitted for the degree of

Doctor of Philosophy

If the Lord Almighty had consulted me before embarking on creation,
I would have recommended something simpler.
Alphonso X, King of Castille and León, 1263-1297

Acknowledgements

Many people have affected the development of this thesis; I appreciate all their effort and advice.

In the Department of Biochemistry: My supervisor James Milner-White, for allowing me the freedom to analyse proteins from a wide range of viewpoints while discussing every stage of my research, and John Coggins and David Leader for reminding me that proteins can exist outside computers (sometimes),

In the Department of Computing Science: John Patterson and Ron Poet, for providing algorithms, coding advice, and access to some excellent computer facilities, and especially James Logie for teaching me everything I needed to know about Unix,

At the University of North Carolina at Chapel Hill: Fred Brooks, Bill Wright, and Larry Bergman, for giving me access to some of the best toys in the world and showing what dialogue between computer scientists and biochemists can achieve,

At Edinburgh Parallel Computing Centre: my excellent managers Alan Simpson, Steve Booth and Neil MacDonald, for improving my data handling and document preparation skills just when they were needed,

And at Home: Sally, Eilidh, and Sophie for putting up with my long vampire impersonation while this thesis evolved.

Hydrogen bonding and the stability of the polypeptide backbone

by
Peter Hugh Maccallum

Submitted to the Department of Biochemistry
on 28 September 1996
for the degree of
Doctor of Philosophy

Abstract

The tertiary structures of globular proteins are crucial in determining reactivity and specificity as biological catalysts and signalling systems. The rules determining the final fold of a protein are still unknown, but some progress has been made in defining tertiary structure in terms of the secondary structure, the conformation of the polypeptide chain. Perhaps surprisingly, not all of the conformational properties of this backbone are known, and several new approaches to studying these are described.

Most studies of peptide structure have focused on hydrogen bonding, and this is used as a starting point for this study. Different descriptions of the hydrogen bond, from geometric rules to *ab initio* calculations, are considered, and an approach based on analysing contributions of individual polar groups to the potential energy using semi empirical Lennard-Jones calculations is chosen on grounds of accuracy, flexibility, and ease of calculation.

Using this approach, it is shown that electrostatic interactions between main chain atoms stabilise the right handed twist found in β-strands and similar interactions between main-chain atoms not hydrogen bonded to each other influence the geometries of hydrogen bonds in α-helices and β-sheets. A role for water and tertiary hydrogen bonds in determining backbone conformation is suggested.

The same technique makes it possible to investigate interatomic repulsions as well as attractions. A detailed analysis of the attractions and repulsions in an idealised polypeptide explains many of the features of helical structures in proteins, and suggests a hitherto unexpected *directional* helix forming pathway, which is supported by a range of kinetic and structural data.

Software for automated searching of a hydrogen bond database is developed, and used to identify hydrogen bonded rings formed by amide side chains and main chain peptides. Integrating the database with novel visualisation techniques allows a previously unidentified property of beta sheets, the hydrophobic ridge, to be detected.

A range of different computational approaches was used surging this research, from molecular modelling to database searching. Several pieces of software were developed, and these are described together with some observations about the types of software and working environments which were found to be useful in structural biochemistry, and what types of software technology could be developed to make this task easier.

Thesis Supervisor: E.James Milner-White
Title: Doctor

Contents

List of Figures

Chapter 1

Introduction

Summary

Hydrogen bonding is identified as the major structural stabilisation in proteins. Previous extensive reviews based on hydrogen bonding patterns have provided useful insights into many aspects of protein structure. In particular, they have been used to identify properties such as the role of saturation of hydrogen bonding potential as an important constraint on all adopted conformations, and to show the value of approximate energy assignments in identifying stabilisation in loops and strained regions of proteins.

This work extends the use of approximate energy calculations to study the conformational preferences of side chains and the polypeptide backbone, providing evidence that:

(1) potential energy functions based on atom-centred potentials without explicit directional constraints are effective in describing hydrogen bonding as judged by the match of a new study of potential energy functions with observed spatial distribution of hydrogen bonding partners,

(2) β twist is an intrinsic property of all polypeptide backbones, with intra-strand electrostatics and mainchain/mainchain and mainchain/solvent steric effects possibly more significant in this respect than sidechain effects,

(3) the extended 3_{10} helix and π helix are not stable secondary structures, the 3_{10} helix because it experiences no barrier to folding to the more stable alpha helix and the π helix because it is sterically unfavourable. These conclusions are based on a set of potential energy calculations on long polypeptides, rather than on N-acetylated N'-methylated amino

acids, which also serve to explain the observed distributions of secondary structure ϕ/ψ distributions,

(4) helix growth is likely to be stepwise, because of an observed potential energy barrier to concerted folding, and directional (in the C' to N' direction) because of a steric/electrostatic block to helix growth at the C-terminal end. Evidence from other sources for this directional pathway is presented.

Searches based on hydrogen bonds are demonstrated to be useful for identifying significant structural stabilisations, not only in cases where the hydrogen bonds are themselves part of the stabilisation but also where they form a geometric framework underlying some other stabilising structure. For example:

(1) rings involving the amide side chains asparagine and glutamine hydrogen bonded to distant sections of the main chain are surveyed, the results suggesting that tertiary interactions involving side chains may be important determinants of local secondary structure.

(2) techniques for visualising the hydrogen bonds which stabilise beta sheets also identify ridges of hydrophobic residues running perpendicular to the strand direction, which seem to be a ubiquitous and structurally significant property of all beta-containing protein structures.

This work has given a clearer picture of the types of computer software which are needed if analysis and interpretation of protein structure is to be simplified and extended to include the increasing database of solved crystallographic structures and the large number of related conformations generated from molecular dynamics simulations. Such software should allow both automated searches based on user-defined regular structural patterns and a flexible visualisation system allowing both realistic, detailed pictures of interactions and simplified but information-rich cartoons for pairwise comparison of related structures. Many of the parts of this system exist already, in isolation, but the nature of software development suggests that more work is needed in the area of geometry description and algorithm representation, that is in file and interchange formats, rather than the development of new software which inevitably has a short lifespan and well documented problems with maintenance if that lifespan is extended.

1.1 Studying biomolecular structure

Since the initial successes in the study of biomolecular structure, based on repetitive hydrogen bonding patterns [1, 2], many attempts to explain the three dimensional structures of biological macromolecules in terms of hydrogen bonds have been made. The structure of proteins has proven to be particularly intractable, particularly frustrating in view of the facts that tertiary structure is known to be entirely defined by the increasingly easy to obtain sequence information, and that the majority of biological reactions and signalling systems are based on proteins' shape-dependent properties.

The difficulty in predicting protein structures from sequence information probably has two main causes. The first is that folding occurs in aqueous solution, and the effects of highly polar solvents on molecular dynamics are still poorly understood. The second is the huge range of possible structures, which make it difficult to detect significant patterns in protein behaviour. Even as late as 1981 the development of a new molecular visualisation technique made a significant contribution to the classification of protein structure families [3], while throughout the last 15 years developments in computer display technology have made the analysis of proteins more routine and methodical. I hope to show that there is still some way to go in fully understanding the forces which control protein conformation, but at the same time suggest a number of constraints on the polypeptide backbone which may make the prediction of tertiary structure more tractable.

This work focuses primarily on intra-protein forces with a strong electrostatic component. These are often (but not exclusively) those which are classified as hydrogen bonds. A number of different methods for analysing these interactions are examined, and a Lennard-Jones potential similar in form to those used in many biomolecular force fields [4, 5, 6, 7] is selected as a tool for analysing the conformational space of the polypeptide backbone and for a more sensitive analysis of tertiary interactions.

1.2 Studies of hydrogen bonds in proteins

Identifying patterns in protein structure requires some sort of simplifying scheme. The polypeptide backbone has two free rotations per residue, which in a 100 residue protein gives a 200 dimensional phase space to describe the backbone alone. Direct comparison of

geometric structures involves costly computations, and gives rise to particular complications when the topology of two related structures is different (for example through residue insertion, deletion or mutation). Higher level patterns such as hydrogen bonds are useful here, since they form an intermediate description which can be used either to identify related geometries through a simple matching scheme and are themselves a significant part of the process which determines the structure under investigation.

1.2.1 Observed structural properties

A comprehensive survey of hydrogen bonds was carried out by Baker & Hubbard [26] in 1984. They looked at 15 of the highest resolution structures available at the time and found a number of properties of protein hydrogen bonds which have since been shown to be generally true of all new protein structures.

The elements of this survey which are most relevant to the results in this work include

Globular proteins exhibit saturation of hydrogen bonding potential. In globular proteins, almost all of the C=O and N-H groups of the polypeptide backbone (almost 90%) which could make hydrogen bonds in fact do so. This is a significant constraint on protein folding, especially for those residues whose hydrogen bonding sites are buried in the hydrophobic interior with no prospect of making hydrogen bonds to water molecules, and so have to find other main chain hydrogen bonding partners. The consequences of this constraint are investigated in molecular models developed in chapter 3.

Most hydrogen bonds are mainchain to mainchain. This generally means in regions of secondary structure, helices, sheets or turns. This observation on its own does not help in predictions of protein structure, since the backbone structure and hydrogen bonding pattern is independent of the sequence (except in the case of proline residues, discussed in section 3.7), but provides a framework into which any theory of protein folding would have to be fitted. A particular consequence of this property is that transitions between secondary structure types are likely to be more significant than the stability of any given secondary structure element alone.

Polar sidechains exhort a directing influence on main chain structure. In regions without repetitive structure, these sidechains give rise to extended networks of hy-

drogen bonds with near-optimal geometry (examples of which are studied in chapter 7), while hydrogen bonding to turns and the ends of alpha helices (particularly the N-terminal ends) has a significant role in "mopping up" the potential hydrogen bonding sites, (and perhaps a role as helix initiators or terminators) which is important to the conclusions of chapter 6 on the pathway of helix formation.

1.2.2 Observed thermodynamic properties

Accurate measurement of the significance of individual interactions in proteins is still difficult, but reviews of the available thermodynamic data suggest that we are close to a reasonable interpretation of the factors involved. For example the review by Williams [9] extends work on small molecule association to give a rough quantitative estimate of the factors involved in the free energy of protein folding.

At first examination, hydrogen bond enthalpies seem to have little to offer as tools for the interpretation of protein folding, since their net contribution to association of small peptides or protein folding is a mere 0-2kJmol^{-1} per residue, apparently because of the need to break water-protein hydrogen bonds first. In fact, hydrogen bonds have a significant contribution (-20kJmol^{-1}) when free energy is considered, with the extra contribution provided by the favourable entropy change on freeing water molecules from their ordered state when hydrogen bonded to the peptide.

Other significant contributions to the free energy of folding come from the restriction of internal rotations, with a decrease in entropy and an unfavourable free energy of around +25kJmol^{-1}/residue, and the enthalpy of strain in the folded state of around 7kJmol^{-1}/residue.

The hydrogen bonding contribution is therefore the major factor, with hydrophobic effect (-9kJmol^{-1}/residue) and van der Waals effects (-3.5kJmol^{-1}/residue) providing the fine tuning which discriminates between huge range of different possible folded states.

The very low overall stabilisation of proteins probably plays an important functional role in controlling the rate and reversibility of folding. In this context, hydrogen bond networks define the space of allowed conformations and perhaps of allowed transitions between states, but will rarely be the sole determinants of the final structure.

1.3 The identification of hydrogen bonding patterns

Looking at hydrogen bonds, a number of useful features become clear. Primarily, they are a very quick and reliable way of identifying patterns of secondary structure. For example, alpha helices are identified as regions with multiple hydrogen bonds between main chain peptides four residues apart, a feature which can reliably be used to assign these structures even without knowledge of the backbone torsion angles. Parallel and antiparallel beta sheets can likewise be identified by their hydrogen bonding patterns, as demonstrated in chapter 8.

A more significant use for hydrogen bonding information is in the identification of non-repetitive structures (loops and turns). Although these can also be identified by their torsion angles, torsion angle information is difficult to interpret in tabular form, and is difficult to display in relation to tertiary structure. Since it turns out that loops can be defined in terms of their hydrogen bonding patterns, these can be used to identify novel patterns and find known patterns in new protein structures.

At Glasgow University, research on Protein structure has been carried out in collaboration with the department of Computer Science, to enable new display techniques to be developed in conjunction with the researchers, who in turn can provide the software requirements and feedback on the success or failure of different approaches. Milner-White, Poet and Belhadj-Mostafeda produced new techniques for visualising protein backbone structures [10] and were able to identify new classes of loop [11] using this approach. My own work placed me in both camps, as programmer and end user, which gave some valuable insight into possible improvements in communication and sharing of expertise between the two fields.

Hydrogen bonds define a topology which is an effective first approximation to the full 3D structure, and although this work looks at many wider factors important to protein stability, hydrogen bonds provide the framework for comparison of similar structural motifs throughout.

1.4 Force fields, conformational spaces and constraints

At one level the forces defining protein structure can be said to be fully understood. There are a number of effective force fields developed for energy minimisation of crystallographic structures, and these can now be used for molecular dynamics calculations as well. All are based on a range of parameters which are either experimentally determined or found by *ab initio* quantum mechanical calculations, covering bond lengths and vibrations, three-body (bond angle) and four-body (bond torsion) effects, electrostatics, and other non-bonded interactions. There are a range of different force fields widely used, but all are fairly similar in their properties, barring only a few minor details of their parameterisation.

Typically, these are used in a number of different ways:

Energy minimisation. The force field is used to determine the energy of an estimated structure, and some iterative technique is used to lower the energy of the system without major topological changes being introduced. This is done either as an attempt to improve a crystallographic structure determination, or to relax constraints imposed in the building of a molecular model to provide a more accurate structure.

Simulated Annealing. The force field can be used to generate a dynamic trajectory, approximated by integrating Newton's equations of motion. Energy miminisation on its own is hampered by the fact that protein structure shows a very large number of local minima, and hopefully one single global minimum which corresponds to the final folded structure. To escape from local minima, the dynamic trajectory is calculated for an un-physically high temperature (typically a few thousand degrees Kelvin) for a short time, followed by another cycle of energy minimisation, hopefully into a lower energy conformation. While this approach and variants of it are very effective at finding the enthalpic minimum, they do not provide a certain route to the true structure of a molecule because they do not model the free energy of the system.

Molecular Dynamics. Until recently, the computational effort involved in integrating the equations of motion over the time scales at which proteins exhibit significant conformational change meant that little useful information could be obtained from such simulations. However, increases in computer power and availability have meant

that direct molecular dynamics simulations of whole proteins have become more common, and although protein folding is a process which is still impossible to simulate, properties of protein folding which depend on protein flexibility can now be investigated.

One possible use of these force fields has not been widely adopted to date. Since it is accepted that the final protein structures refined using these parameter sets are realistic, the individual components of the force field are likely to be reliable estimates of intramolecular forces. This is one of the main themes of this work: by taking the individual interactions which together stabilise a structure, it is possible to gain valuable insights into their relative importance and detailed conformation dependent properties.

This was largely carried out using molecular modelling, with reference to structures from the Brookhaven Protein Database [12] to verify whether models corresponded to real patterns seen in proteins. One important feature of these models is that they were not tied to a single force field wherever possible. In particular, constrained energy minimisation was not carried out, even though this improves the quality of simple molecular models. The rationale for this was twofold;

1. the changes in bond lengths and torsions would have made the results sensitive to the properties of the force field used, and hence less general.

2. many of the significant conclusions of this work relate to conformations which are at the edges of permitted regions of conformational space, and in some cases actually depend on the identification of the major *unfavourable* interactions, which would have been very hard to identify if the model structures had become distorted through degrees of freedom other than the ones explicitly being studied.

1.5 Conventions

Throughout this work a variety of different types of display are used to show structures and energy calculations. These are explained as they are used, but the conventions of the three most common types of figure, colour potential energy diagrams for absolute energy calculations, monochrome potential energy diagrams for constrained energy calculations,

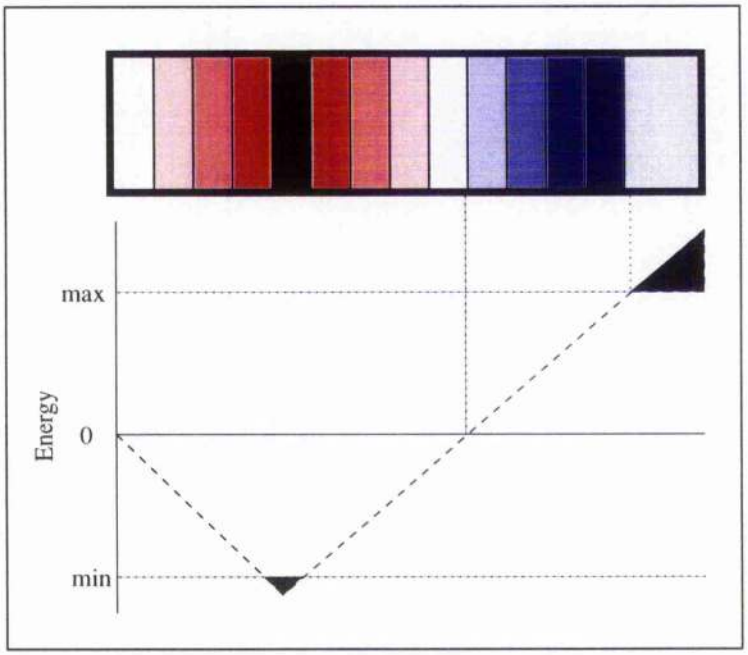

Figure 1.1: Conventions for Energy Contour Diagrams

Throughout this work, many systems are reduced to a two dimensional conformational space and potential energy plotted as a function of two paramters. There are two classes of energy plot, and they are displayed using two different conventions. For plots which represent a full sampling of the conformational space, such as Ramachandran plots, the lowest value found is the absolute minimum for the system. In these cases, energies are contoured as greyscale values relative to this minimum, from white $(H - H_{min} < 0 + \delta H)$ *to black* $(H - H_{min} > H_{max})$. *For plots which represent a spatial interaction between two free groups, the energy is relative to the two groups at infinite separation. Contours are drawn relative to* $H = 0$, *with negative values in the range white→red, and positive in deepening shades of blue. Where* $(H > H_{max})$, *all values are uniform grey.*

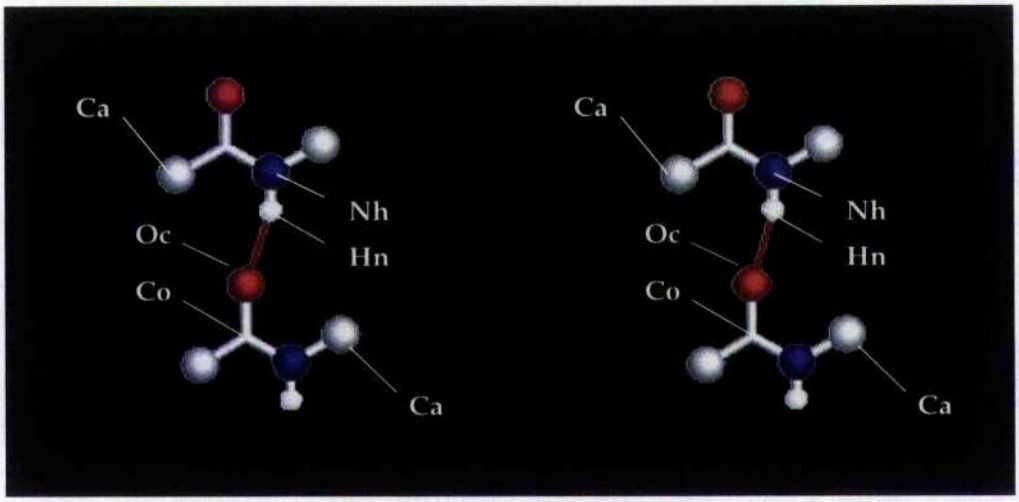

Figure 1.2: Conventions for Molecular Graphics

Molecular diagrams in this work are designed to be seen in relaxed stereo, that is with the left eye seeing the left image and the right eye the right image. The display convention for the ball and stick models is Oxygen–red, Nitrogen–blue, Hydrogen–white, Sulphur–yellow, and Carbon–grey, with amino acid C^{α} atoms drawn as balls and all other carbon drawn at the same width as bonds, giving them the appearance of "joints". Chemical bonds are grey, and hydrogen bonds are red. In most cases, main chain segments are drawn from C_i^{α} to C_{i+N}^{α} rather than by the more usual residue assignments (N_i^{h} to C_{i+N}^{o})

Section I

Section I begins with a review of the various definitions of hydrogen bonds which have been used in the past. From this, it is clear that a flexible but physically realistic technique for examining the geometry dependent properties of intra-protein forces would be useful. Such a technique is developed, and proven to be powerful even in the analysis of systems which were thought to be well understood, particularly the forces stabilising (and in some cases destabilising) strands and helices.

Chapter 2: Hydrogen bonding

Chapter 2 reviews the definition of the hydrogen bond, and contrasts different ways of assessing them in protein structures. Empirical Lennard-Jones 9-6-1 and 12-6-1 potentials are identified as the most reliable way of obtaining a realistic picture of intra-protein forces interactively. A new technique for evaluating the hydrogen bonding potential of polar groups, based on neutral dipolar probes with van der Waals parameters, is shown to capture the main features of hydrogen bond geometry and is used in chapters 3 and 4.

Chapter 3: Geometry constraints and the twist of the β strand

Chapter 3 looks at the four dipoles, two of N-H and two of C=O, which make up the "dipeptide" system of the Ramachandran plot which is used to interpret properties of protein structure. Studying the four possible interaction sets they provide, it is shown that any dipeptide Ramachandran plot is likely to have its minimum energy in a conformation which is seldom seen in real structures. An explanation for this is provided by a model in which tertiary hydrogen bonding is approximated using blocking groups. The resulting potential energy suggests the beta strand has an intrinsic electrostatic predisposition to twist caused by the carbonyl-carbonyl interaction, a feature which has previously been ascribed to side chain interactions. This work has been the subject of a paper, (Maccallum et al I, Journal of Molecular Biology, [13]).

Chapter 4: Hydrogen bond geometries for polar sidechains and peptides.

Chapter 4 takes the potential and probe system developed in chapter 2 and applies it to assess the hydrogen bonding potential and geometry of polar sidechains. This shows

that these simple calculations can reproduce the observed patterns of hydrogen bond distribution around these groups, all properties other than sp^2-directing effects of carbonyl oxygen lone pairs being accurately represented: these results are complementary to recent exhaustive database analyses of sidechain hydrogen bonding properties [35, 27]. The same technique applied to interactions between pairs of peptides shows that electrostatic effects provide the explanation for the observed deviations from ideal hydrogen bonding geometry in crystal structures. The work on peptide-peptide interactions has been the subject of a separate paper, (Maccallum et al II, Journal of Molecular Biology, [14]).

Chapter 5: The stability of different types of helix

Using the results that structure can be interpreted as a set of separate additive interactions (as shown in chapter 3) and that the geometry of peptide/peptide interactions can be explained in terms of steric effects and electrostatics alone (as shown in chapter 4), the stabilities of different classes of secondary structure elements are investigated using model polypeptides and assessing the two residue, three residue, four residue and five residue interactions separately. This reveals three significant features: the 3_{10} helix is not a local minimum on the polypeptide conformational potential energy surface, the π helix is excluded because of atomic collisions, and there is a significant barrier to concerted folding of strands to helices.

Chapter 6: The ends of α helices and the dynamics of helix formation

Chapter 6 takes the observation that simple concerted folding of helices is unfavourable and looks for evidence that a stepwise pathway, adding residues at the ends of helices, is adopted. Common distortions at the N' and C' ends of helices are identified as possible folding features, and a model of stepwise folding is developed which shows that the ends must indeed be distorted. The C' end in particular has a conformational block to helix formation, and it is suggested that the absence of a similar block at the N' end means that helix growth has a preferred C' to N' direction. Evidence from experimental studies is gathered to support this suggestion.

Section II

The second section approaches the problem of protein architecture from the top down, relating the techniques for studying individual interactions to the more elaborate tools needed for tertiary structure analysis. Methods from analysis of text databases to Virtual Reality visualisations are applicable here, and both approaches are shown to provide new insight into the forces stabilising proteins - searches of hydrogen bonding patterns showing how sidechains can stabilise main chain conformations, and visualisations of hydrogen bonding networks helping to identify *hydrophobic* interactions which appear to stabilise beta sheets.

Chapter 7: Tertiary ring structures stabilising proteins

With the assurance that hydrogen bonding is a useful conceptual framework for studying protein structure, a database of hydrogen bonds for proteins from the Brookhaven protein database was developed as an aid to structure pattern identification. Software for automated searches of this data is described in chapter 7, and an application to identifying and classifying ring structures formed by main chain segments hydrogen bonded to amide side chains is shown. The results of this chapter form part of a paper, (le Questel et al, Journal of Molecular Biology [17]).

Chapter 8: Showing hydrogen bonds in relation to protein backbones

One immediate problem which arises from a database of hydrogen bonds is that it is hard to interpret. Chapter 8 demonstrates a number of ways in which hydrogen bonds can be shown in relation to overall protein fold, and suggests a simple way in which these and related displays could be defined for inclusion in general molecular graphics systems. A novel display technique based on the midpoints of peptides is shown to be particularly useful for the display of beta sheet structure. An implementation of this technique has been developed using the VIEW system by collaborators at the University of North Carolina at Chapel Hill (Bergmann et al, Journal of Molecular Graphics [18]).

Chapter 9: Hydrophobic ridges in β sheets

The techniques developed in chapter 8 made it possible to make a quick visual analysis of the sheets in a number of proteins. This review, presented in chapter 9, revealed that the known propensity for hydrophobic residues to pair up between strands actually extends to the formation of ridges of hydrophobic residues running perpendicular to the strand direction. An extensive set of examples is provided, and it is suggested that these ridges are an important factor in the overall stability of β sheets.

Chapter 10: Problems in Protein Visualisation

Most of this work was carried out using specially written software, and it is not clear how a more general package of suitable flexibility could be written to reduce the programming effort for this type of research. However, there are a number of general issues arising from my experience, in particular in the matter of software prototyping and the conflict between the demands of working biochemists and the attitudes of programmers, which suggest some general rules for developing software in the field of biomolecular structure. Some impressive attempts have been made to provide a generic, flexible structure analysis tool [19], but my experience as presented in the final chapter suggests that a satisfactory solution is still some way off, and may lie in better agreement about file formats rather than the development of ever more elaborate specialised software tools.

Section III

Finally, the appendices contain details of parameter sets used, some of the modelling techniques developed for this study, a description of the hydrogen bond format and search software, and two of the many display algorithms which were included in tools developed specially during my research.

Part I

Hydrogen Bonding, Electrostatics, and the Conformational Space of the Polypeptide Backbone

Chapter 2

Hydrogen Bonding

A hydrogen bond is an interaction in which a hydrogen atom bonded to an electronegative atom is exposed to a second electronegative atom and, through a combination of electrostatics and weak quantum mechanical bonding, binds to that atom with a characteristic directionality and short donor–acceptor distance, typically less than the combined van der Waals radius of the two atoms involved.

A full treatment of the hydrogen bond requires ab initio quantum calculations, but these are too time consuming for interactive investigations and in fact show that the major part of the stabilisation is provided by the electrostatic part of the interaction. The electrostatic part has been used on its own as a tool for studying hydrogen bond patterns, but is deficient in regions of close packing or poor hydrogen bond geometry where it is most needed.

An approach based on combined electrostatics and Lennard-Jones potentials is shown to have the right qualitative features to explain hydrogen bond geometry. The close donor–acceptor contact is handled by having a zero radius for hydrogen bonding hydrogen, which compensates for the lack of bonding effects. This simple model shows the correct directionality, has approximate quantitative justification from small molecule data, and is the basis of a range of widely used force fields, so is adopted throughout this work.

The examination of directionality is carried out using a set of fixed atoms and a probe system which is constrained to have a two dimensional interaction potential, which can be visualised easily and captures many of the significant features of the higher dimensional spaces it approximates.

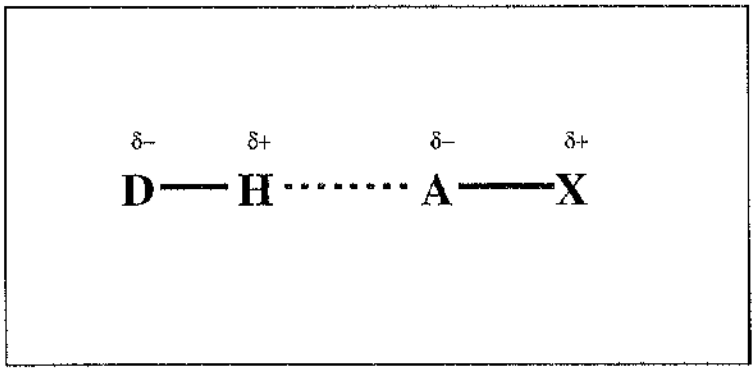

Figure 2.1: A Simple Picture of Hydrogen Bonding
This figure shows the simplest interpretation of hydrogen bonding, as a purely electrostatic effect occurring when hydrogen is bonded directly to a strongly electronegative atom such as oxygen or nitrogen. The resulting charge asymmetry can be treated as a pair of point charges with values which are not multiples of e, so-called PARTIAL CHARGES. *In this situation, hydrogen adopts a positive partial charge, while the atom covalently bound to it (the* HYDROGEN BOND DONOR D*), takes on a negative partial charge. When this system is brought near to another charged or polar group, the atoms tend to line up so that the hydrogen is as close as possible to the atom of that group with the highest effective negative charge (the* HYDROGEN BOND ACCEPTOR A*), while the donor atom maximises its distance from the same group.*

2.1 Introduction

Hydrogen bonds are invoked to explain anomalously short D-H...A distances. In proteins these distances do not seem as extreme as for the classic HF...HF or small molecule case, but hydrogen bond type interactions are ubiquitous determinants of biomolecular structure.

It has long been recognised that the physical properties of molecules containing hydrogen bonded to a strongly electronegative atom are unusual, and that such molecules exhibit unusual association properties and behaviours in polar solvents. This has been attributed to the very low electronegativity of hydrogen, and is regarded primarily as an electrostatic effect of the polarity of the D-H bond. Some of the more extreme effects of hydrogen bonding in materials have led to more elaborate models in which the electrostatic effects are enhanced by a specific bonding term. This has been regarded as necessary to explain the short D-H...A distances seen in some systems, and the unusually strong directional dependence of the bond in materials such as those containing HF species.

2.2 Methods: Models of the hydrogen bond

Models of the hydrogen bond fall into a number of classes, ranging from the trivial geometric description of observed geometric preferences from known protein structures to a full quantum mechanical treatment. The purpose of this chapter is to identify a model which exhibits as close a match as possible to the observed properties of hydrogen bonding in proteins while still being simple enough to analyse and implement in a range of different situations. Models based on empirical atom-centred potentials are regarded as acceptable for the analysis of protein structure, and nearly all structural determination and dynamics simulation is currently done using these methods. Quantum mechanical techniques have historically been difficult to apply correctly and have required unfeasably long computation times, but new approaches, better software and faster computers are now making *ab initio* calculations of systems of the size investigated in this work feasible.

2.2.1 Empirical Models

Since hydrogen bonding can be directly related to the electronegativites of the atoms involved, most models have focused on charge distributions, with some consideration of ideal geometries included in an empirical way.

The first models of hydrogen bonding were purely electrostatic, and attempted to obtain realistic results by distributing point charges to represent valence electrons and unshielded atomic nuclei. for example the simple Pople/Lennard-Jones model which was developed to explain the O-H...O bond distance in water had five different charged objects:

- a single unshielded (+1) charge on the donor oxygen nucleus
- a single unshielded (+1) charge on the donor hydrogen nucleus
- a single point, doubly charged (-2) to represent the bonding electrons
- a single point, doubly charged (-2) to represent the lone pair electrons of the acceptor oxygen
- a double unshielded charge (+2) on the acceptor oxygen nucleus

Using this model Pople and Lennard-Jones were able to assign positions to the electron pairs to explain the observed positions of the nuclei. However, such a model is quite hard

to justify from first principles, and for complex systems the positions of all the electron pairs could not be reliably assigned.

More usable empirical potentials have been proposed, with potentials refined against the observed properties of crystals of small organic molecules and tailored to fit the observed structural properties of proteins - some of these are discussed in section 2.2.4. In particular, systems of point charges are supplemented by Lennard-Jones or Morse style potentials, which allow for the repulsive cores of atoms and also provide a more realistic treatment of the interactions of non-bonded atoms which are very close to each other.

Since a hydrogen bond only contains some 10% actual bonding character, many observers have chosen only to consider the electrostatic component. One case which is of particular significance here is the work of Kabsch and Sander [20], who studied the strength of hydrogen bonds in proteins using a simple model which assigned single point charge to the atoms of the peptide bonds. Some of the properties of this model will be discussed later: there are obvious oversimplifications involved in ignoring the positions of lone pair electrons, but a simple electrostatic model is surprisingly useful. The only extra consideration which will be made here is the effect of non-bonding interactions.

2.2.2 Quantum mechanical models

A complete model would have to provide a full quantum mechanical treatment of the bonding as well as the electrostatic terms. As a step towards this Tsubomura proposed a set of canonical structures to describe the contributing molecular orbitals of the D-H...A system (see figure 2.2). Spectral studies and ab initio calculations suggest that the first two which contain explicit hydrogen bonding terms, ψ_b and ψ_c, contribute roughly 10% of the observed hydrogen bonding energy when the hydrogen bonding geometry is close to linear [23]. Although it is now becoming possible to solve the wave equations for small peptides and sections of proteins, such results are hard to interpret for small sections of large molecules and the match between the theoretical results and the physical properties of the system is hard to assess, particularly when trying to measure attractions between polar groups which are part of the same molecule.

An exhaustive study of the precise nature of the hydrogen bond is outside the scope of this work, but possible bonding effects could play a role either enhancing or interfering with the results based on empirical force fields. The data presented here are qualitative

interpretations of the structural properties of proteins: if it were possible to define the minimal set of significant protein structures which define folding and stability then effort could be focused on these using the best available quantitative techniques. Here it is argued that the problem is not yet well enough defined for such an approach to be successful.

One of the most significant contributions which a covalent model of the hydrogen bond would make would be to enhance the directionality of the interaction: there might be expected to be a stronger tendency for the three atoms most closely involved to be co-linear and in the case of systems involving pi-orbitals, a tendency to be coplanar also. Later chapters show that in proteins this is rarely the case, but it must be stressed that there are considerable conformational constraints placed on any interacting system in a protein. On the other hand, many systems do succeed in maximising electrostatic interactions; a simple point charge based model explains many observed patterns, and quantum bonding effects may only be significant as quantitative corrections to the empirical model presented throughout this work.

2.2.3 Hydrogen bonding sites along the polypeptide backbone

The polypeptide backbone is a sequence of alternating hydrogen bond donors and acceptors, as shown in figure 2.3. Each peptide can act as a donor and acceptor, and in most cases acts as both at once. Saturation of hydrogen bonding potential means that the arrangement of these positions is important in determining the final folded state, and possibly also in the control of the protein folding pathway. Most of these hydrogen bonds are mainchain-mainchain interactions, although bonding to sidechains and solvent also play an important role.

There are, of course, other hydrogen bonding sites in proteins; polar sidechains, the charged ends of the polypeptide backbone, and charged residues can all participate in hydrogen bonding. These will be discussed further in chapter 4.

2.2.4 Electrostatic Models

The success of the Pople/Lennard-Jones model for water, and the fact that the main stabilising interaction in hydrogen bonds is an effect of molecular polarity, have encouraged several workers to adopt a simple electrostatic model of the hydrogen bond. In particular

Kabsch and Sander [20] used a simple four-point model to provide an energy to use as an extra hydrogen bond classification criterion, giving an interaction as shown in figure 2.4. Clearly this model is only a rough approximation, but even an approximate energy assignment turned out to be useful in analyses of hydrogen bonds - for example, see the work of Milner-White in the analysis of the gamma turn interaction found in some main chain conformations [30].

2.2.5 Lennard-Jones potentials

Lennard-Jones potentials provide a more realistic alternative to geometric or electrostatic methods. The electrostatic system obviously has a problem with the absence of a repulsive core on atoms, but also shows clear deviations from the expected behaviour near to the ideal hydrogen bond length, one of the regions we want to investigate in detail.

Many Lennard-Jones type potentials are available, as these are the functional forms used in force fields for structure refinement and molecular dynamics programs, the most commonly used being AMBER [6], CHARM-M [7] and GROMOS [4]. The basis for the non-bonded interactions in these was a series of studies by Lifson, Hagler and co-workers[21, 22, 23] who empirically derived a force field based on the observed structures and thermodynamic properties of crystals of small organic molecules. Later force fields include terms designed to force the energy calculations to match the observed structures of proteins, some containing specific directional hydrogen bonding terms, for example. While this is reasonable for a force field which is to be used for structural determination, there is a risk that properties of proteins which are effects of constraints on proteins' conformational space and hence potential clues to the kinetics of folding may simply be masked out as part of the force field, so I have chosen to use the original transferable force field because it is closer to verifiable caloric measurements and has fewer assumptions about hydrogen bond geometry in proteins.

2.3 Results: Visualising possible interactions

Any potential energy function can be interpreted visually, but the ease with which it can be understood depends on the number of dimensions required to describe it and how intuitive any dimensionality-reducing constraints are. In particular, any approach which

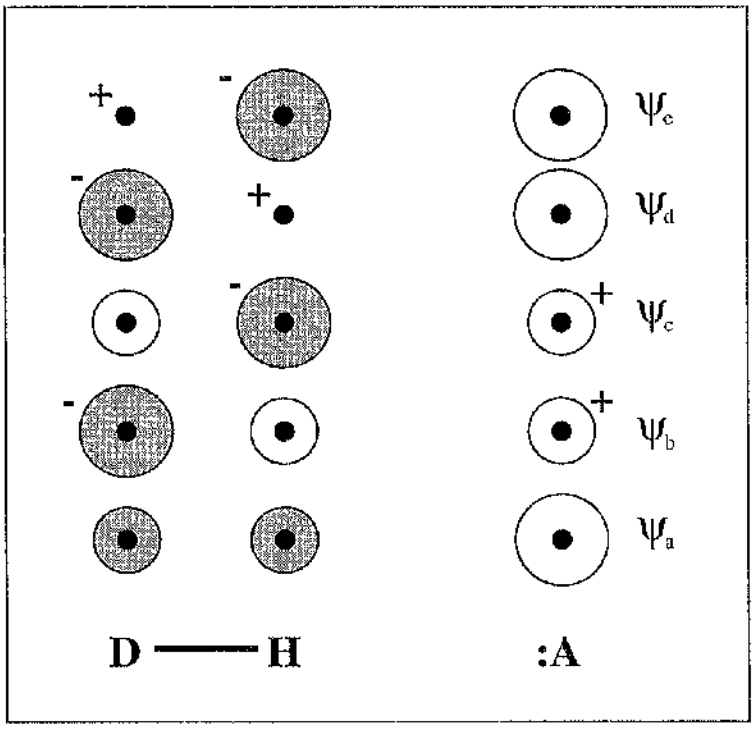

Figure 2.2: Tsubomura Molecular Orbitals

Tsubomura proposed that the short intermolecular distances seen in hydrogen bonded systems were due to weak but genuine bonding effects between the donor and acceptor systems. According to the molecular orbital approximation, any bonded system can be seen as linear superpositions of single-atom wave-functions, and further any bonding system can be seen as a superposition of the resulting many-atom wave-functions, molecular orbitals. This figure shows the possible molecular orbitals for the simplest hydrogen bonding system, ranging from the fully bonding, non hydrogen bonded orbital with lowest energy to the fully anti-bonding systems with higher energy. Hydrogen bonded systems actually exhibit a superposition of the lowest three of these molecular orbitals, but the overall energetic contribution of true bonding is quite small - between 10 and 25% for systems such as proteins.

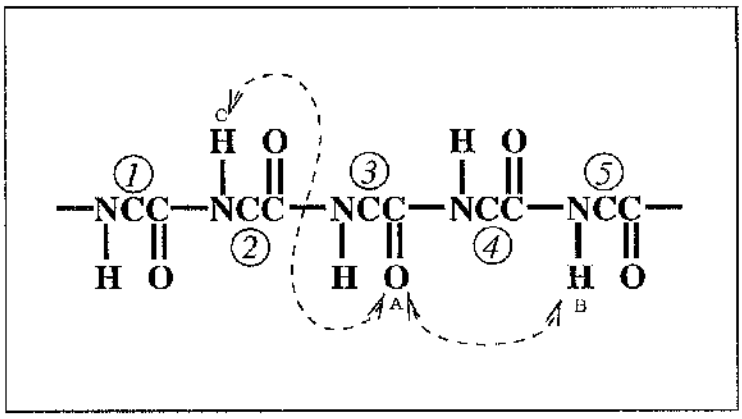

Figure 2.3: Hydrogen Bonding Sites on the Polypeptide

The polypeptide backbone is a series of hydrogen bond donor and acceptor sites. Each peptide has one N-H group which can act as a hydrogen bond donor, and one C=O group which can act as a hydrogen bond acceptor, the oxygen being the acceptor atom. The exceptions are the ends of the strand, where NH_3^+, CO_2^-, or acetylated C termini provide even stronger donor and acceptor sites, and the residue proline, which has no N-H group and hence cannot act as a hydrogen bond donor. The nomenclature is traditionally to number donor or acceptor atoms according to the amino acid residue they belong to, with hydrogen bonds numbered from the acceptor atom, so a hydrogen bond between atoms A and B in the diagram is a 3→5, or i→i+2, hydrogen bond, while one between A and C would be a 3→2, or i→i−1, hydrogen bond.

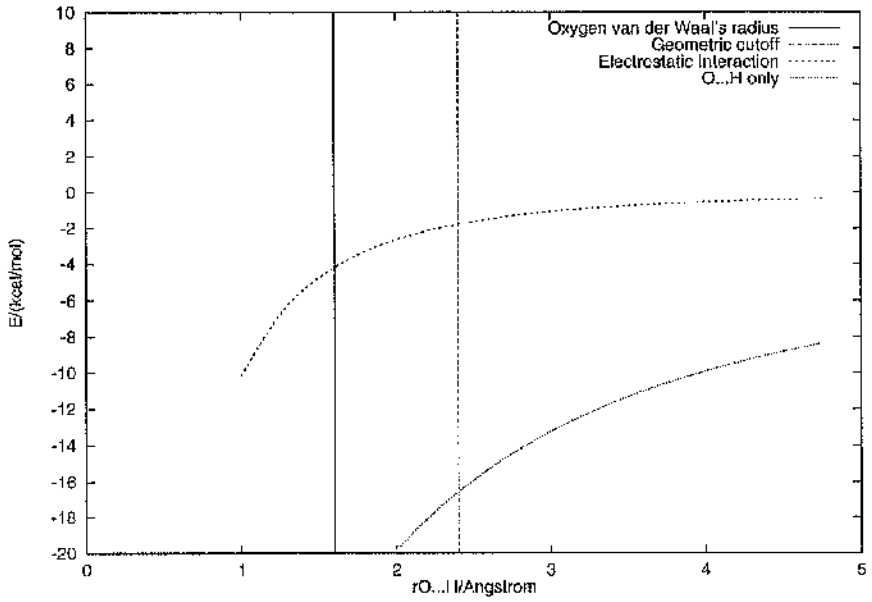

Figure 2.4: Electrostatic Hydrogen Bond Energy

This graph shows the potential energy of the system N-H...O=C as a function of H...O distance assuming the four atoms to be co-linear and taking the partial charges used by Kabsch & Sander as described in the text. At short H...O distances this function is dominated by the $1/r^2$ H...O attraction, while at larger distances the effect of the two groups N-H and O=C being dipolar is to cause the attraction to tend to zero more quickly than for the simple H...O case.

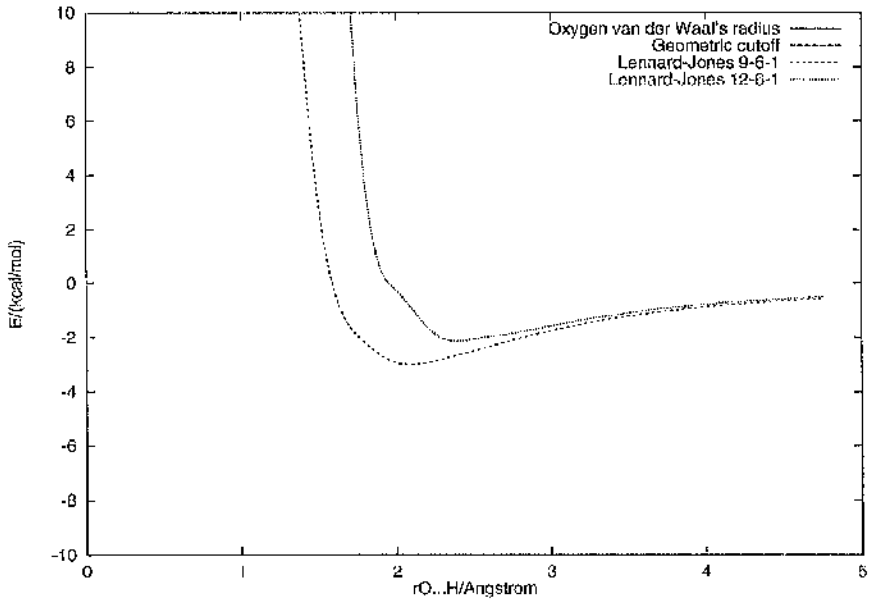

Figure 2.5: Lennard-Jones Potentials

This graph shows the same co-linear N-H...O=C system as the previous figure, using the LENNARD-JONES *12-6-1 and 9-6-1 potentials of Lifson & Hagler as described in the text. The electrostatic part of the potential is very similar to that in the previous example, but the additional van der Waals and London forces which are included can be seen to effectively describe the repulsive short range behaviour too. Notice the way the N...O repulsion can be seen to push into the H...O minimum in the 12-6-1 case. This strong repulsion needs to be countered by explicit hydrogen bonding terms in calculations based on this strength of potential, but is less strong in 9-6-1 potentials.*

results in a one or two dimensional potential energy surface can be useful, especially if the constraints applied are clear enough.

The best way to look at the interactions involving polar groups in proteins is to take one of the groups as reference and display the potential energy of a typical hydrogen bonding partner relative to it. In this way the geometric properties of a given potential energy can be shown. One example is the ϕ, ψ conformational space of a single residue in a polypeptide, first studied by Ramachandran [24], where the significant degrees of freedom can be clearly seen and the rest of the system assumed to be fixed by inflexible bonds. This system is looked at in detail in chapter 3: here a more general case, with no assumed bonding constraints between the hydrogen bonding partners, is considered.

2.3.1 Interaction Plots

Two dimensional interaction plots can be drawn if one of the partners can be held fixed in a plane and each point relative to it restricted to a single conformation, whose potential energy can then be plotted. The test donors and acceptors should fulfil these criteria:

- they should be neutral. If there is a net charge, the subtle effects of the electrostatic distributions will be lost: the potential energy will be overlaid with a simple distance dependent function. If the two units are neutral, any non-zero energies will be a function of the geometry of the interaction alone.

- they should have a simple shape. If the probe is composed of many atoms, some regions of the potential energy surface may be stabilised through contact interactions between parts of the system other than the hydrogen bonding partners under investigation (see for example section 4.3 which explicitly looks at this effect for peptides).

- they should be representative. Wherever possible, realistic representations of polar groups should be chosen, and for example partial charges on any atom likely to have a directing effect on a hydrogen bond should be included wherever possible.

The obvious choice is the simplest possible case, N-H as donor, peptide C=O as acceptor. Neither of these is a stable species, but in the scheme of Lifson et al, each is a net

neutral object. They have simply defined geometries, and correspond to the most common donor and acceptor species in proteins, so make ideal probes.

One problem which arises is the number of degrees of freedom of the system. In this case, the geometry could be completely described by four parameters - H...O distance, N-H...O angle,, H...O=C angle, and (N-H...O)=C out-of-plane angle, but such a four dimensional dataset would be time consuming to calculate and very hard to interpret. Instead, it is best to specify a simpler subset of the system we wish to consider. The first simplification is to consider the cases where (NHOC) are all coplanar, which is likely to give rise to the strongest interactions. Secondly, it is worth making a rough estimate that the hydrogen bonding potential is strongest when the system is linear, and considering only the cases where the N-H...O or H..O=C angles are 180°. This makes it possible to investigate the potential energy in two two-dimensional systems, one where the N-H...O atoms are co-linear and the H...O distance varied, another where the H...O=C atoms are co-linear and the H...O distance varied. In figures 2.8 and 2.9 which show these results, you can consider the group shown to be held in place while the other group has a geometry defined by

- the position of the donor (or acceptor) atom of the free group relative to the acceptor (or donor) atom of the fixed group, which is the x,y position in the potential energy field
- the second atom of the free group, which is placed along the line joining the donor and acceptor atoms, pointing away from the fixed group.

The bond length is held constant. The system is shown in figure 2.6.

2.3.2 2-D representation of geometric hydrogen bond criteria

The simplest case to apply this type of display to is the set of geometrical constraints based on observation of proteins [26]. These can be treated as providing an all-or-nothing potential energy function. The configurations classified as hydrogen bonds can be seen in figures 2.7 and 2.8.

The interesting feature of these is that the observed values of the N-H...O angle cover a much smaller range than the C=O...H angles. This feature is a test of any model of the

potential holding together hydrogen bonding groups; the ability to reproduce this effect is important for any technique which is to be used for analysing more complex structural interactions, as later chapters do.

One further effect which is seen is that there is a tendency for hydrogen bonds to cluster around two positions which correspond roughly to the expected positions of centres of charge associated with the two lone pairs on the carbonyl oxygen (see, for example, Ippolito *at al* [27]). This is not a strong effect, and it may be simply a statistical result - since the potential is wide enough to accommodate two hydrogen bonding partners, and those partners will be mutually repulsive, they will tend to bind at opposite sides of the potential minimum. Judging whether treating the charge as being concentrated at two points for a more directional model as has been attempted in some force fields is useful or not would require a sensitive quantum mechanical treatment of the single bond case.

2.3.3 Electrostatic 2-D interaction plots

As a next step, it is instructive to investigate the properties of the Kabsch and Sander calculations as applied to the sort of system which will be studied in depth using the more complete Lennard-Jones potential. The Kabsch and Sander calculation is widely referred to, as it is available in the DSSP format files for describing polypeptide backbone structure [20]. So what does it actually do, and how does this relate to the more complete calculation? Figure 2.8 shows the two-dimensional form of this potential.

This method calculates the electrostatic component of a hydrogen bond assuming that the N-H and C=O groups are neutral overall with an uneven charge distribution as described in appendix A. To investigate the effectiveness of this model, it is possible to examine the energy it calculates for a wide range of geometrical arrangements. Since the two groups are neutral overall, it is possible to treat them as if they were separate molecules interacting in vacuo, a reasonable assumption if we agree only to draw conclusions from results at short ranges where there is no question of solvent molecules or other polar species intervening to interfere with the interaction.

The results show quite neatly the benefits and disadvantages of the model. In the absence of a repulsive core, short range interactions are possible where overlap effects would have been expected to prevent the atoms approaching each other. This is not a problem in the examination of crystallographic coordinates, where these repulsions have already

been accounted for in the refinement procedure, but they mean that an examination of disallowed conformations in model structures using this technique is inappropriate. Another feature of the results is that there is no strong directionality predicted if the electrostatic energy alone controls the hydrogen bonding geometry. The most important factor determining the potential is the donor-acceptor distance, but in a real protein the observed hydrogen bonds show quite distinct directional preferences which must be related to the potential energy.

2.3.4 Lennard-Jones 2-D interaction plots

Finally, figures 2.11 and 2.12 show the interaction plots for the Lennard-Jones 9-6-1 potential. Two important properties of these plots are the realistic repulsive core, and the very close match of the shapes of the potentials with the geometric rules which are based on observation of proteins. The N-H...O potential is narrower than the C=O...H potential, as represented in the lower allowed range of N-H...O angles relative to the range of C=O...H angles. This is even the case in the absence of distinct sites for lone-pair electrons, which are normally invoked to explain the wide range of C=O...H angles observed.

In addition to providing a repulsive core region as expected, notice that the shapes match the empirical geometric constraints much more closely than the simple electrostatic model. This is due to the form of the Lennard-Jones parameters - in particular the van der Waals radius of hydrogen is effectively zero, since in hydrogen bonded systems the acceptor atom can come very close to the hydrogen. Sufficient repulsion is provided by the hydrogen bond donor atom, Nitrogen in this case, that the '9' and '6' parameters on H can be set to zero. For the fixed N-H case, this means that O close to H but as far as possible from N is favoured, and hence a restricted range of N-H...O angles is seen. For the fixed C=O case, the repulsive core is more symmetrical and a wider range of C=O...H angles is allowed. There is no repulsion in this case between the fixed group and the "probe" object, the hydrogen of N-H, and since the constraints always have the hydrogen pointing at the oxygen, the potential is dominated by the electrostatics except where the O...H distances are short enough for the N...O repulsion to take effect.

Clearly, the 9-6-1 potential presented here is an adequate model of the hydrogen bond in proteins. The shape of the energy minima for both the fixed N-H and fixed C=O cases is a close match for the observed distributions, and the potential contains terms to represent

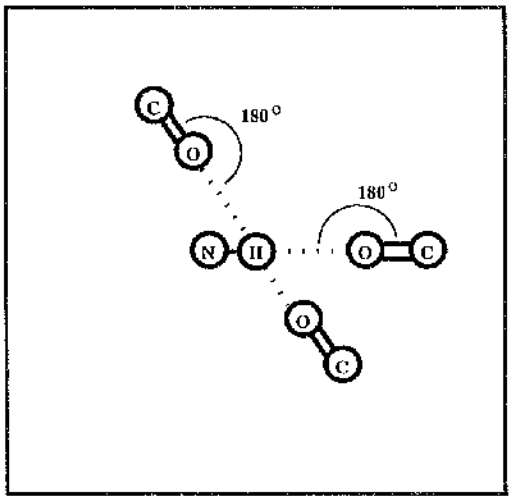

Figure 2.6: An Idealised Geometry for Hydrogen Bonding
The N-H...O=C system has 5 degrees of freedom, so the potential energy surface cannot be visualised easily. However, by fixing the system to be coplanar the degrees of freedom are reduced to 3, and adding the further constraint that the angle H...O=C be fixed to 180° reduces the system to 2 dimensions, so a 2D potential energy surface can be drawn. Each point on the surface represents the position of an oxygen atom relative to a fixed N-H group, with the C=O always pointing directly at the hydrogen in the system.

both the dipolar interactions and the van der Waals contact effects. The calculation involved for each conformation is quite simple, and many different conformations (or many different interactions in a single large system) can be calculated in a reasonable time.

2.4 Conclusion: Lennard-Jones potentials are adequate for examining hydrogen bonds

This chapter has shown how there are a range of different models of hydrogen bonds, each having its own problems with either realism or calculation time. Examining the properties of these different systems involves finding a set of representative conformations which show the behaviour of the potentials in different situations. Enforcing planarity and providing angle constraints seems to be a good way of getting a flavour of the various systems, and the results show that a Lennard-Jones potential of the type refined by Lifson & Hagler [23] shows the correct geometric properties, and is believed to provide energies for hydrogen

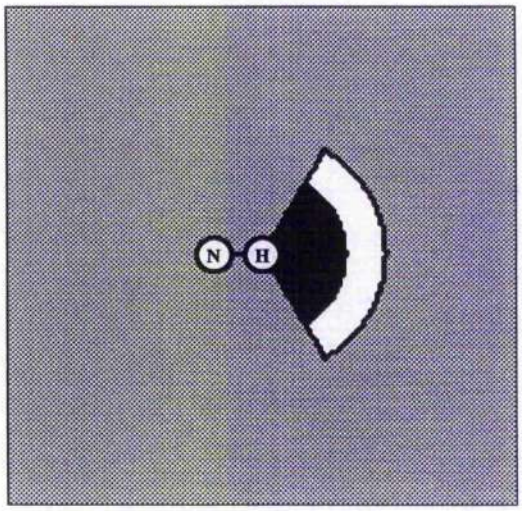

Figure 2.7: Classical H-bonding Regions – N-H fixed
This figure shows a 2D interaction plot as shown in the previous figure, with fixed N-H and C=O...H constrained to be linear The clear regions show the conformations which are classified as hydrogen bonds by the geometric criteria of Baker & Hubbard, the black region shows where the constraints would give N-H...O distances less than the van der Waals radius of Oxygen + the van der Waals radius of Hydrogen.

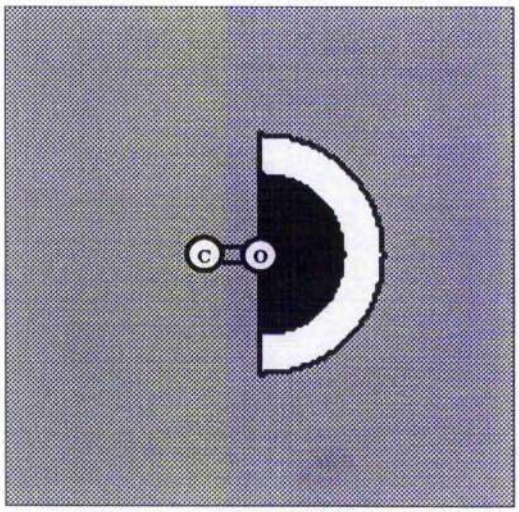

Figure 2.8: Classical H-bonding Regions – C=O fixed
This figure shows the same 2D interaction plot as in figure 2.7, but with C=O fixed and O...H-N linear, again using the hydrogen bond criteria of Baker & Hubbard.

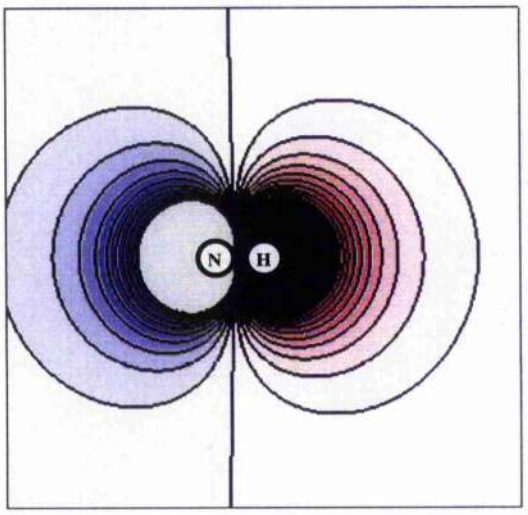

Figure 2.9: Electrostatic Interaction Potentials – N-H fixed
2D interaction plots with fixed N-H and C=O...H linear. The potential energy has been calculated using the electrostatic potential of Kabsch & Sander. Energies are contoured at 0.5 kcalmol^{-1} intervals, with a minimum of -5.0 kcalmol^{-1} and a maximum of 5.0 kcalmol^{-1}. The steep attractive core is apparent, but it is also clear that the potential shows very limited directionality.

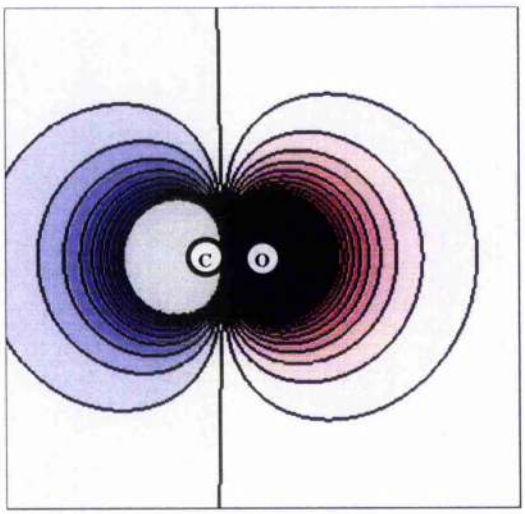

Figure 2.10: Electrostatic Interaction Potentials – C=O fixed
2D interaction plots, with C=O fixed and O...H-N linear, potential energy using the electrostatic potential of Kabsch & Sander. Energies are contoured at 0.5 kcalmol^{-1} intervals, with a minimum of -5.0 kcalmol^{-1} and a maximum of 5.0 kcalmol^{-1}. This and the previous figure suggest that N-H...O and C=O...H angles could show a very wide variation if electorstatics were the only significant part of the potential energy function.

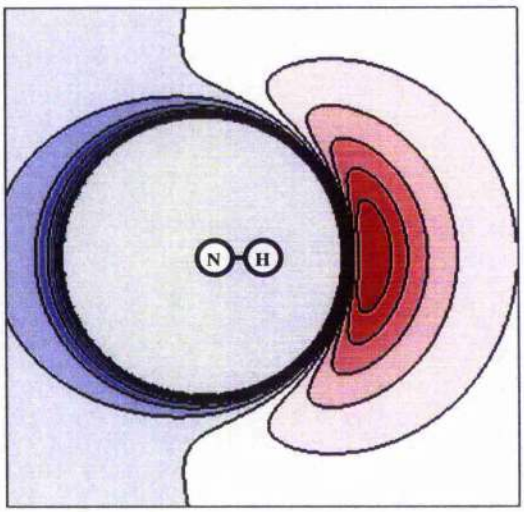

Figure 2.11: Lennard-Jones Interaction Potentials – N-H fixed
This figure shows a 2D interaction plot with fixed N-H and C=O...H constrained to be linear, the potential energy calculated using the Lifson & Hagler Lennard-Jones 9-6-1 potential. Energies are contoured at 0.5 kcalmol^{-1} intervals, with a maximum of 3.0 kcalmol^{-1}. The minimum covers a small range of N-H...O angle values, although there is no specific directing term in the potential energy.

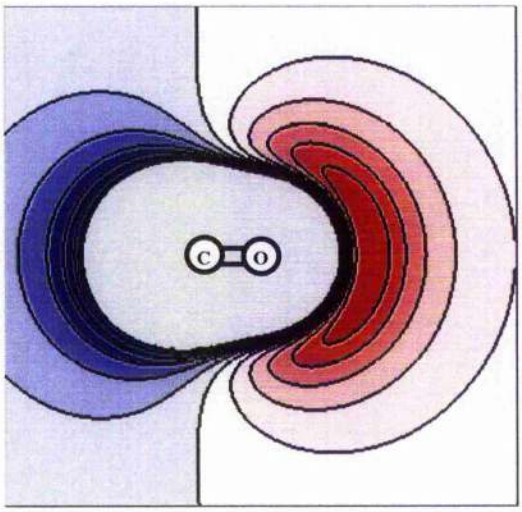

Figure 2.12: Lennard-Jones Interaction Potentials – C=O fixed
This figure shows the 2D interaction plots, with C=O fixed and O...H-N linear. The potential energy has again been calculated using the Lifson & Hagler Lennard-Jones 9-6-1 potential. Energies are contoured at 0.5 kcalmol^{-1} intervals, with a maximum of 3.0 kcalmol^{-1}. The minimum covers a wider range of C=O...H angles than the fixed N-H case.

bonds which are acceptably close to the actual values.

The directional properties of the hydrogen bond have been shown to be adequately reproduced by this simple functional form based on pairwise additive potentials. This is important, because it means that we are now in a position to look for more complicated shape based potentials safe in the knowledge that there is no intrinsic reason why special features such as lone pair electrons need to be taken into account. This is not to say that pairwise potentials are the only right way to study the forces in proteins, nor that the results cannot be superseded by more elaborate techniques, but given a problem where the complexity of the phase space is more than enough of a challenge, it is useful to have a valid, simple function to evaluate at each conformation.

Chapter 3

Geometry constraints and the twist of the β strand

Interaction potentials based on two-atom probes and fixed groups suggest a method of interpreting the conformational energy of two adjacent peptides as four separate interactions.

Analysis of the four interactions shows that all solvent free calculations will necessarily give the gamma turn conformation as the potential minimum for dipeptide, but also that there is an asymmetry between the right and left twisted β strand conformation as a consequence of the high dipole of the carbonyl group.

When tertiary hydrogen bonds are taken into account by a simple model which treats hydrogen bonding groups as fixed external constraints, the gamma turn interaction is destabilised and the two peptide system shows that there is an intrinsic tendency of the beta strand to twist even before side chain effects are included, partly a function of intrinsic electrostatics, partly a function of the external environment of strands in solution or protein cores.

3.1 Introduction

The first question which arises from the adoption of a new way of analysing hydrogen bonds clearly has to be "does it give any different results from existing methods?". Here the system is applied to the simplest possible geometry of the polypeptide backbone, and novel results are demonstrated which help to explain the twist observed on beta sheets.

3.1.1 Non-optimal hydrogen bond geometries.

Hydrogen bonds based on electrostatic models will always give the potential energy minimum where the dipoles are aligned end-to-end. However, other geometries are possible, and some turn out to be significant where other factors such as backbone bonding constrain the backbone. Identifying which of these is significant requires at least an approximate value for the energy in each conformation, and methods for examining systems which may be rare or unobserved in proteins. To start with, the simple N–H...O=C system is likely to be useful, although here we impose a quite different set of constraints on the system to show the potential energies of systems which are far from ideal hydrogen bond geometries.

3.1.2 Antiparallel dipoles

Can any system which cannot be explained by electrostatics or hydrogen bonding geometries alone be better understood using this slightly more complex approach? The answer is yes, there turn out to be a number of aspects of protein conformation which can be investigated in this way.

The first case to consider is a purely electrostatic interaction which cannot be described as a hydrogen bond but which is widely observed in chemical systems. The donor and acceptor probes we are considering constitute pairs of charges distributed along an axis, and this is of course the description of a dipole. Dipoles can interact in the classical 'hydrogen bonding' sense, lined up end to end with opposite charges near to one another, but they have another significant stable conformation: they can line up side by side, with opposite charges adjacent, the *antiparallel* configuration.

The figure 3.1 shows the potential energy for our simple N–H...O=C system, but in this case the H...C=O angle, instead of being 180°, is fixed at 90°. Thus the values where the oxygen is at the bottom of the plot represent the antiparallel conformation, those where it is in line with the N–H axis represent an extreme of the allowed hydrogen bonding geometries, and those at the top of the plot represent parallel dipoles, where like charges are close to each other and hence there is net repulsion.

There is clearly a favourable interaction by this model, but the geometry is such that a typical geometric hydrogen bond description would suggest no stabilising interaction. Including the electrostatic and steric forces shows that the conformation is not only

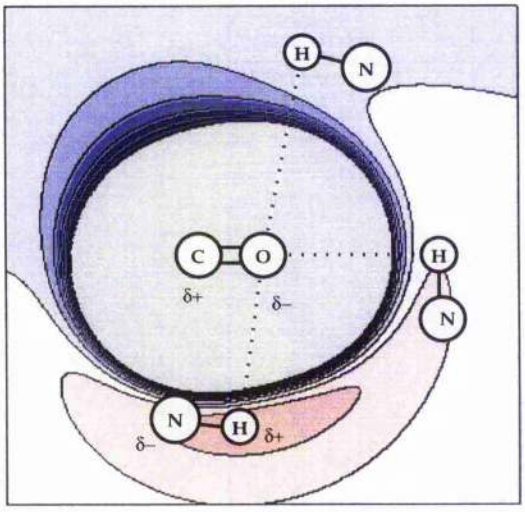

Figure 3.1: Antiparallel Dipoles Represent a Stable Conformation
This is a 2D interaction plot for the same system as figure 2.12, but with the angle O...H-N fixed at 90° rather than 180°. This changes the nature of the interaction completely: where angle C=O...H is linear, the interaction is weak, but where it is near 90°, the C=O and N-H dipoles are antiparallel and a significant minimum is found. Where C=O...H is near -90° the dipoles are parallel and the interaction is destabilising.

favourable, but in principle accessible, so is worth looking for in proteins to see if it plays a role. The electrostatic model of Kabsch and Sander would have found the same favourable interaction, but would not have been able to suggest whether it was a short-range anomaly of the simple model or a possible configuration.

3.2 Observations: Backbone constraints on neighbouring peptides

Antiparallel dipoles may be allowed in proteins, but they are weaker than hydrogen bonds, and as a result will not be found stabilising systems which are free to make hydrogen bonds. They are only likely to be significant interactions in cases where there are strong constraints, for example those provided by bond angles in interacting systems which are close to each other along the polypeptide chain.

Since most possible hydrogen bonds are satisfied in proteins, any weaker interaction is either going to be a less significant or will only occur under special circumstances where

other features prevent hydrogen bonding. The most common constraint on potential donors or acceptors is that the majority of them are part of the polypeptide backbone. This means that they are not free to form the best possible hydrogen bonding geometries, and also are held close to their neighbours by bonds, and so have to act in concert with the peptides nearby along the chain.

Such a system is the set of interactions between the atoms of adjacent peptides, those which share a single alpha-carbon atom. Each peptide can be treated as a set of two dipole units, one N–H and one C=O, and the resultant set of four possible interaction pairs can be examined. All of these interactions could show stable end to end or antiparallel dipole effects, and all of them are heavily constrained by the rigid bonds of the polypeptide backbone, held in close proximity to each other with no real opportunities to make hydrogen bonds between themselves, as will be shown later. The local electrostatics do make a significant contribution, in ways which can be related to classical hydrogen bonds but are subtly different.

The system described below is the pair of peptide bonds on either side of the alpha-carbon of a single residue either in a polypeptide or in the N-acetyl N'-methyl alanine system which is often used as a test of potentials and contrasted with the Ramachandran plot. Thus all the main chain atoms between the alpha carbon of residue $i-1$ and $i+1$ are included, where residue i is the central amino acid residue.

3.2.1 The Gamma Turn conformation

The most significant interaction in the adjacent dipeptide system is the one between the C=O of residue $i-1$ and the N–H of residue $i+1$, that is an $i \rightarrow i+2$ hydrogen bond. This is the only one of the four interactions which has enough free rotation and a wide enough separation to form a genuine hydrogen bond, and in this case the possible hydrogen bond is actually seen in proteins. It is known as the *gamma turn* [29, 30], and is seen mostly at the ends of beta strands in proteins with extensive beta sheets, although it is not common. Its rarity is mysterious when the ease with which it can be formed is considered - only a single amino acid residue has its conformation fixed by this interaction, compared with the four required to make a single alpha helical hydrogen bond, so it should be entropically favoured.

3.2.2 Carbonyl to carbonyl interactions

Looking at dipolar interactions does not restrict us to looking at possible hydrogen bonding systems. Any dipole can adopt the antiparallel configuration to minimise its potential energy, and the most significant dipole in a peptide is that of the amide carbonyl group. The two carbonyl groups in adjacent peptides, those of residues $i-1$ and i, can try to line up in this way, or even to try to form an end-to-end alignment. There are only three bonds between the two carbon atoms, though, and this would be expected to provide a strong constraint on how close to the optimal configuration the system can get.

3.2.3 Other symmetry related interactions

Since the peptide is being treated as two polar units, there are four possible interactions between a pair of peptides. The other two interactions can also be treated separately.

The first is the C_5 interaction which has been studied in some detail by Benedetti and Toniolo [28]. This is the interaction which is predicted to occur between the N-H and C=O groups of a single amino acid residue, the nearest pair on adjacent peptides. The groups cannot form a hydrogen bond, but may be expected to form a dipolar interaction providing strongest stabilisation when the H-N-C^{α}-C=O five-membered ring structure (hence C_5) is planar.

Finally, there is one other possible interaction, closely related to the carbonyl-carbonyl interaction. The two N-H groups each have an effective dipole, and thus they can interact in exactly the same way as the carbonyl groups might be expected to. In fact, in a sense they form the exact complement of the carbonyl-carbonyl interaction: swapping C=O for N-H in glycine residues is equivalent to swapping ϕ for $-\psi$. If the two interactions had the same potential energy profiles, then these two groups would provide potential energies which would be symmetric about the line $\phi = -\psi$ on the Ramachandran plot.

3.3 Methods: Measuring the relative stability of stabilising interactions.

A survey of the observed phi/psi distribution in known protein structures shows that there are significant regions of the backbone parameter space which are rarely adopted

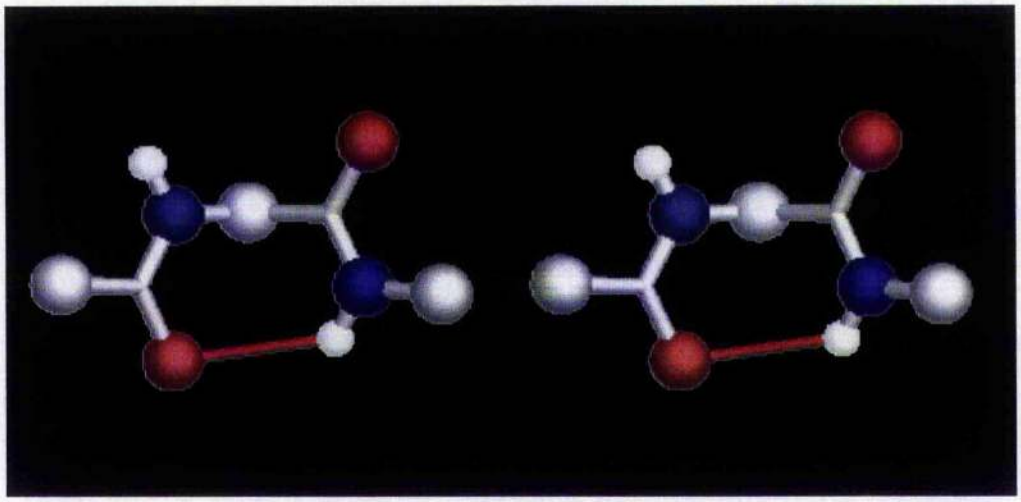

Figure 3.2: The Gamma Turn Interaction
As described in the text, a peptide can be approximated by a pair of dipoles. For the "glycyl dipeptide" this means that there are four interactions to consider: the most significant is shown here, the $C{=}O_{i-1} \rightarrow N\text{-}H_{i+1}$ interaction. This can have hydrogen bonding character, and conformations stabilised by this $i \rightarrow i{+}2$ hydrogen bond are known as "gamma turns" when seen in proteins. For the central resiude $\phi = -90°, \psi = 90°$.

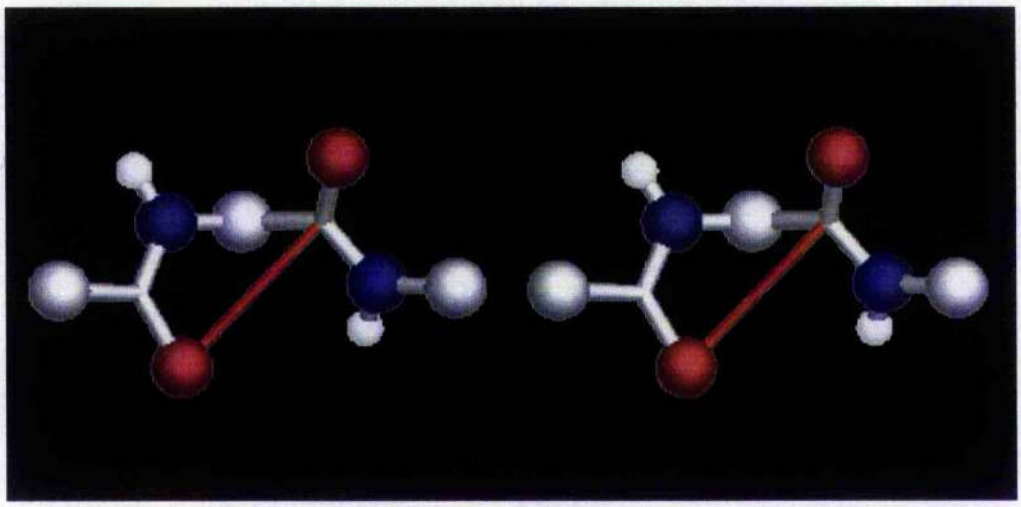

Figure 3.3: The Carbonyl-Carbonyl Interaction
The carbonyl group has the largest dipole of the two units, so the $C{=}O_{i-1} \rightarrow C{=}O_i$ interaction would be expected to be significant. There are only three bonds seperating the two groups, so the freedom of these dipoles to line up favourably is severely constrained. The conformation shown, $\phi = -90°, \psi = 120°$ is allowed.

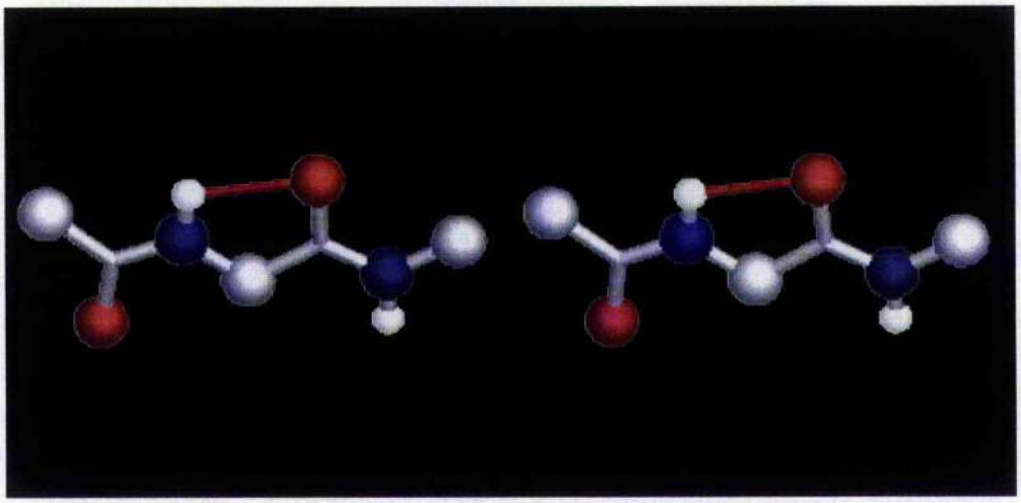

Figure 3.4: The C_5 Interaction
The most strongly constrained interaction in the dipeptide is the N-H_i →C=O_i interaction. In fact, the only free parameter here is the H...O distance, so the potential energy is effectively monopolar rather than dipolar. The H...O distance is shortest in the conformation shown, with $\phi = \psi = 180°$.

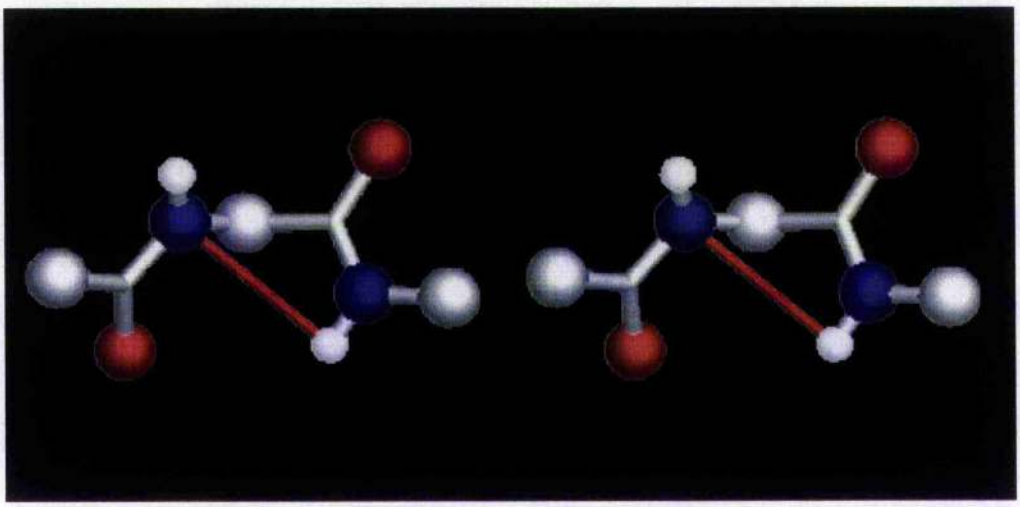

Figure 3.5: The N-H...N-H Interaction
If the charge and bulk of the N-H groups were the same as those of the carbonyl groups, the N-H_i →N-H_{i+1} would just be a symmetry-related form of C=O_{i-1} →C=O_i. Shown here is $\phi = -120°, \psi = 90°$, the equivalent of figure 3.3. In fact, the hydrogen atom is smaller and has a lower charge magnitude than oxygen, so there is an asymmetry, discussed in the results section.

by any residues. This means that as an analytic tool, structural data is at best incomplete. In addition, this plot shows the cumulative effects of many different contributions to polypeptide stabilisation, including local interactions, secondary structure stabilising interactions, tertiary mainchain interactions, sidechain interactions, solvent and substrate effects, and hydrophobic effects. In many cases the number of external constraints to the backbone conformation are such that unfavourable regions of the phi/psi space are adopted by residues, which show bond lengthening and bond angle distortions to accommodate these unfavourable conformations. These effects combine to mean that the phi/psi plot, and associated energies, taken from coordinates in real proteins are not directly useful in the analysis of the interactions of neighbouring residues because coverage of the available phase space is patchy, incomplete and inconsistent. What we are actually interested in here is the full phi/psi plot without relaxation effects, and the appropriate technique to use is therefore molecular modelling.

It is simple enough to generate coordinate sets for a dipeptide with only two free structural parameters. The coordinates of the first peptide are fixed, and the plane defined by these atoms and its normal define the basis space in which the coordinates for the second peptide can be generated: the full algorithm is presented in appendix A. Since we are interested here in the local peptide-peptide interactions, it is possible to leave out the positions of hydrogen atoms and sidechains. C^{α} atoms are not included in the energy calculations but are required as part of the coordinate generation.

Once the coordinates have been generated, it is a simple matter to calculate the energy, since the potential form used is only dependent on the distances between pairs of interacting atoms. The coordinates can be generated for phi/psi values covering the whole range from -180° to 180°. The results are very high energies - typically several thousand kcalmol^{-1} - because Lennard-Jones potentials are not strictly valid for species which are actually bonded to each other. Fortunately, it turns out that overall shape of the potentials is still meaningful, and the energies can be interpreted if they are taken relative to the lowest energy found for the system - so the lowest energy found can simply be subtracted from all the energies found, which is reasonable so long as a conformation near to the lowest energy conformation has been sampled.

3.4 Results: Four interactions which affect polypeptide conformation

The results presented in this section show that this technique works very well, providing a detailed breakdown of the forces between pairs of neighbouring peptides, and also identifying a number of surprising features, one with implications for tertiary structure effects which are dealt with in the following section and another which provides new insight into as well known but poorly explained feature of protein structure, the observed ubiquitous right-handed twist on beta strands and sheets.

The results here have been calculated without inclusion of side chain atoms, so in each case there is rotation symmetry about $\phi = 0^o, \psi = 0^o$, and each minimum is present twice. This means that they approximate the potentials for a glycine based polypeptide. A preliminary generalisation to other amino acid residues would be simply to look at the minima which fall in the top left quadrant of the phi/psi plots, since that region is least disrupted by side chain effects. When I refer to the minimum of these plots, the one accessible for L-amino acids with will be the one of interest. Side chains are not included to allow the intrinsic dynamics of the polypeptide backbone to be studied, to help distinguish side chain effects from the core nature of the polypeptide.

3.4.1 The Gamma Turn conformation

Figure 3.6 shows the potential energy calculated for the gamma turn interaction, which is the strongest of the four interactions, as we will see. The interaction is symmetric, and quite easy to interpret.

The minimum for this interaction might be expected to lie where the two peptides are coplanar with the hydrogen bonding H and O as close to each other as possible. This would form a seven-membered ring structure, which should be relatively stable (see, for example, chapter 7, which discusses hydrogen bonded rings in more detail). The calculations show that this structure is impossible; because all of the bond angles in the system are close to 120^o, only a six-membered ring could be accommodated in the available space. The $\phi = 0^o, \psi = 0^o$ structure is destabilised by the repulsion between the carbonyl oxygen of residue $i-1$ and the amide nitrogen of residue $i+1$. (There is no steric repulsion between the carbonyl oxygen and the amide hydrogen since the effective van der Waals radius of

hydrogen bonding hydrogens is zero.)

The region where the dipole interactions can be strongest while still accommodating this repulsion is around $\phi = -90°, \psi = 90°$ (and $\phi = 90°, \psi = -90°$ for glycine residues). This is reassuring, since it corresponds exactly to the beta-pleated sheet structure of Pauling and Corey [2] which was predicted as the most stable beta strand structure - their crucial insight being that the extended conformation, $\phi = 180°, \psi = 180°$, would not provide an ideal hydrogen bonding geometry. In fact, pleating is observed in all beta sheets, although not to the extent predicted either by Pauling and Corey or by this potential energy calculation.

Other studies of the interactions between two neighbouring peptides focus on full implementations of the various force fields used for structural studies of biomolecules, and usually consist of a calculation of the potential energy of the Ramachandran plot for N-acetyl-N'-methyl glycylamide (or alanylamide). These calculations have not previously been broken down into their component parts, but it is still possible to identify in each case that the predicted minimum is near $\phi = -90°, \psi = 90°$; in other words, whatever the force field the conformation of the alanyl dipeptide is dominated by the gamma turn interaction. This is curious, because the gamma turn conformation is actually quite rare in proteins (see for example the review of Milner-White [30]). There must be some other effect in globular proteins which masks this interaction. A likely explanation is provided in section 3.5.

3.4.2 Carbonyl to carbonyl interactions

Figure 3.7 shows the plot for the C=O...C=O interaction. The minimum lies well above and to the right of the diagonal line $\phi = -\psi$, such that it favours a β-strand with a right-handed twist. In this conformation the two carbonyl groups are not far from being antiparallel, and the close match of this potential energy contribution for the observed conformations of beta sheet residues is very suggestive. The minimum is not as pronounced as that for the gamma turn interaction, but if as seems likely the gamma turn is forbidden by tertiary structure effect, then the carbonyl-carbonyl interaction may well be the most significant stabilising effect between adjacent peptides, and as a result the twist on the beta strand may well an intrinsic property rather than a statistical effect or one guided only by side chain crowding as has previously been proposed [31, 32]. It is worth remembering

that previous descriptions of the forces stabilising the polypeptide have almost exclusively focused on interactions which could be classed as hydrogen bonds. Here is a case where there is a very significant interaction, quite clear to see, which is purely electrostatic in nature.

3.4.3 Other symmetry related interactions

Figure 3.8 shows the C_5 interaction. It is hard to justify a study of this interaction using a non-bonded potential, since the species involved are clearly interacting through bonds, but it is still possible to get a quick interpretation of the relative importance of this interaction as compared to the other three. It has been shown to be significant in a number of sterically restricted polypeptides using doubly C^{α} substituted amino acids [28], but here it can be seen that in the absence of such constraints, the interaction displays only a shallow minimum at $\phi = 180°, \psi = 180°$, where the C_5 atoms are in a flat five-membered ring structure. The bonding constraints mean that the interaction cannot get close to an ideal hydrogen bond, so this interaction has very little directing effect on polypeptide conformation.

The twisting effect of the carbonyl/carbonyl interaction would be exactly cancelled by the N–H...N–H interaction if the 9-6-1 parameters for the two types of group were identical. However, there are two significant differences. The first is that the amide hydrogen has zero effective van der Waals radius, and so there is no strong repulsive core region for the interaction as there is for the carbonyl interaction. This means that the potential energy profile is much flatter, and no sharp minimum is provided to distort the overall energy profile of the phi/psi plot. The second is that the partial charges on the N–H group are much lower, and as a result the interaction is much weaker. This has a very significant effect on breaking of the intuitive symmetry about $\phi = -\psi$. The shape of this potential is shown in figure 3.9.

3.5 Methods ii. Tertiary hydrogen bonding effects on the dipeptide potential

Since the gamma turn is such a strong interaction, dominating the Ramachandran plot for the dipeptide in every set of calculations of its potential energy (including those using

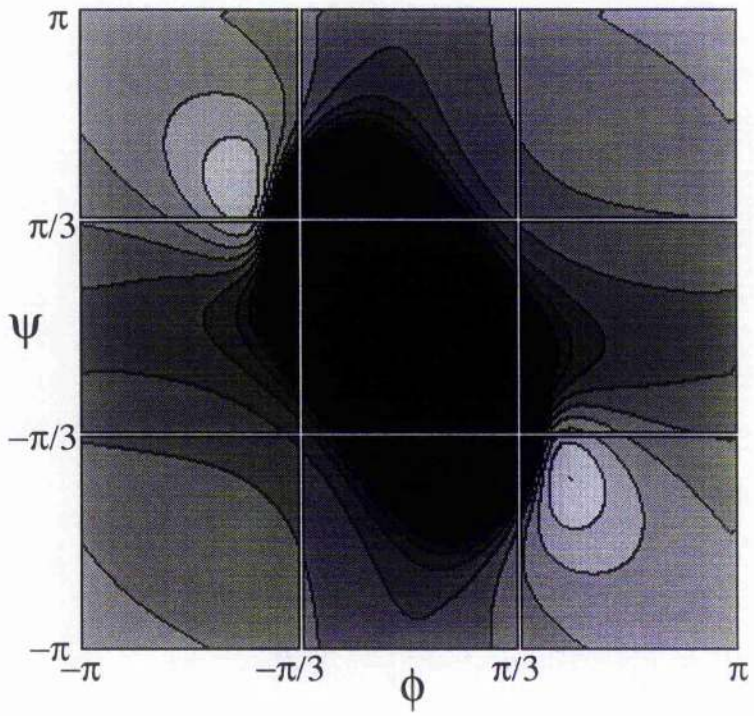

Figure 3.6: Energy of the Gamma Turn Conformation

This shows the value of $H - H_{min}$ for $(C{=}O)_{i-1}(N\text{-}H)_{i+1}$ as a function of ϕ and ψ for the glycyl dipeptide. Note the clear, strong minimum around the $\phi = -90°, \psi = 90°$ region, (and the symmetry related form at $\phi = 90°, \psi = -90°$) where a hydrogen bond is made. In this and all subsequent figures sidechains are deliberately not considered as discussed in the text. The effect of adding them would be to exclude conformations where $\phi > \pi/3$ and $\psi < -\pi/3$: leaving them off allows properties inherent in the main chain to be studied.

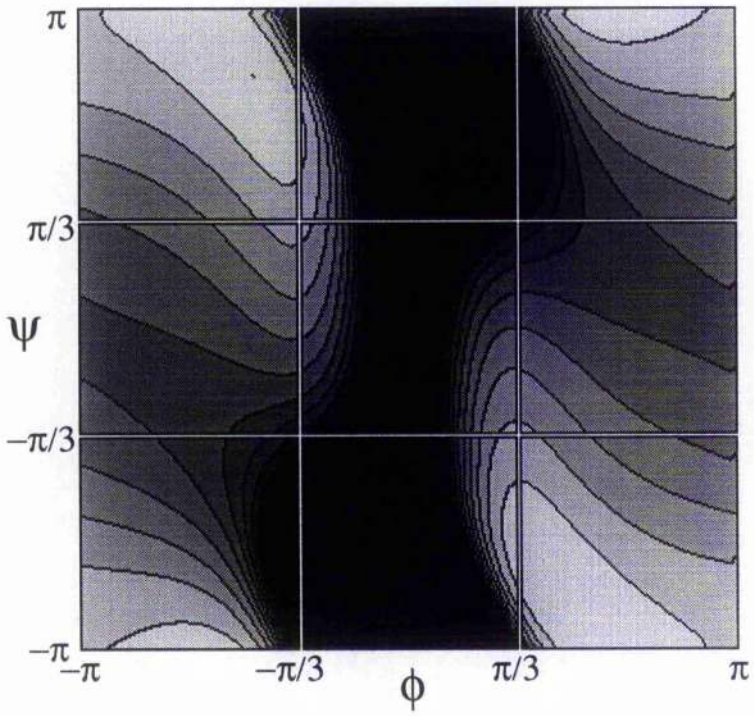

Figure 3.7: Energy of the Carbonyl-Carbonyl Interaction.
This shows a striking and somewhat unexpected result. The $(C{=}O)_{i-1}(C{=}O)_i$ interaction has a clear minimum above the line $\phi = -\psi$, in the region of right-twisted beta sheets and polyproline helices. This is a significant interaction, and depending on its strength relative to the other three cases could exert a considerable directing effect on all extended strand conformations.

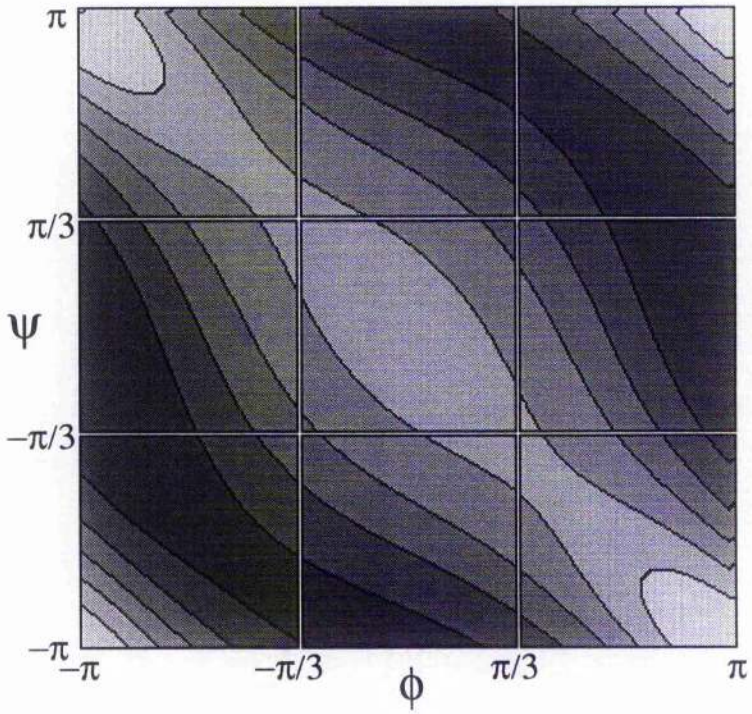

Figure 3.8: Energy of the C_5 Conformation
As expected, this interaction shows a minimum at $\phi = \psi = 180°$ where the H...O distance is shortest. Notice that, since the effective van der Waals radius of hydrogen is zero in the force field used, there is no strong repulsive core in this interaction.

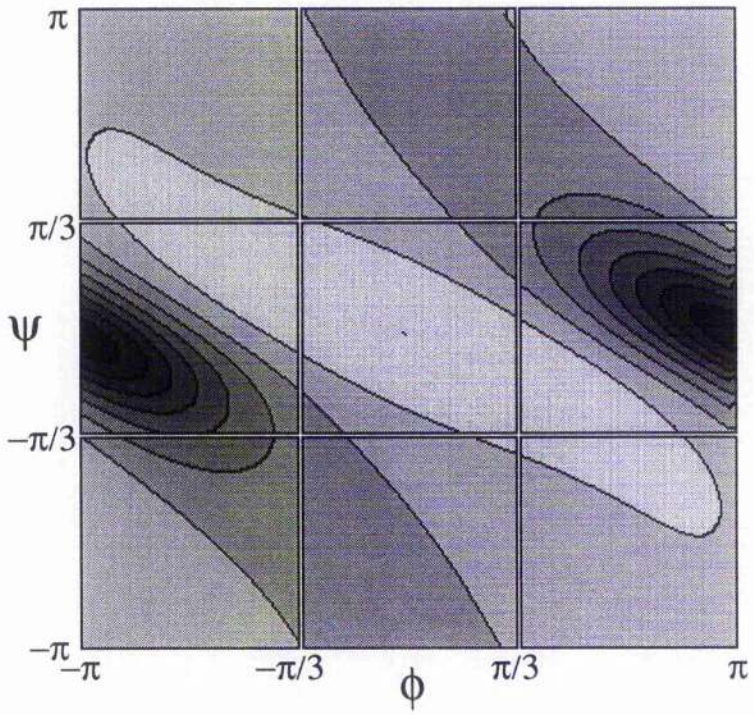

Figure 3.9: Energy of the N-H...N-H Interaction.
As discussed, this interaction is significant because it is not *a symmetry related form of fig3.7. The altered charges and atom sizes conspire to make a quite different potential energy contour - with a minimum not as deep or as sharp as for the C=O...C=O case. In this force field, the traditional "neck region" exclusion ($\psi \approx 0°$) is not present, as this is ascribed to H...N collisions, not possible with this force field.*

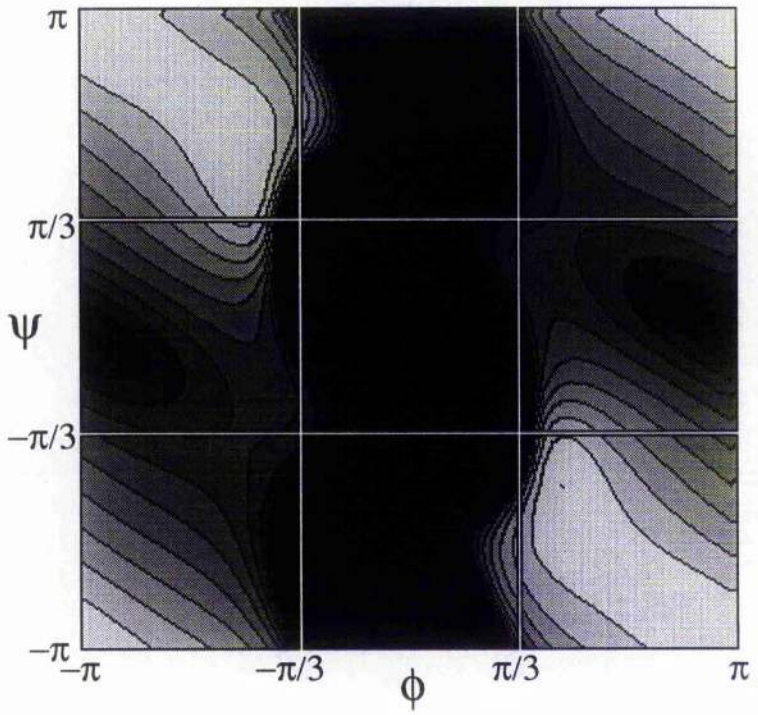

Figure 3.10: Conformational Energy of the Glycyl Dipeptide.
This shows the result of a simple linear addition of the results in figures 3.6 to 3.9. At first appearance, the results are dissapointing: the interaction is dominated by the gamma turn hydrogen bond, and any asymmetry due to the C=O...C=O interaction is masked by this effect. Compare this result with the next figure: clearly something is lacking from this model. Notice that although accessible, the alpha helical region $\phi = -60°, \psi = -40°$ is a saddle point, not a local minimum.

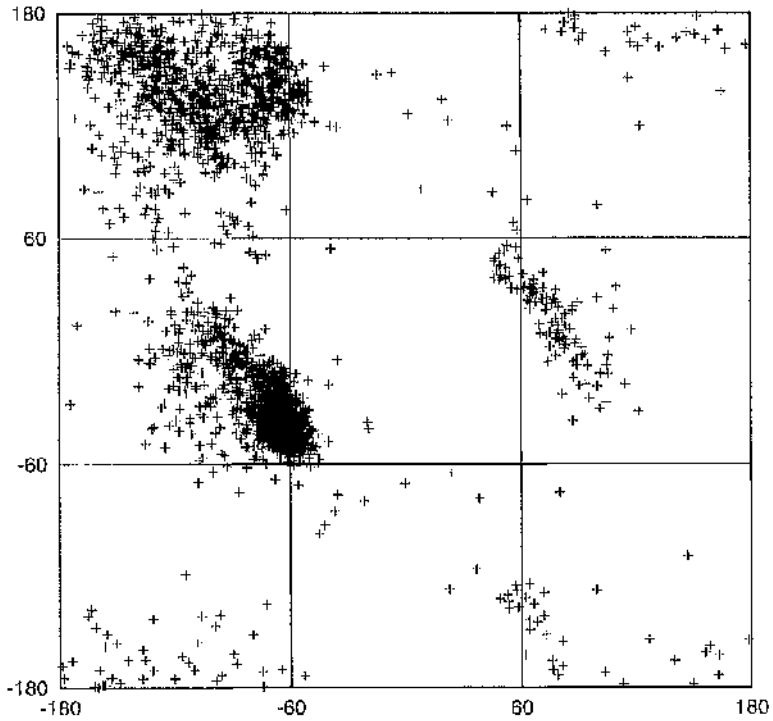

Figure 3.11: Observed ϕ, ψ values.
This figure shows backbone torsion angles taken from 10 high resolution proteins in the Protein Data Bank, including glycine. Contrast this with figure 3.10: in particular, notice that the real values are excluded from region of the gamma turn apart from a small cluster of single residues at the centre of that region.

the force fields adopted here), it seems strange that the observed phi/psi distribution shows no marked concentration at the gamma turn region. Is it possible that simple force fields are missing some significant interaction? It would seem not. Even though the $(C{=}O)_{i-1}(C{=}O)_i$ interaction seems to be significant and may give rise to a local energy minimum in a different region, there is thus far nothing to suggest that it is anything other than a secondary interaction, and reviews of all commonly used force fields agree that the gamma turn is the most stable conformation for the dipeptide. There must be some external constraint present in proteins which displaces the gamma turn.

Clues to the nature of this constraint can be identified in several previous works. First, the observation that 70-80% of all possible hydrogen bonding sites are satisfied in proteins is crucial. In regions where the protein is in an extended conformation, it will be hydrogen bonded to a distant part of the protein, the substrate, or water molecules. Second, when gamma turns are found in proteins they tend to be in characteristic positions, most frequently at the ends of beta strands. This is a situation where the hydrogen bonds tend to become frayed and donor-acceptor lengths are longer than elsewhere. This suggests that typical hydrogen bonds in some way interfere with gamma turn conformations, and the gamma conformation is only stable where the residues involved cannot make any other hydrogen bonds.

This suggests a simple test which can be carried out. If we assume that each of the peptides in our dipeptide system has its full hydrogen bonding potential met, we can make a reasonable guess as to where the atoms from the external bonding source lie by assuming linear hydrogen bonds with optimal geometries. The Ramachandran plot can then be redrawn with these new atoms included and the effects of tertiary hydrogen bonding seen.

The models were constructed as shown in figure 3.12. One bulky atom was added to represent the hydrogen bonding partners of each of the four sites in the dipeptide. For the hydrogen atoms, the partners were treated as water or carbonyl oxygen atoms, and placed 1.0 Angstrom away along the lines defined by the N-H bonds. For the oxygen atoms of the dipeptide the partner should be a hydrogen atom, but the force field in use assigns hydrogen a radius of zero. To model the blocking group, the hydrogen bond donor atom was used instead. To simplify the system, the same oxygen parameters were used: the partner atoms were placed 1.6 Angstroms along the direction of the C=O bond. In each

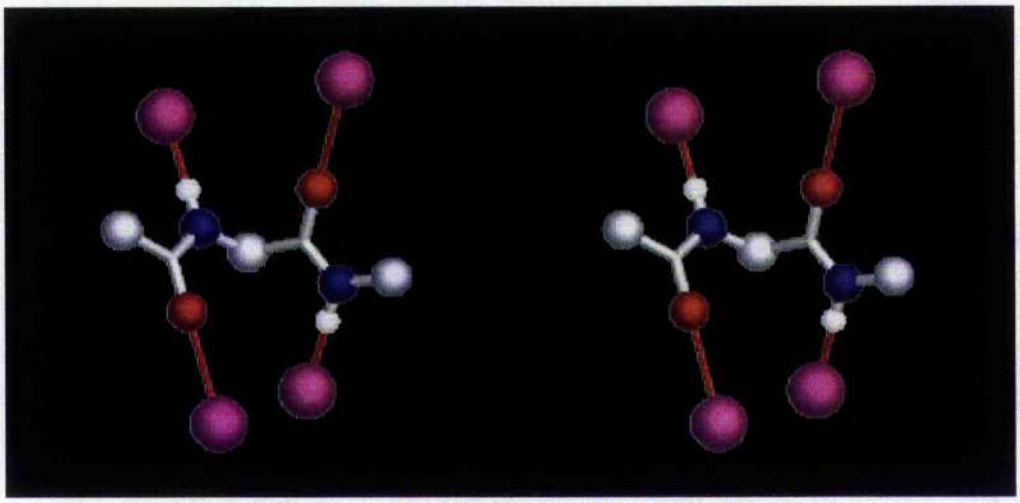

Figure 3.12: Tertiary Bonding Partners for the Dipeptide.
This shows the dipeptide with virtual tertiary structure, placed to mimic perfect hydrogen bonding surroundings as discussed in the text. The conformation here is a twisted, extended beta strand. In the beta-pleated (gamma turn) conformation the partner *of* NH_{i+1} *would clash with* O_{i-1}.

partner atoms.

The system can be interpreted in a number of ways, either as a fully solvated strand of an unfolded protein or a strand buried in a folded protein with hydrogen bond partners provided by other strands or sidechains. To concentrate on the properties of the strand under investigation, and to allow for the cases where one pair of partners, donor and acceptor, actually correspond to a single water molecule acting as both species, collisions between the partner atoms were not considered. This means that any changes seen in the potential energy were down to collisions between partner atoms and parts of the dipeptide.

3.6 Results ii. Partner-blocked conformations

Figure 3.13 shows the simplest version of this system which can account for the absence of the gamma turn. Ignoring the charge on the hydrogen bond acceptor (and indeed it turns out that any atoms other than that acceptor can safely be ignored), the steric blocking of the non-bonded atom can be seen to interfere with the gamma turn sufficiently to completely prevent it in this type of system.

to completely prevent it in this type of system.

Once the effect is seen, it can be easily interpreted. Clearly, the gamma turn arises when the carbonyl oxygen of residue i-1 is near to the amide hydrogen of residue i+1 as discussed before. When the amide hydrogen is already hydrogen bonded to another atom (usually oxygen), the remote atom fills the site where the bond to O_{i-1} would normally be. The gamma turn is easily displaced wherever any other hydrogen bonding candidate is found: figure 3.10 gives the gamma turn region less than 0.5 kcal/mol stabilisation relative to the other strand conformations, while a secondary or tertiary structure hydrogen bond is given an energy of around -5 kcal/mol if the geometry is acceptable. This has great significance when considering the importance of the carbonyl-carbonyl interaction, because the dominating effect of the gamma turn is removed and the remaining accessible regions of the conformational space are dominated by the carbonyl interaction. It also shows a general principle in the interpretation of forces in proteins, that protein structure is rarely defined as a local minimum of a simple system but rather by a set of choices between a range of secondary and tertiary stabilisations, balanced by statistical and the entropic effects of exclusion.

3.7 Conclusions. Beta Sheet twist is enhanced by electrostatic effects.

The analysis in this chapter suggests that there are two significant and easily represented effects which combine to favour the intrinsic right handed twist of beta strands and hence of beta sheets. Previous studies have concluded that a combination of entropic effects and side chain packing interactions provide the main source of the twist. The possibility of its being an intrinsic strand property has been ignored, due partly to the misconception that the peptide-peptide interaction is approximately symmetric about $\phi = -\psi$.

In fact, there is considerable asymmetry induced by the carbonyl dipole which favours increased twist, and when the dominant gamma turn interaction is excluded (as it must be in most cases where the main chain has hydrogen bonds to solvent, other chain segments, or sidechains) then the twisted strand becomes the minimum energy conformation.

A simple model of solvation or tertiary bonding has been shown to have considerable use as a structural tool. One possible avenue for further investigation would be to ex-

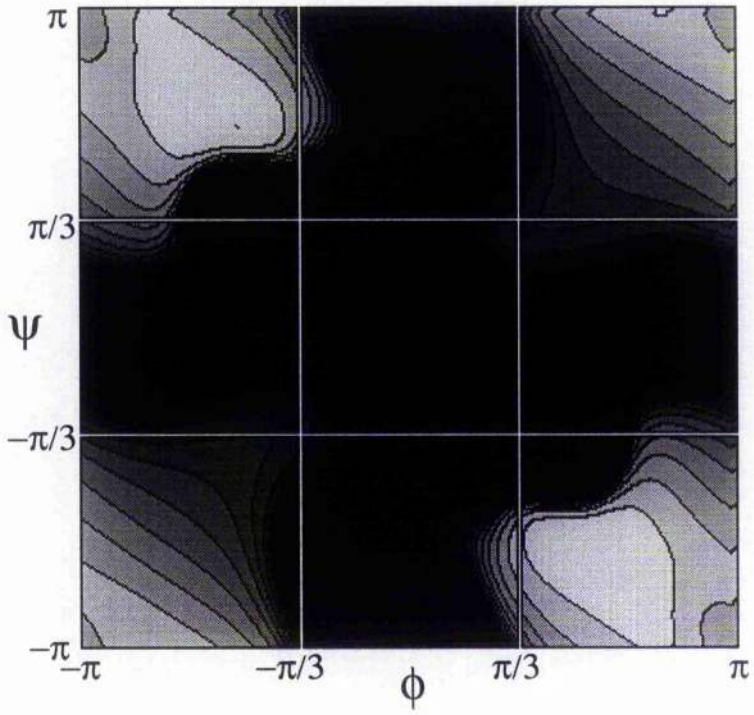

Figure 3.13: Conformational Energy of Tertiary-bonded Dipeptide.
Using the tertiary bonding partners shown in fig3.12, with no charge but the steric parameters of secondary nitrogen, the gamma turn region is completely excluded. Now compare the shape of the minimum with fig3.11 – the contour from $\phi = -120°, \psi = 60°$ to $\phi = -60°, \psi = 90°$ is an almost perfect match for the real data points. It is now possible to see the asymmetry introduced by the carbonyl-carbonyl interaction: the results suggest a spontaneous twist in the observed direction not only for $\phi < -\pi/3, \psi > \pi/3$ but even for $\phi > \pi/3, \psi < -\pi/3$, a glycine-only region.

tend the realism of this external binding model, perhaps even develop a mean field with combined continuum and discrete properties as a tool for assessing the solvation energy of model structures (which could also be used in substrate binding studies or even long timescale molecular dynamics, where water interactions are often the dominant calculation even in studies of small proteins).

3.7.1 Situations of gamma turns in proteins

Since the gamma turn is predicted as the energetic minimum for the beta strand yet is rarely seen, its observed pattern of occurrence must explicable, and the suggestion that it is excluded by secondary and tertiary hydrogen bond effects should be supported by evidence from real proteins.

A survey of gamma turns in proteins classifying them as "strong" (those with $E<1.0 kcalmol^{-1}$) or weak ($E\geq1.0 kcalmol^{-1}$) shows a significant set of distribution preferences which corroborate the suggestions made here.

Extended sets of gamma turns ("compound gamma turns") are seen, but only very rarely and then only of the weak kind. An extended strand is almost always right twisted and part of a sheets hydrogen bonding network, never a "pure" compound gamma turn stabilised only by this interaction.

Among strong gamma turns (which are nearly always isolated), fully 68% occur at the N or C termini of strands or helices, with strand termini roughly twice as common as helix termini. This is consistent with their only occurring where other significant stabilising interactions force a section of the polypeptide backbone to do without external hydrogen bonds, allowing the intrinsic minimum of the backbone to be adopted.

3.7.2 Increased twist predicted, and observed, in proline rich strands

In order to test the effect of the carbonyl-carbonyl interaction and its significance for the polypeptide backbone, it would be useful to remove one or other of the interactions in the polypeptide chain and see the effects on protein structure. There is of course one residue which does this, proline. Since there is one amide N–H group missing in backbones including proline, the directing (and hence twisting) effect of the carbonyl interaction is predicted to be considerably stronger. It has frequently been observed that proline residues

adopt left-handed helical conformations corresponding to to the part of the beta region occupied by the collagen triple helix.

Proline is constrained to adopt ϕ values around 60° by its side-chain, which confuses the issue a little by requiring that residue to have a left-handed helical conformation when in extended structures. However, a review of regions including proline (not just the proline residues themselves) [33] has shown that residues around the proline residue also show increased twist as predicted. In particular, residue $i-1$, where i is proline, has no N–H...N–H interaction, and also cannot make a gamma turn type interaction. As a result this residue would be expected to show a strong left-handed twist guided by the remaining strong interaction, the carbonyl-carbonyl interaction, and this is indeed observed. It is interesting that short proline rich regions with a twisted strand conformation are frequently seen in proteins, and this work suggests that inclusion of proline in a region of a protein has a very strong directing effect on the conformation that region will adopt.

The polyproline conformation encourages external hydrogen bonding by exposing the strand hydrogen bond donor and acceptor sites optimally, but the presence of proline means it is not favoured as part of a beta sheet because of the missing hydrogen bonds: hence a proline rich region is solubilised, as water can provide the necessary hydrogen bonds without sacrificing any of the hydrogen bonding potential of the protein.

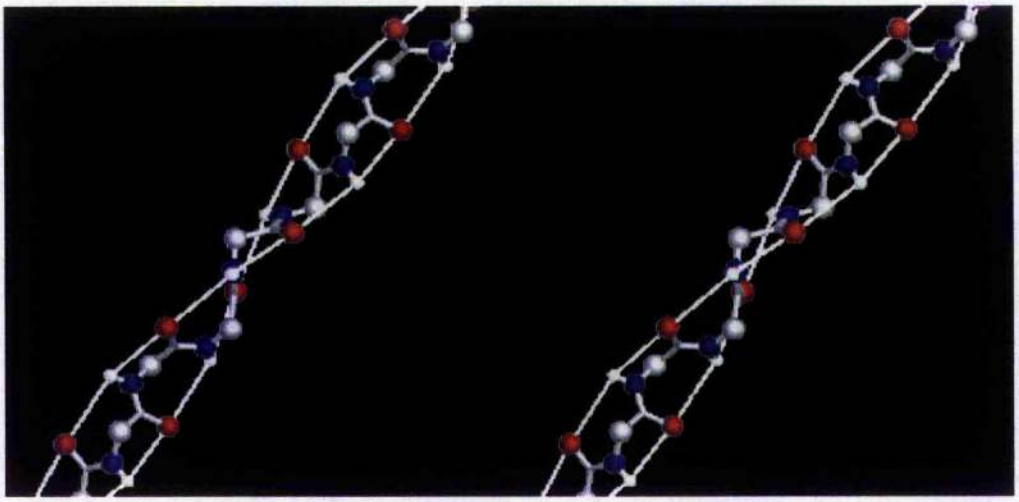

Figure 3.14: A Right-Twisted Beta Strand.
This figure shows the most commonly observed conformation for the beta strand, slightly twisted in a right-handed sense (clockwise looking along the backbone from N' to C'). See the figures in chapter 8 for the effect of extending this twist to whole sheets. $\phi = -90°, \psi = 130°$. *The white lines are simply guides to show the sense of the twist.*

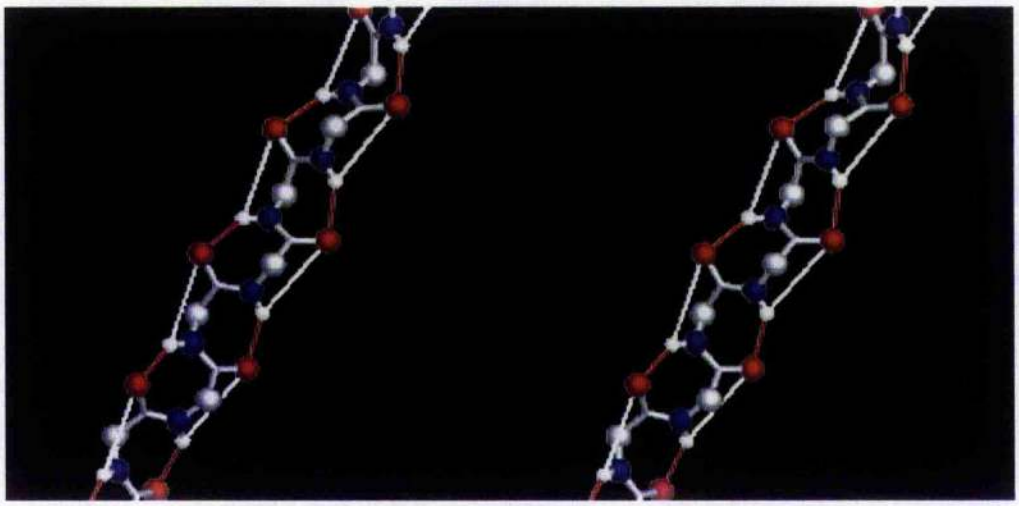

Figure 3.15: A Pleated "Gamma turn" Strand.
This shows the minimum energy extended structure for the polypeptide chain in the absence of solvent or tertiary structure. Only very short (typically 1-2 residue) stretches of this conformation are ever seen in proteins. $\phi = -90°, \psi = 70°$. *White lines are simply guidelines, red lines are hydrogen bonds fulfilling the Baker & Hubbard criteria.*

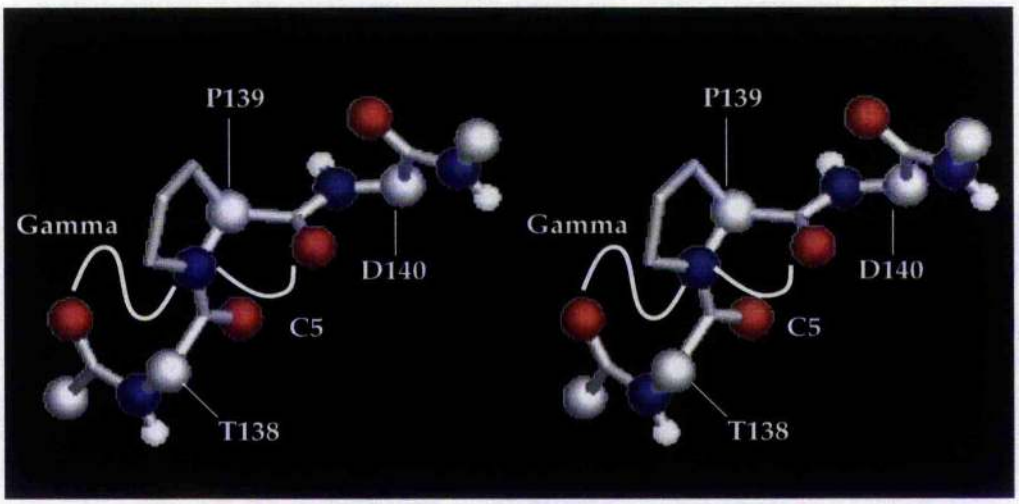

Figure 3.16: A Polyproline type II helix.
The backbone of a polyproline type II helix, T138/P139/D140, from Cytochrome c peroxidase. The twist on strands involving proline is enhanced by the C=O...C=O interaction. For Pro_i, residue i-1 has no gamma turn interaction while i has no N-H/N-H interaction. Both of these effects are twist-enhancing.

Chapter 4

Hydrogen Bonding Geometries for Peptides and Polar Sidechains

Extending interaction plots to all polar sidechains shows that the simple atom-centred potential is adequate to explain observed hydrogen bonding geometries in nearly all cases. Both donor and acceptor groups can be modelled as two atom probes, and constraining them to always be coplanar with the sidechain groups and always point at the approximate centre of charge gives potential energy contours which closely match the observed distribution of hydrogen bonding partners for each sidechain.

The approach is particularly successful in the cases of arginine, where the predominance of three centres for hydrogen bonding is shown to be a simple consequence of the charge distribution of arg$^+$, histidine, where the hydrogen bond acceptor geometry is correct without extra directional constraints again as a consequence of the charge distribution alone, and asparagine and glutamine, where the high acidity of the anti *hydrogen is shown to be a consequence of the carbonyl dipole.*

The main failure identified is in carboxylate sidechains and to a lesser extent the hydrogen bond acceptor activity of asn and gln, where the absence of explicit lone pair charge centres means that the two centres of hydrogen bonding around each carbonyl oxygen become a single broad region.

For main chain peptides, the charge distribution gives rise to a dipole correction to the simple hydrogen bonding geometry, which has been measured as an angle γ describing the in-plane deviation from ideal hydrogen bonding geometry. Using a peptide as a probe and as

the fixed group in interaction plots allows the observed value of gamma for parallel sheets, antiparallel sheets and helices to be explained as a consequence of simple electrostatics.

4.1 Introduction

A picture of hydrogen bonding as electrostatic interactions between polar residues, and a method of estimating their energy for any given geometry, allows us to look at the detailed behaviour of polar residues in proteins. In addition to the polypeptide backbone, which will be studied separately in some depth, many amino acid side chains are polar and have strong tendencies to form hydrogen bonds, sometimes with water molecules but largely within the body of the protein.

The hydrogen bonding residues are cysteine, histidine, asparagine, glutamine, serine, threonine and tyrosine, which can act as donors or acceptors, aspartate and gluatamte which are only acceptors at cellular pH, and lysine, arginine and tryptophan, which act as donors only. Some hydrophobic residues have significant local charge densities, and aromatic residues display interesting electronic interactions involving their pi orbitals, but these are outwith the scope of this work. Here we will only consider polar interactions which closely follow the classical model of the hydrogen bond or have significant effects on hydrogen bonded structures.

The first question to be addressed is, "what is the preferred hydrogen bonding geometry for the binding sites of the sidechains of these residues?" This is quite a hard problem to address even for a simple electrostatic or combined electrostatic/Lennard-Jones potential. The problem is, even with the constraints imposed by the polypeptide backbone and bonds holding the sidechain in place, any of the possible donor/acceptor pairs may be found in proteins, in nearly any conformation. Some attempts have been made to address this problem, notably the exhaustive work of Thornton et al which seeks to tabulate all the observed pairwise interactions and make some deductions from their spatial distribution.

There is a simpler approach to this problem. The potential functions we are using describe a set of energies associated with points in a phase space with six dimensions for every pairwise interaction, corresponding to displacement in the { x,y,z } directions and rotations about the { x,y,z } axes of the two units relative to each other. It is clear that interpretation of these results depends on being able to reduce the dimensionality of the

system to two or three so that the potential energy surface can be easily visualised.

In studying the effect of hydrogen bonding on protein conformation, we are helped by the observation that in crystal structures hydrogen bonding geometries are often near to the predicted optimal geometry. Also, we are interested in the general properties of sidechains, and can investigate individual interactions separately when systems of interest are identified. This suggests using a generic 'hydrogen bond donor' or 'hydrogen bond acceptor' to probe the potential energy space.

4.1.1 The electronic distribution on protein sidechains

One area in which quantum mechanics is in marked conflict with the empirical method of Lifson et al is the charge distribution, especially for peptides. The empirical force field was generated with the apparently reasonable assumption that the hydrogen bond donor and acceptor groups had a net zero charge. This gave a stable result and a transferable force field, but is significantly at odds with the actual picture of the charge distribution as found from a quantum analysis of the charge density of many polar groups. For the side chains, many better models of the charge distribution exist. Among the first exhaustive charge derivations for biomolecules was that carried out for the AMBER force field [6] using a 6-31 G* basis set on individual amino acids, and these charges were used here in conjunction with the empirically derived Lennard-Jones 9-6 parameters of Lifson *et al* [23]. Details of all parameters and charges used are given in appendix A.

4.1.2 Charged residues

A full description of polar interactions in proteins would be incomplete without some treatment of charged residues. There are four sidechains which may be charged at intracellular pH, Lysine, Glutamate, Aspartate and Histidine. Glutamate and aspartate are carboxylic acids, and lysine is a primary amine. Under most conditions (excluding the extremes of pH found in some regions of the gut, for example), these three will adopt the same charge as they would free in aqueous solution, glutamate and aspartate losing a proton to become negatively charged, lysine gaining one to become positively charged. The ionisation potential or electron affinity can be affected by the local environment of a residue in a protein, and this effect can be observed by examining the dependency of

overall charge of the protein on pH, but in practice only histidine commonly titrates at physiological pH. (It is commonly found in active sites which require a temporary sink for protons or electrons - the precise behaviour of the system can be tailored by surrounding the active histidine with charged or polar residues to control its ionic state in the native enzyme.)

It may be expected that charged interactions in proteins would be among the most important in determining protein stability, but this is not the case; in fact, salt bridges are quite rare in globular proteins, and are very poorly conserved as compared to hydrogen bonds, for example [34]. This seems counterintuitive at first, since a lysine-carboxylic acid interaction has an electrostatic interaction energy an order of magnitude stronger than that for the other polar interactions considered in this work. However, with fully charged residues there are a number of effects related to interactions with water which must be taken into consideration. Ions in solution are stabilised by an extensive shell of water molecules extending 3 to 6 angstroms from the ion itself. A buried charge loses this shell entirely, and a pair of opposite charges interacting but exposed to water at the protein surface have their solvation shells disrupted. It is not currently possible to assess the contribution of water to these interactions quantitatively, but the fact that salt bridges are poorly conserved suggests that the net contribution to protein stability is close to zero.

In all four of the charged residues, the unit of charge is not concentrated on a single atom but is rather distributed as partial charges across a number of atoms. For this reason it is necessary to consider the whole region over which this charge is distributed rather than the sort of two atom system usually associated with hydrogen bonding, but other than that hydrogen bonds involving polar residues can be treated the same way as other polar interactions, given that the quantitative result of any energy calculation is tempered with caution given the contribution of water outlined above.

4.2 Methods: Interaction plots

In order to get a better picture of the interactions which stabilise side chain interactions, each side chain needs to considered individually. The dictionary of side chain interactions shows the observed geometric dependence of interactions for each side chain. This work shows that the systems are hard to interpret because there are so many free parameters

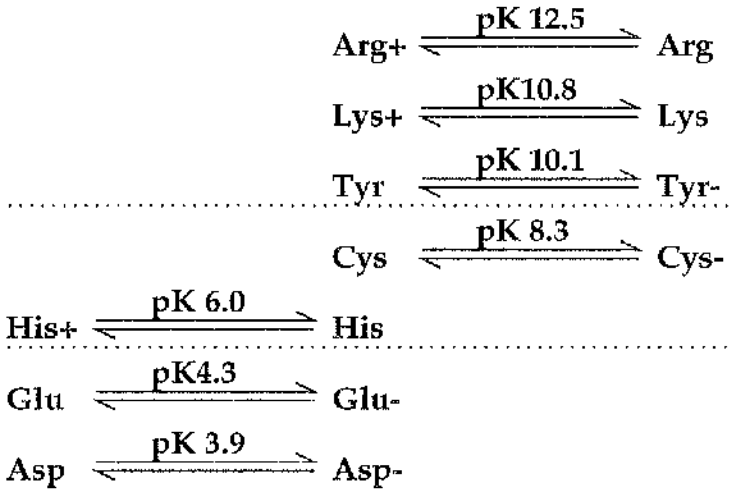

Figure 4.1: Oxidation States of Polar Sidechains
At normal physiological pH, only Histidine and Cystine show variable oxidation states. The pK values for individual residues may vary due to local electrostatic effects – an important form of reactive site tuning for many enzymes.

even in a single pair of interacting groups. (The dictionary looks at all residues, not just hydrogen bonding groups and finds a similar complex situation for each of them.)

Chapter 2 shows that it is possible to study these interactions by imposing constraints which in most cases hold the hydrogen bonds in the most favourable conformation - generally pointing at the hydrogen bond acceptor (when using a donor probe) or donor (when using an acceptor probe). In the systems considered here there is often more than one donor or acceptor atom to take into account, and in these cases a more central atom is chosen as the target. In most cases this means that the probe will also be pointing directly at the correct donor or acceptor atom when the four-atom geometry is optimal, although in some, low-angle deviations from optimal geometry may mean that the minima found are not the global minima for the system. It seems reasonable to assume that so long as conformations in the areas of the global minima have been found, these deviations will not be significant.

9-6-1 potentials are used throughout this chapter, with parameters and charge distributions as given in appendix A.

4.3 Results: The interactions of polar sidechains

The potential energy plots presented here are an attempt to see if the potential energy calculations being used throughout this work are useful for all polar interactions in proteins. In each case the interaction plot is for a fully coplanar system, because in all cases the energy minimum is likely to be a coplanar system. In one or two cases (most notably for arginine) there are likely to be significant stabilising interactions perpendicular to the plane, but by far the most common stabilising effects in proteins come from systems which are almost coplanar, or at least show deviations of less than 30° from planarity.

The results should be compared with the tabulations of interactions carried out by Singh and Thornton [35] and Ippolito et al [27].

4.3.1 Interaction potentials for charged sidechains

Interaction plot for Arginine(Arg^+)

Arginine is a hydrogen bond donor, and the preferred arrangement of hydrogen bond acceptors is best observed using a C=O probe system in which the carbonyl group is always pointing towards the centre of the group, the C^ζ atom; the results of this calculation are shown in figure 4.2.

Ippolito et al suggest a nomenclature for the three distinct regions where hydrogen bond acceptors are observed to cluster, $N^{n2}H\rightarrow$:anti$_I$, $N^{n1}H$,$N^{n2}H\rightarrow$:syn, $N^{n2}H$,$N^{\epsilon}H\rightarrow$:anti$_{II}$, and these three positions are shown to have strong minima. In the scatter-plots the syn and anti$_{II}$ minima are peanut shaped, with two barely discernible centres in each minimum rather than an oval as shown here, but this could be attributed to the interactions where {twin nitrogen}/{twin oxygen} arginine/carboxylate pairs form. Otherwise, the syn/anti$_I$/anti$_{II}$ preferences are largely proportional to the number of hydrogens contributing to the interaction locally, and hence also to the additive local positive charge of those hydrogens: the potential shown is an almost perfect match for the observed hydrogen bond acceptor distribution.

Interaction plot for charged Aspartate and Glutamate(Asp^-,Glu^-)

Aspartate and Glutamate both have similar charge distributions in the force field used, and so only one plot (with N-H as the probe, constrained to point at the carboxyl carbon)

is shown. There are no sp^2 directing effects in this model, suggesting that explicit lone pair charge placement may be important – although the width of the potential may mean that several hydrogen bond donors are always present and statistical effects of crowding give the sp^2-directed distribution seen.

The survey of Ippolito et al showed that there was significant favouring of the lone pair directed positions, which is not reproduced in the results here, but also showed that the ratio of syn to anti hydrogen bonding showed a distinct preference, with a syn:anti ratio of 0.51:0.49 for aspartate, 0.57:0.43 for glutamate, and an overall 0.53:0.47 preference for syn hydrogen bonding. The syn position is clearly favoured by the electrostatic model used here.

The preference for syn hydrogen bonding is strongest with arginine and lysine, supporting the conclusion that the directing influence is mainly electrostatic. Even among other side chains the ratio is higher than 0.53:0.47. This means that the main anti contribution comes from mainchain hydrogen bonds, which in many cases means a highly constrained system – particularly for aspartate, which very frequently makes hydrogen bonds to the local backbone peptides, possibly unable to make an ideal electrostatic contribution.

Interaction plot for charged Histidine(His^+)

Charged Histidine has two hydrogen atoms which act as hydrogen bond donors, and figure 4.3 shows the potential energy of a C=O acceptor probe pointed at the centre of the Histidine ring. The results are as would be expected for individual N-H groups. The scatter-plots of Ippolito et al could not distinguish the three forms of histidine as they appear identical to X-ray crystallography, so there is no simple way to judge whether this potential energy form is a perfect match. There is no significant cluster of hydrogen bond partners between the two regions shown in this figure, though, so it is fair to assume that this is a good match for those among the histidine residues which are charged.

4.3.2 Interaction potentials for charged sidechains

Interaction plot for Uncharged Histidine as Hydrogen Bond Donor

Figure 4.5 shows the hydrogen bonding potential for histidine with N^δ protonated - although no C^β atom is included so this is almost indistinguishable from the N^ϵ case. Here

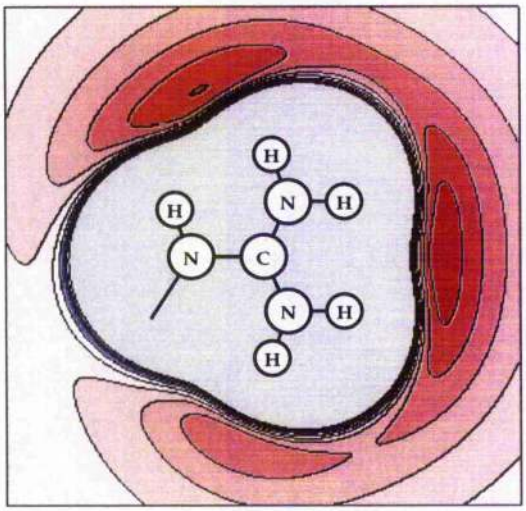

Figure 4.2: Interaction plot for Arginine(Arg^+)
Arginine can be treated as a hydrogen bond donor, and the ideal arrangement of hydrogen bond partners can be observed using a C=O probe and a 9-6-1 potential as in figure 2.11, with the probe always pointing at the centre of the charged group. Contours are drawn at 2kcal/mol intervals.

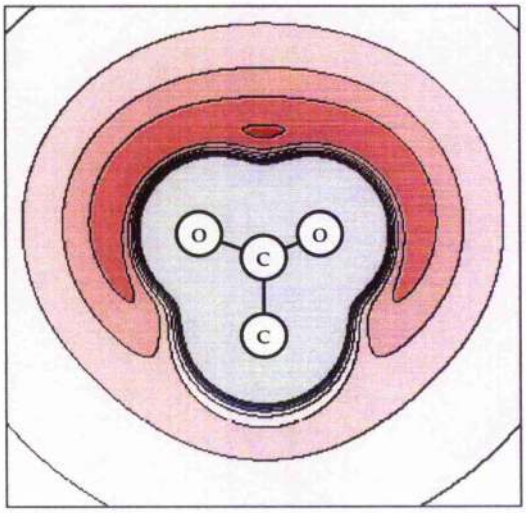

Figure 4.3: Interaction plot for charged Aspartate and Glutamate(Asp^-,Glu^-)
Plot of potential energy of aspartate and glutamate sidechains as hydrogen bond acceptors, with N-H as probe. Contours are drawn at 2kcal/mol intervals. Note especially that there are no sp^2 directing effects in this model.

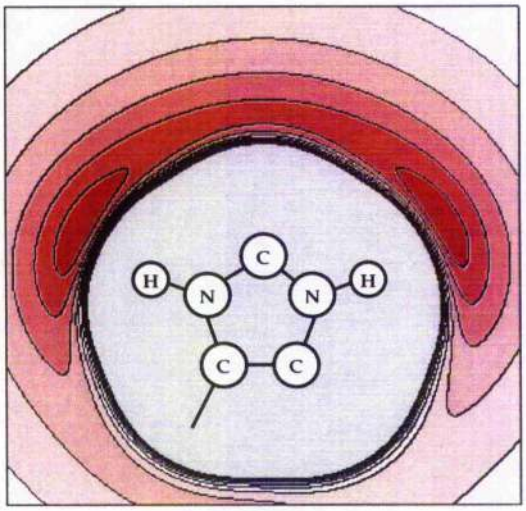

Figure 4.4: Interaction plot for charged Histidine(His^+)
Charged Histidine has two hydrogen atoms which can be regarded as hydrogen bond donors, and this plot shows the potential energy of a C=O acceptor probe. Contours are at 2kcal/mol intervals.

Interaction plot for uncharged Histidine as Hydrogen Bond Acceptor

Figure 4.6 shows the same system, with the same charges and parameters as in figure 4.5 but using N-H as the probe to identify hydrogen bond acceptor regions. The strong clustering of hydrogen bond donors around the N^{δ} atom was thought to indicate a distinct lone pair electron concentration, but the results here suggest that the ring charge distribution alone is enough to encourage a directional hydrogen bond. The carbon atoms flanking N^{δ} are slightly positive, and this has a significant directing effect.

Interaction Plot for Asparagine and Glutamine as Hydrogen Bond Donors

Asparagine and Glutamine polar groups have identical charge distributions in this force field, so figure 4.7 serves for both. In the first calculation, hydrogen bond donor activity is measured and so a C=O probe is used. The carbonyl dipole is significant here, providing an added electrostatic contribution to any $A...H^{anti}N$ hydrogen bond. This effect is experimentally verified by the fact that *anti* hydrogen atoms of carboxamide groups are more acidic than *syn*: the carbonyl group dipole gives the *syn* hydrogen a higher effective positive charge.

imentally verified by the fact that *anti* hydrogen atoms of amide groups are more acidic than *syn*: the carbonyl group dipole gives the *syn* hydrogen a higher effective positive charge.

Interaction Plot for Asparagine and Glutamine as Hydrogen Bond Acceptors

Using an N-H probe identifies hydrogen bond acceptor regions; the results are shown in figure 4.8. As in the aspartate/glutamate case, there is no sp^2 directing effect observed. The clustering seen is weaker than that for the hydrogen donor positions. Overall the syn/anti ratio for both donor and acceptor groups is 0.46:0.54, but the directing effect of the lone pairs is less significant than that of the amide protons: the strongest directing effect is that of the anti N-H allied with the C=O dipole.

Interaction Plot for Serine and Threonine as Hydrogen Bond Donors

X-ray crystallography does not provide explicit hydrogen positions for alcoholic sidechains, and most studies of hydrogen bond partners focus on techniques for deducing the hydrogen position. Figure 4.9 shows the hydrogen bond donor region for alcoholic sidechains and suggests the hydrogen bond acceptor distribution which could be seen if hydrogen positions were known.

Interaction Plot for Serine and Threonine as Hydrogen Bond Acceptors

For both serine and threonine the most significant clustering seen is the steric hindrance which causes donor and acceptor clustering in the staggered conformations. The expected preference for the *trans* position is not seen, probably because local sidechain - mainchain hydrogen bonds dominate, favouring the gauche conformations. The cases where alcoholic side chains act as both donor and acceptor at the same time are not best envisaged by superimposing the two diagrams given here, but would require a more complex three dimensional function balancing the preference for the staggered conformations with the need to keep donors and acceptors of opposite polarity as widely separated as possible.

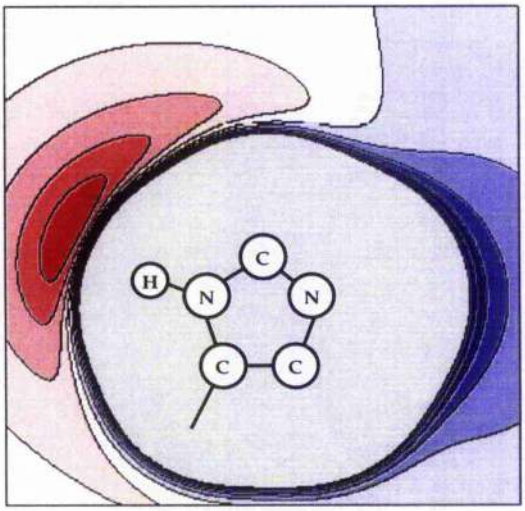

Figure 4.5: Interaction plot for Uncharged Histidine as Hydrogen Bond Donor
Uncharged Histidine can exist in two forms, with N_δ or N_ϵ protonated. Both forms have similar charge distributions, so a model with no explicit joint to the main chain is considered here to cover both cases. C=O is used as probe in this calculation.

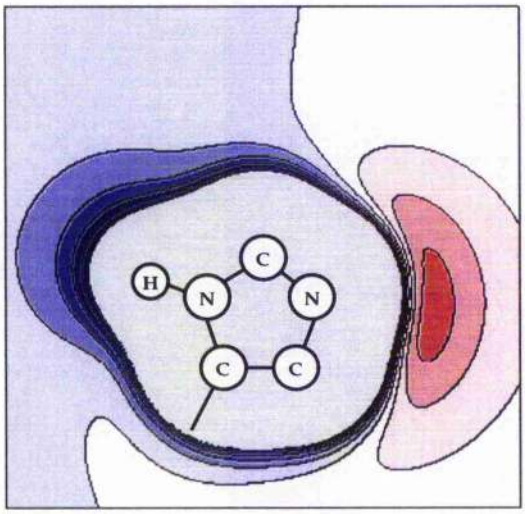

Figure 4.6: Interaction plot for uncharged Histidine as Hydrogen Bond Acceptor
Using N-H as a probe, a centre of negative charge on N^ϵ flanked by positive charges is identified as a hydrogen bond acceptor. It has a distinct geometry, tailored by the positively charged atoms on each side. Contours are at 2kcal/mol intervals.

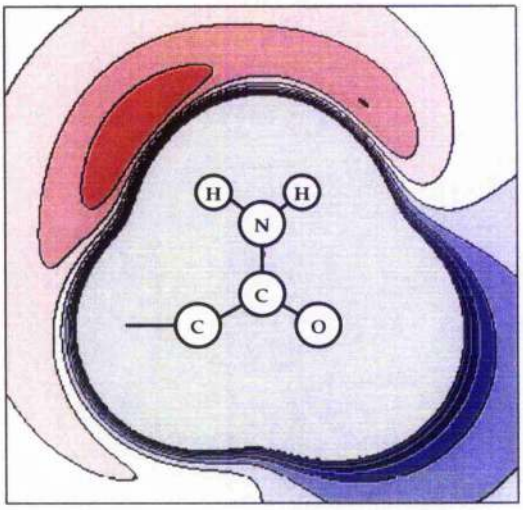

Figure 4.7: Interaction Plot for Asparagine and Glutamine as Hydrogen Bond Donors
C=O is the probe here, this time pointing at the centre of the C-N amide bond. Notice the relative strength of the anti hydrogen as donor here compared with the syn hydrogen, thanks to a directing effect of the carbonyl dipole.

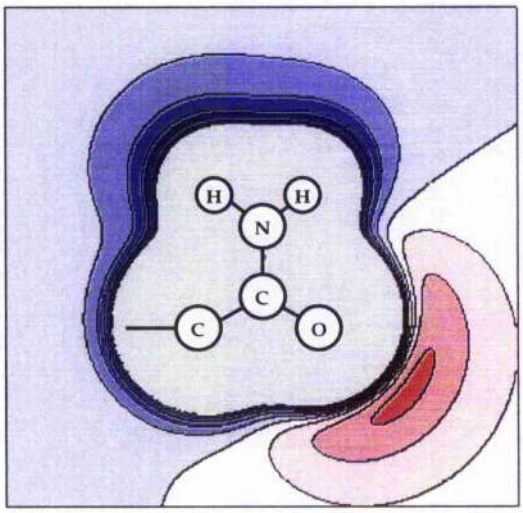

Figure 4.8: Interaction Plot for Asparagine and Glutamine as Hydrogen Bond Acceptors
Asparagine and Glutamine can act as hydrogen bond acceptors as well as donors, and shown here is the same system as in the previousfigure but with N-H as probe. As in the aspartate/glutamate case, there is no sp^2 directing effect observed.

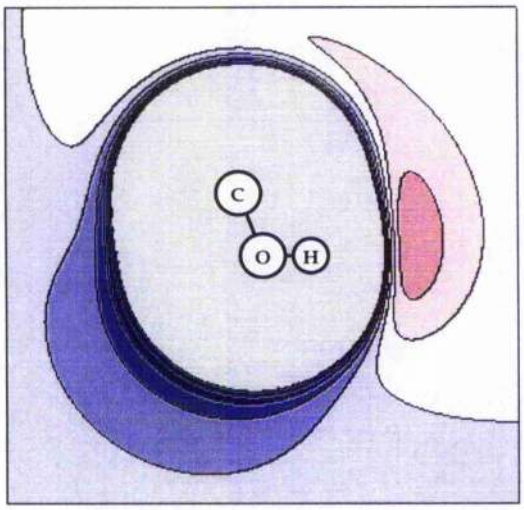

Figure 4.9: Interaction Plot for Serine and Threonine as Hydrogen Bond Donors
This figure suggests the hydrogen bond acceptor distribution which could be seen if hydrogen positions were known, with C^{β}-O^{γ}-H in the plane and a C=O probe directed at he oxygen atom.

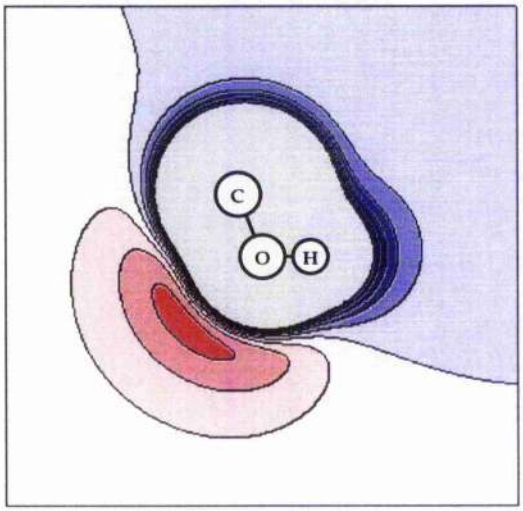

Figure 4.10: Interaction Plot for Serine and Threonine as Hydrogen Bond Acceptors
Partial charges on the oxygen atoms are enough to make these sidechains act as hydrogen bond acceptors even in the absence of explicit lone pair charge centres. Here N-H is the probe.

4.4 Results ii: Secondary and tertiary peptide/peptide interactions

Sidechains are not the only polar parts of proteins which show geometry dependent polar effects based on their partial charge distributions. The peptide itself consists of atoms with effective partial charges, and these give rise to effects which are distinguishable from simple hydrogen bonds. The case of two adjacent peptides has been considered in the previous chapter, but it there are also other places where polar interactions other than the hydrogen bond can have a conformational directing effect when the peptides are more widely separated along the polypeptide chain. Both in secondary and tertiary mainchain-mainchain interactions there are carbonyl-carbonyl effects which cause deviations from linear hydrogen bonds.

The structure of the peptide bond can only be explained in terms of the orbital structure. A more traditional picture of the peptide bond, as a resonance structure, actually predicts that the peptide would be unable to take part in hydrogen bonding, since the canonical form with a C=N double bond whose contribution would explain the planar nature of the bond would give rise to a positive charge on the nitrogen atom, and hence no tendency to form the sort of charge distribution which allows N-H to act as a hydrogen bond donor. Clearly something else is happening: N-H *does* act as a hydrogen bond donor, but at the same time the peptide bond is planar.

The answer lies in the structure of the molecular orbitals formed from the p-orbitals of nitrogen, carbon and oxygen. Rather than forming an sp^3 hybrid structure, the nitrogen lone pair mixes with the pi bonding p-orbitals of the carbonyl bond to form a set of extended orbitals which favour the flat bond and encourage increased electron density on the electronegative nitrogen and oxygen atoms. This gives a system which, when treated as point charges mapped onto the atom centres, has negative centres on N and O, positive on C and H, as required for electrostatic hydrogen bonding. Notice also that the charge on the nitrogen is not equal and opposite to the charge on Hydrogen, but is considerably higher. This is because nitrogen has an electron withdrawing effect from the alpha-carbon also, so there is some deviation from the simple picture of electroneutral groups acting as hydrogen bonding donors or acceptors.

The potential of Lifson et al [23] is unusual in that it assumes net neutral N–H and

C=O groups, while most other force fields such as AMBER [6], based on *ab initio* quantum mechanical calculations, have larger negative charges assigned to the amide nitrogen atom, balanced by a small positive charge on the C^{α} position.

In (Maccallum et al II, [14]) a comparison of these two contrasting charge distributions and their effect on the geometry of peptide/peptide interactions is shown. In this case there is very little difference between the two systems, so it is safe to take the electroneutral approximation as a starting point (which turns out to be vital for easily interpreting results from whole proteins) to investigate systems, with the more realistic charge distribution used to verify the generality of the more important results. This is useful because being able to treat peptides as neutral objects means that forces between them can be estimated in the absence of other atoms, giving a "clean" potential energy estimation allowing the intrinsic dynamics of the polypeptide backbone to be studied without need to estimate the effects of other atoms.

4.4.1 The angle γ in peptide interactions

In proteins, this nonlinearity has been measured by an angle, γ [36], which is the *in plane* C=O...H angle in a peptide to peptide hydrogen bond, positive when the hydrogen is nearest the C_{α} atom, negative when it is nearest the nitrogen of the acceptor peptide. It has been shown that in α-helices and parallel beta sheet its value is usually negative, while in antiparallel beta sheet it is mostly positive.

To show that carbonyl-carbonyl electrostatic interactions can have the correct directing effect on peptide hydrogen bonds, a more elaborate probe system needs to be used. In this case, we are interested in the interactions between *all* the atoms in the pair of interacting peptides. The probe and the fixed group both need to be full peptides (including the C_{α} atoms). Rather than attempting to guess the optimal hydrogen bond geometry as is possible for simple two-atom probes, the geometry is fixed so that the acceptor C=O and donor N–H are in each case antiparallel, so that the co-linear hydrogen bond can be formed and any other stabilised arrangement easily interpreted. The probe can have two orientations, one parallel to the fixed peptide, as it would be in parallel sheet or alpha helices, one antiparallel as it would be in antiparallel beta sheet.

Figures 4.12 and 4.13 show the plots for these two conformations. The energy minima are found with the NH and CO groups oriented for hydrogen bonding, but not co-linear.

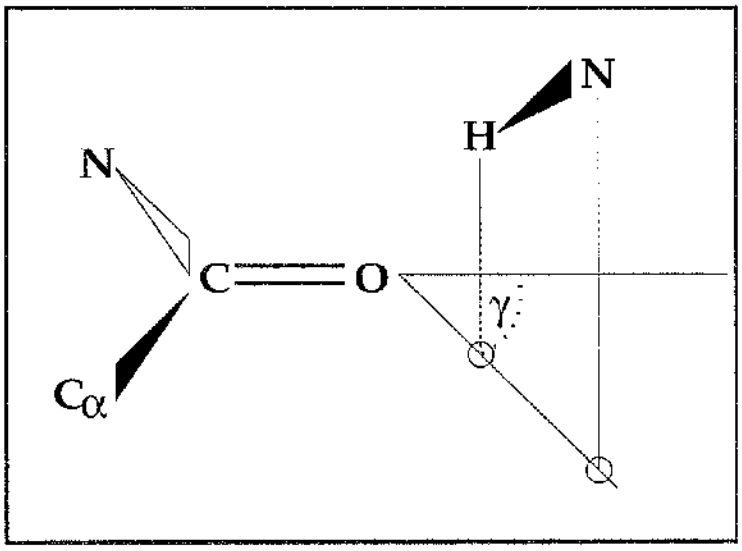

Figure 4.11: The angle γ definition
γ is the angle C=O...H in the plane of [$C^{\alpha}C^{o}O^{c}$] for a hydrogen bond to a polypeptide backbone. Note that N-H...O are not necessarily co-linear.

The relative positioning of the groups at the minimum is such that in each case the carbonyl oxygen of the lower, fixed group is as close to possible to the carbonyl carbon which has a permanent fractional positive charge. In the parallel case this gives rise to a minimum with a gamma angle of -23°, close to the observed values of -19° for parallel beta sheets, -20° for alpha helices. The minimum for the antiparallel case is 15°, comparable to the observed value of 20° in antiparallel sheets in real proteins.

4.5 Conclusions

A full interpretation of the hydrogen bonding sites of a protein would require looking at each in turn and calculating the contributions from all nearby charged or polarised atoms. Looking at isolated polar groups and calculating the hydrogen bonding potential with simplistic probes gives loose predictions of the hydrogen bond partners' distribution which is surprisingly close to that actually seen.

Two conclusions about the quality of the results can be drawn. One is that atom-centred potentials with no explicit directional terms are sufficient to reproduce the observed distribution of hydrogen bond geometries. No specific directional terms seem to

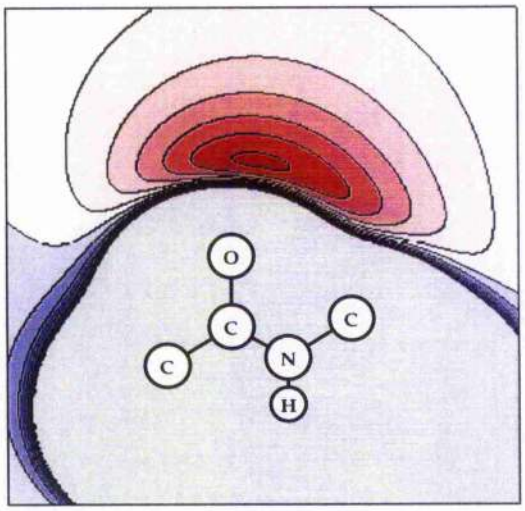

Figure 4.12: Peptide-peptide interactions: parallel
This figure shows the effect on the interaction plot of including the extended charge set of a whole peptide rather than just the C=O and N-H units. The probe peptide is in the same orientation as the fixed one, parallel and coplanar with potential energy plotted relative to the hydrogen position. The minimum is shifted to give a negative value to γ, *with the minimum corresponding to* $\gamma = -23°$.

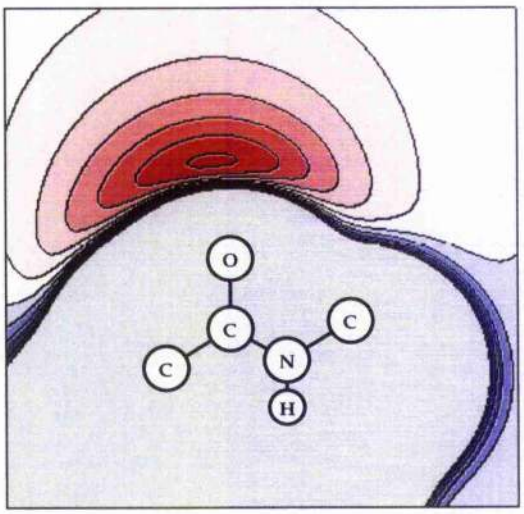

Figure 4.13: Peptide-peptide interactions: antiparallel
This shows the same system as the previous figure, but with the probe peptide now turned over to correspond to an antiparallel stand rather than a parallel one. The minimum now gives a postive value for γ, $+15°$.

be required. The other is that lone pairs, particularly on carboxylate groups but to a lesser extent on all unsaturated oxygens, are significant and should perhaps be included as separate points with partial charges of their own.

Chapter 5

The Stability of Different types of Helix

Extending energy calculation based analysis of peptide interactions to systems with more than two groups and two free parameters requires a new approach to reduce the size of the possible solution space. This chapter describes the first step in this direction, applying constraints to an infinite polypeptide chain which require the bonds of each residue which are free to rotate to adopt the same conformation. The conformational energy of an infinite polypeptide homo-conformer can then be calculated.

Each contribution to the conformational energy is examined, classifying each pair of interacting peptides by the corresponding hydrogen bond interaction. Individually, the i+3, i+4, and i+5 give minima for the expected helices (3_{10}, α_R, and π_R respectively), but in combination they show that there is no local minimum for the 3_{10} helix – it will always fold to an α_R helix if possible – and there is a strong steric repulsion destabilising the π_R helix, explaining in a simple to interpret system why extended helices of these types are never seen in proteins.

Another feature of the combined potential energy is that there is a significant barrier to concerted strand to helix transition, suggesting that helix folding must be either catalysed, stepwise, or both. Models dealing with this problem are the subject of chapter 6.

5.1 Introduction

Some of the intrinsic properties of the polypeptide can be explained in terms of the classic Ramachandran plot dipeptide interactions as discussed in chapter 3, and some of the long range effects seen in beta strands and helices can be explained as seen in chapter 4. However, looking in particular at the figures in chapter 3, it is clear that there is no interaction which specially stabilises the alpha helical conformation, even though that region of the phi/psi plot is accessible. This means that only longer range forces have a significant directing influence on helical conformations, which has significant implications for the folding pathway of proteins; if there is no intrinsic tendency for the residues on their own to adopt helical conformations, helix folding can only be some type of cooperative process. This chapter uses the techniques developed in earlier chapters to look at the individual polar contributions to the conformational energy of a *polypeptide*, rather than the glycyl dipeptide.

5.1.1 The three predicted helical structures

Beta sheets and strands only comprise one major class of protein structure. As early as the 1940s it was suggested that the polypeptide backbone could be stabilised by hydrogen bonds along the length of the chain to form helices. The structures which were believed to be possible included

- the 3_{10} helix,
- the α or 3.6_{13} helix and
- the π or 5.4_{17} helix

These will be described fully in the next section, but the basic features are:

The 3/10 helix is stabilised by main chain hydrogen bonds three residues apart. Individual residues are frequently found in the conformation predicted for the right handed 3/10 helix, but extended 3/10 helical structural elements are not found in globular proteins. Most 3/10 helices found are either distortions in alpha helices, often at the ends, or are independent but less than five residues long. The same situation is found in small peptides, although α-disubstituted amino acids can be made which force the peptide to adopt a 3/10 helical conformation.

The alpha helix is stabilised by hydrogen bonds between main chain groups four residues apart. This is the only common helical structural element found in proteins: it was one of the earliest structures confirmed and nearly all globular proteins contain one or more long alpha helices. Some are comprised entirely of alpha helical units connected by loops.

The pi helix should be stabilised by hydrogen bonds between main chain groups five residues apart. No pi helices are found in proteins. There are a few places where alpha helices include a widening to incorporate a five-residue hydrogen bond, but the conformation of the residues involved is not that predicted for an extended pi helix. In fact, even individual residues rarely adopt the pi-helical conformation, although the dipeptide phi/psi plot suggests that such conformations should be possible.

5.1.2 The concerted folding model

The *concerted folding model* is a picture of helix folding in which all of the residues which are to be involved in a helix adopt that conformation cooperatively, with all the residues participating from the start of folding through the change from random coil to helix. It was originally invoked to explain the results from time-resolved circular dichroism which show that 100% (or more) of the final helical content of a protein is adopted within a few microseconds of the start of folding. It was felt that the only way this could be explained was through a cooperative process.

At one extreme, the concerted folding model could be interpreted as implying that all the residues fold in a straight path from random coil to helical conformation at the same rate, so if all of the residues were in (say) an extended conformation to start with, then at each time along the folding pathway all the residues would have similar phi/psi values. This is clearly an extreme interpretation, but not neccesarily an unreasonable one. A looser interpretation might be that helix folding may exhibit multiple nucleation points, with the other residues adopting complementary conformations in response to these as folding progresses.

The naive picture of concerted folding is useful, because it gives a restricted phase space which can be visualised in two dimensions. In this chapter, potential energy plots for individual interactions are calculated in the context of a polypeptide chain in which all the residues share the same phi and psi values. This provides an energy profile for a fully

concerted strand to helix transition, and although few hard conclusions can be drawn from this somewhat unlikely folding pathway, suggestions are made as to how those constraints can be relaxed: this is dealt with in chapter 6.

5.2 Methods: Analysing the conformational energy of concerted folding

Again, we are interested in systems which are not necessarily realistic for globular proteins, and forbidden conformations may have as much to tell us as those which are energetically stabilised. As a result we can make little use of real protein coordinates at this stage, and artificially generated coordinate sets are more appropriate.

Our main interest is in the folding from extended conformation to helix, and it would be a simple matter to generate coordinate sets just along a straight monotonic transition from one to the other. But it is in fact possible to generate full phi/psi plots of these interactions in a few minutes on a reasonably powerful workstation, so this is what has been done.

5.2.1 Non-bonded and Electrostatic Contributions to the Stability of Helices

Using a 9-6-1 potential for the interactions means that a range of effects can be seen. Not only are the classic hydrogen bonds identified, but other polar effects involving the amide carbonyl groups can be included alongside a reasonable treatment of non-bonded repulsions.

A significant amount of work has focused on the electrostatics of the alpha helix itself, in particular on the permanent dipole it exhibits and the apparent use of this in substrate and charged sidechain binding. In this work, this subject is not investigated because the long range charge cooperativity is dependent on polarisabilities and dielectric effects which cannot be treated by the Lennard-Jones 9-6-1 model. A conclusion about the relative importance of short range and long range electrostatic interactions would require a full *ab initio* quantum mechanical treatment, and so is not attempted here. Instead, only those interactions which involve direct contact between near peptides are treated, in which dielectric effects are likely to be small. Hydrogen bonds for peptides 1,2,3,4 and 5

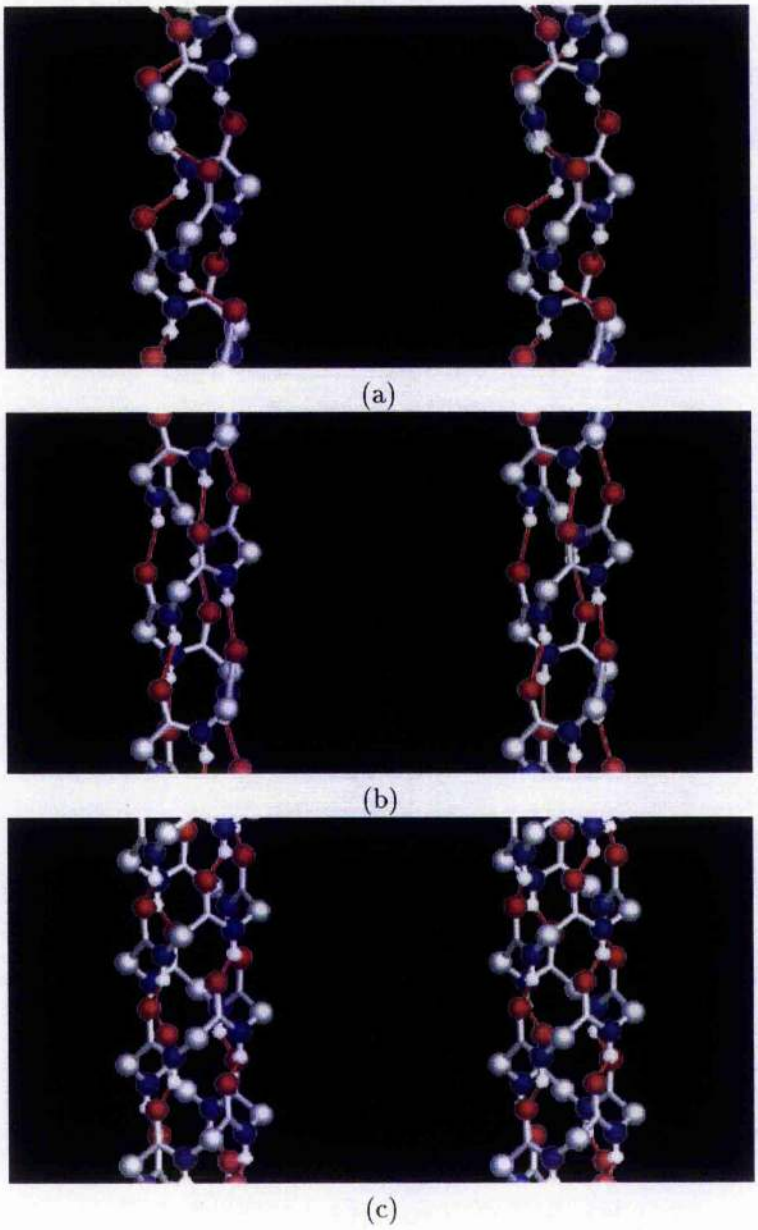

Figure 5.1: Classes of Right-handed Helix

The three classes of helix predicted on the basis of hydrogen bond geometry are shown here. Coordinates were generated as models, since extended stretches of 3_{10} and π helix are not observed in globular proteins. a) A 3_{10} helix, $\phi = -60^{o}, \psi = -30^{o}$, stabilised by $i \rightarrow i+3$ hydrogen bonds. b) An α helix, $\phi = -60^{o}, \psi = -40^{o}$, stabilised by $i \rightarrow i+4$ hydrogen bonds. c) A π helix, $\phi = -60^{o}, \psi = -65^{o}$, stabilised by $i \rightarrow i+4$ hydrogen bonds.

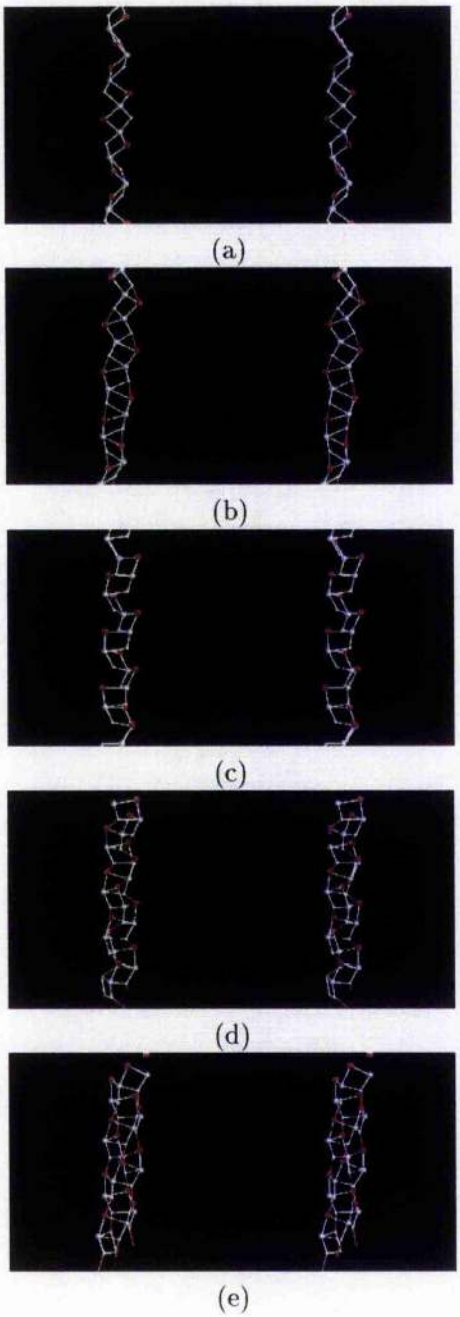

Figure 5.2: The Concerted Model of Helix Formation

This figure shows the conformations adopted by a polypeptide in which every residue folds from a twisted strand to an alpha helical conformation at the same rate. Only the outlines of the peptide bonds are drawn, with $C_{\alpha i}$, H_{ni}, O_{ci-1}, and $C_{\alpha i-1}$ marked. a) $\phi = -90^o, \psi = 130^o$, twisted beta strand. b) $\phi = -90^o, \psi = 70^o$, gamma turn conformation, with $i \rightarrow i+2$ hydrogen bonds. c) $\phi = -90^o, \psi = 10^o$, the "neck"region of the Ramachandran plot. d) $\phi = -65^o, \psi = -25^o$, a 3_{10} helix stabilised by $i \rightarrow i+3$ hydrogen bonds. e) $\phi = -60^o, \psi = -40^o$, an α helix stabilised by $i \rightarrow i+4$ hydrogen bonds.

residues apart are considered.

5.2.2 Charge Distribution and Non-Bonded Parameters for the Peptide Unit

As discussed in section 4.6, there are two ways of representing the charge distribution on a peptide. One is to use a QM derived set of partial charges, which implies the placement of an effective positive charge on the C^{α} position, the other is to use the empirical force field of Lifson et al [23] which has the amide atoms exhibiting an internal charge redistribution but no polar effects beyond the bond itself. The second was shown to be quite adequate, with the correct patterns of repulsion and attraction and minima in broadly the same positions as given for the *ab initio* case. It also has the great advantage that the polypeptide backbone can be broken down into its constituent peptide units and potential energies calculated between these independent objects. Since the potential energies have a considerable electrostatic component, it is useful to have zero net charge on the groups being considered, and for as few atoms to be shared between the functional units as possible to prevent double-counting of potential energy contributions. In fact the parameters (including the repulsive cores) for the C_{α} atoms can be left out, ensuring that any results are purely interactions between peptides, making them much easier to interpret.

These results below show that even with this simplification, nearly all the features of repetitive polypeptide conformation can be explained, and there is little reason to suppose that adding in the extra carbon atoms would add much to the final picture.

5.2.3 Co-ordinates for Model

The calculations in this section were originally suggested by examination of the interactions between peptides in real polypeptide backbones from the Brookhaven database, but analysis of the more complicated aspects of real structures is saved for the next chapter, where the situations under which $i \rightarrow i+3$ and $i \rightarrow i+5$ hydrogen bonds can naturally occur are identified. The features which arose from that analysis which are relevant for this section are

- the peptide-peptide interaction which corresponds to an $i \rightarrow i+3$ hydrogen bond has a strength roughly 60% that of an equivalent $i \rightarrow i+4$ hydrogen bond
- the situations where $i \rightarrow i+5$ interactions arise are distortions of alpha helices, not residues in pi-helical conformations.

The coordinates are generated using the scheme given in Appendix 1, with the peptides planar (ie angle $\omega = 180^o$) and all bond lengths and angles fixed to standard values [5]. The conformations are fully described by the phi/psi values for the backbone, and the potential energy surface is therefore two dimensional.

Using a SPARC based Sun4 workstation, it is possible to generate 2250 sets of coordinates and their interactions in around 100s, even using unoptimised code, so it is possible to generate a full range of 150×150 phi/psi conformations for each interaction.

5.3 Results: The conformational energy of homogeneous helical structures

Using whole peptides rather than N–H...O=C interactions can lead to conflict with the usual nomenclature used for hydrogen bonds, but in the case of 3/10, 3.6/13 and 5.4/17 helices the strongest part of the peptide-peptide interaction is contributed by a correctly oriented hydrogen bond, so it is possible to use the same labelling scheme. This means that adjacent peptides sharing one alpha carbon atom can be described as having and $i \rightarrow i+2$ type interaction, those with one intervening amide bond have an $i \rightarrow i+3$ type, those with two intervening amides $i \rightarrow i+4$, and so on.

The significance of using whole peptide interactions has already been discussed in chapter 4, in that the whole charge distribution of a single bonding unit contributes to the distribution of hydrogen bonding partners, not just the individual hydrogen bonds which have been picked out in the past. Using a 9-6-1 potential as well as point charges allows the effects of steric hindrance to be quantified, and unfavourable interactions caused by overlap repulsion can be seen.

5.3.1 The $i \rightarrow i+2$ interaction stabilises the beta strand conformation

The interaction between two adjacent peptides was discussed at length in chapter 3. The same results are presented in figure 5.4 for completeness: the interaction is slightly different from the others in that it requires special treatment of the three-bond interactions: as before, the electrostatic effects are kept but the Lennard-Jones 9-6 parameters are removed. Thus, the results in figure 5.4 are the same as those in figure 3.10, the case without tertiary hydrogen bonds, namely a strong stabilisation of the gamma turn conformation and the right-twisted beta strand (and again its mirror image, since in this section we are only considering a polyglycine peptide). As is show later, this chain length stabilisation is quickly swamped by the stronger hydrogen bonds from the longer range interactions.

The quantitative significance of this plot is hard to assess. For this analysis the $i \rightarrow i+2$ interaction is a little sharp and so is softened by multiplying it by 0.5 before adding it to the other components.

5.3.2 The $i \rightarrow i+3$ interaction: stabilisation of the 3/10 conformation

Figure 5.4 shows the potential energy for a pair of peptides separated by one intervening amide group. This corresponds to the $i \rightarrow i+3$ hydrogen bond when $\phi \approx -60^{o}$ and $\psi \approx 0^{o}$, and the potential energy shows a strong minimum in this region as would be expected. However, comparing the shape of the potential to those in figures 5.7 and 5.8 shows that the minimum is not as well defined as that for the $i \rightarrow i+4$ interaction, and it occurs in a region where there are repulsions from the $i \rightarrow i+2$ interaction as well. The $i \rightarrow i+3$ cannot make its N–H...O=C interaction linear: there must always be a marked kink in the hydrogen bond which decreases its maximum possible stability. This is seen in section 5.3.5, looking at the overall energy of the helical conformations, where even in places where the backbone gets its primary stabilisation from $i \rightarrow i+3$ hydrogen bonds the strength is still only around 3 kcalmol^{-1} compared to around 4.5 kcalmol^{-1} for an average $i \rightarrow i+4$ hydrogen bond in an alpha helix.

In spite of these problems, figure 5.6 shows the effect of combining the $i \rightarrow i+2$ and $i \rightarrow i+3$ potentials. There is a fairly clear local minimum for the 3/10 helical conformation, so there does not seem to be any reason *at this stage* why an extended 3/10 helix should not be found.

It is even possible to include a simple model of competition for solvent, by including the tertiary bonding partners of the i+2 case (but not of the i+3). This is equivalent to allowing external species to bind to extended chain but excluding them as the strand coils up, making a reasonable model system the result of which is shown in figure 5.7.

5.3.3 The $i \rightarrow i+4$ interaction: the alpha helix

The alpha helical hydrogen bond has a nearly optimal geometry, and this can be seen in figure 5.6 as a clear, deep minimum at $\phi = -60°$, $\psi = -40°$. This is not surprising, of course, but the other feature which is of great importance in this potential energy is the steep repulsive region near $\phi = -70°$, $\psi = -50°$. This is a repulsion arising from steric hindrance, as in the conformation which this corresponds to, the carbonyl group of residue i is forced too close to the amide nitrogen of residue $i+4$. This means that three successive residues cannot adopt similar conformations near these phi/psi values, and so the pi helix is not possible. There is no stabilisation for a pair of neighbouring peptides adopting this conformation, as shown in figure 5.4, and the $i \rightarrow i+5$ stabilisation described below for four homo-conformer residues is not strong enough to overcome the $1/r^9$ or $1/r^{12}$ repulsive core of the $i \rightarrow i+4$ repulsion. There is therefore a simple explanation of why the pi helix cannot be formed by a polypeptide chain which is based entirely on the interactions along the length of the helix, rather than arguments based on the efficiency of the packing at the core or sidechains of the helix.

5.3.4 The $i \rightarrow i+5$ interaction: too weak for pi helices

With five peptides in the same conformation (which can be achieved by setting four residues to have the same phi/psi values) it is possible to see the $i \rightarrow i+5$ interaction. This has a minimum where the pi helix was predicted to fall, with another sharp repulsion corresponding to the 5-residue C=O...N steric repulsion at slightly lower psi values. However, if all the residues are in this conformation, the repulsion from the $i \rightarrow i+4$ interaction in the previous section must also be included, and as a result there is significant net *destabilisation* of the pi helix. The only way an $i \rightarrow i+5$ hydrogen bond can be formed is through considerable distortion of the chain (see chapter 6), and this means that $i+5$ bonds cannot be a repetitive secondary structure feature.

5.3.5 Overall conformational energy

Putting all these interactions together ought to show the energy profile of the concerted backbone, and this is shown in figure 5.8. This plot is simply obtained by adding together the $i+2, +3, +4, +5$ energies and displaying the results on a single phi/psi plot. Figure 5.8 should be compared with figure 5.5, which shows the same calculation but for the $i+2, +3$ interactions only. The minimum for the $i+3$ interaction is not as deep as that for the $i+4$ hydrogen bond. As a result, there is no energy barrier between the minimum for the 3/10 helix and that for the alpha helix. In fact, if three or more residues adopt the same conformation in a polypeptide, there is no minimum for the 3/10 conformation at all; it simply becomes a potential energy trough with a slope clearly in the direction of the alpha helical conformation. The 3/10 helix, which is only common in peptides with artificial blocking sidechains, is inherently unstable and will always tend to fold into an alpha helix. These results suggest that there is no energy barrier to this process and that it will happen *spontaneously*.

5.4 Conclusion: Concerted strand-helix transition is energetically unfavourable.

It is now possible to draw a potential energy diagram for the concerted strand to helix transition and draw some conclusions about the dynamics of such an event. Figure 5.9 suggests that a straight line across the Ramachandran plot, going from a gamma turn conformation to a helical conformation through the 3/10 region is as good a folding path as any, with no other stabilised folding channels identified by this method. This profile is shown in figure 5.10. What it shows is that there is no particular stabilisation provided by the concerted model along this path until an extensive 3/10 helix has been formed, from where the helix will spontaneously fold as suggested. This means that there is a permanent energy barrier to cross, estimated at 2.0 kcalmol^{-1} per residue. This is too high to allow the sort of rates typically seen in protein folding studies, so some more cooperative system which somehow avoids the unfavourable random coil with no helix or strand character intermediate which the concerted model requires. Possible candidates are the subject of the next chapter.

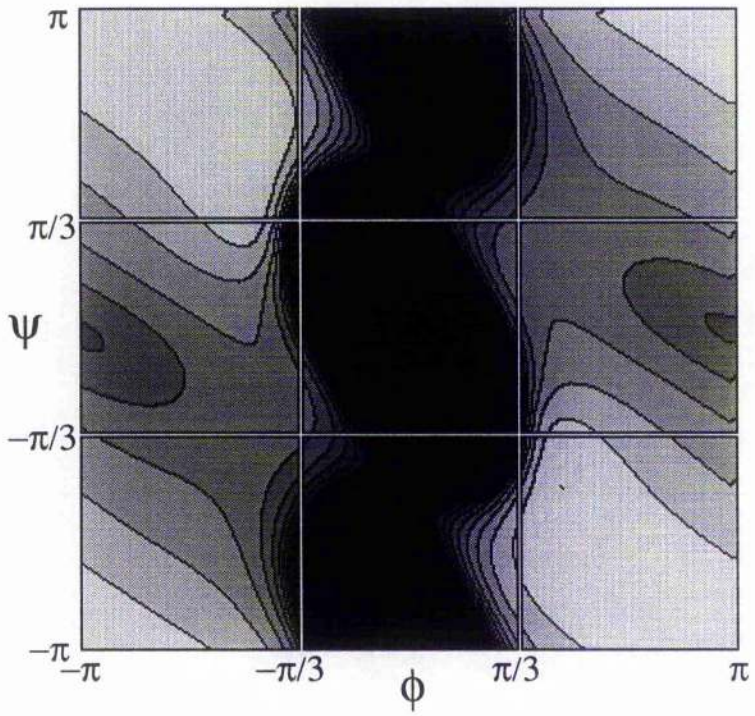

Figure 5.3: Conformational Potential Energy of the $i \rightarrow i+2$ interaction.
This is the interaction between two adjacent peptides, the same result as in figure 3.10 but divided by 2 to provide softening as described in the text. Contours are drawn at 1kcal/mol intervals.

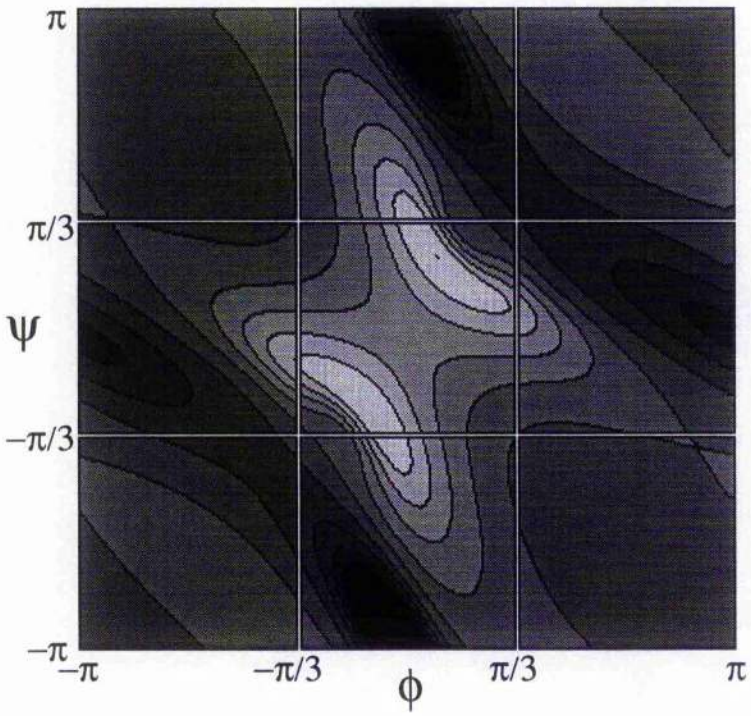

Figure 5.4: Potential Energy of the $i \rightarrow i+3$ Interaction

This shows the interaction energy between two peptides (separated by one intervening peptide, which is not included in the energy calculation) in a homo-conformer strand which is everywhere assumed to have the same ϕ, ψ values. The interaction is $(C_{Oi}, O_{Ci}, N_{Hi+1}, H_{Ni+1}) \rightarrow (N_{Hi+3}, H_{Ni+3}, C_{Oi+2}, O_{Ci+2})$ *The minimum, near $\phi = -40°, \psi = -40°$ corresponds to the strongest hydrogen bond which would form if the intervening peptide could be safely ignored.*

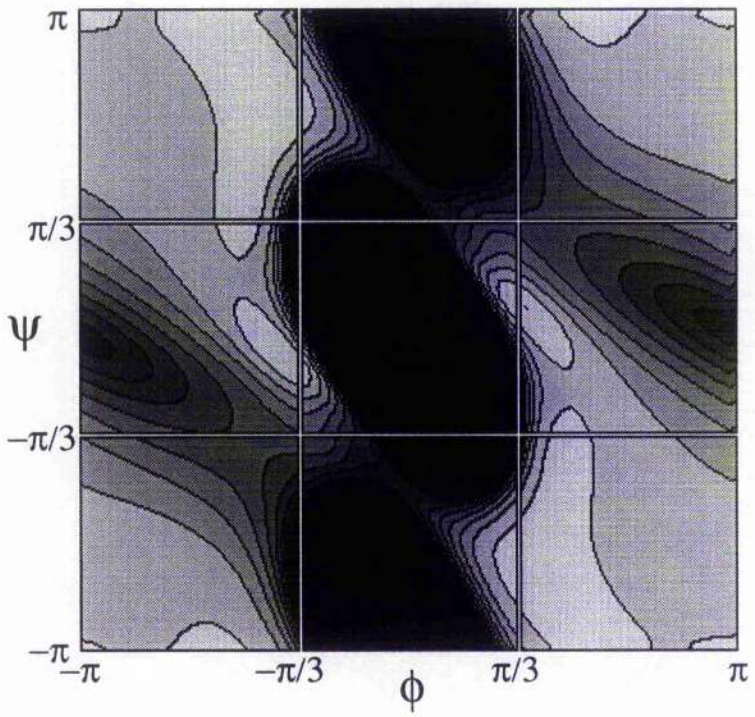

Figure 5.5: Combined $i \to i+2, i \to i+3$ Interaction
This is a simple addition of the results 5.3 and 5.4. The minimum of the $i \to i+3$ interaction is shifted to $\phi = -60°, \psi = -20°$ as observed in short 3_{10} stretches in proteins, while in the β conformations the i+3 interaction adds further electrostatic enhancement to the twisting.

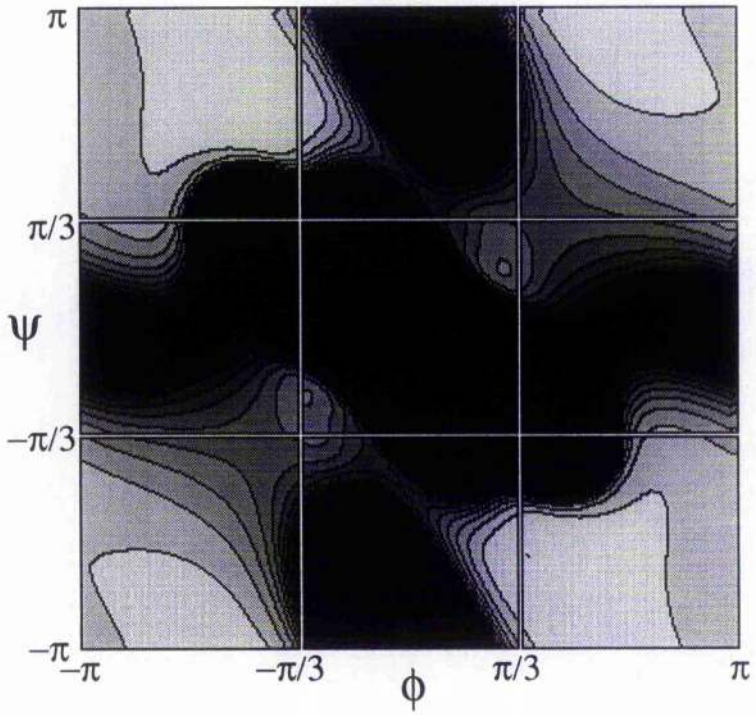

Figure 5.6: Combined $i \to i+2$,$i \to i+3$ Interaction with Tertiary Bonds
Adding the result from figure 3.12, ×0.5 rather than the results in figure 5.3 gives a quick "sketch" of a more elaborate model in which tertiary bonding is allowed in extended conformations but not in helical ones. A full treatment would need to consider the potential energy of breaking the tertiary hydrogen bonds: this figure is just a device to show the effect of excluding the gamma turn $i \to i+2$ hydrogen bond region.

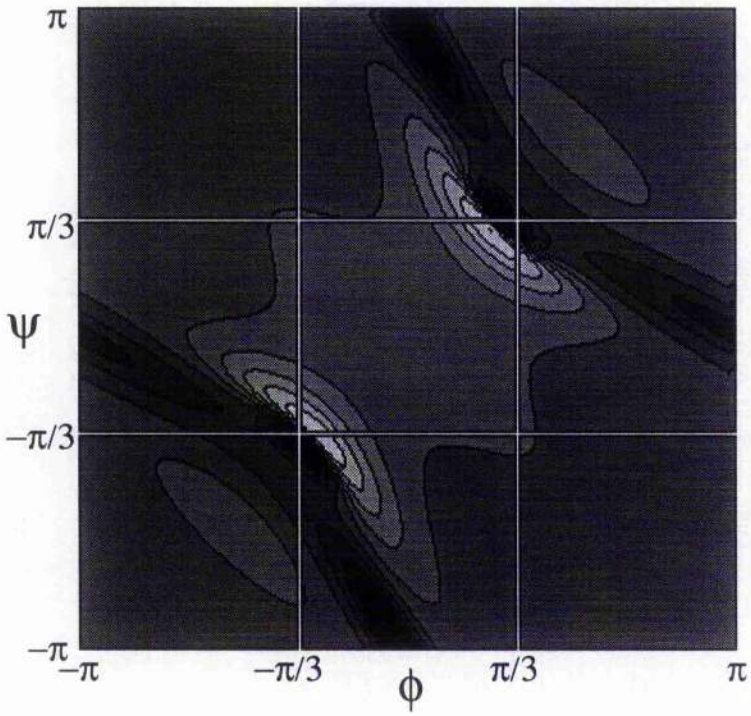

Figure 5.7: Potential Energy of the $i \rightarrow i+4$ Interaction
This shows the same system as figure 5.4, but with two intervening peptides whose interaction is ignored. The minimum corresponds to the strongest hydrogen bond which can form under these conditions, at $\phi = -55^o, \psi = -55^o$. *An unfavourable ridge, caused by* $O_i...N_{i+4}$ *repulsion, has become apparent, passing through the region* $\phi = -60^o, \psi = -60^o$. *The interaction is* $(C_{Oi}, O_{Ci}, N_{Hi+1}, H_{Ni+1}) \rightarrow (N_{Hi+4}, H_{Ni+4}, C_{Oi+3}, O_{Ci+3})$

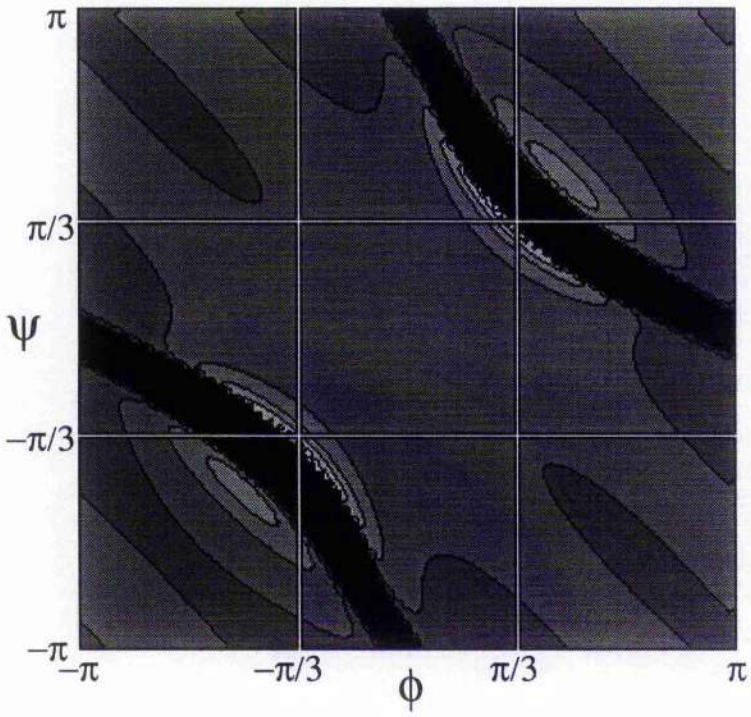

Figure 5.8: Potential Energy of the $i \rightarrow i+5$ Interaction
This is the interaction which should stabilise π helical conformations, $(C_{Oi}, O_{Ci}, N_{Hi+1}, H_{Ni+1}) \rightarrow (N_{Hi+4}, H_{Ni+4}, C_{Oi+3}, O_{Ci+3})$. In fact the best hydrogen bond, around $\phi = -65°, \psi = -65°$ is prevented by $O_i...N_{i+5}$ repulsion. The remaining minimum lies at $\phi = -60°, \psi = -60°$, destabilised by the $O_i...N_{i+4}$ collision of figure 5.7.

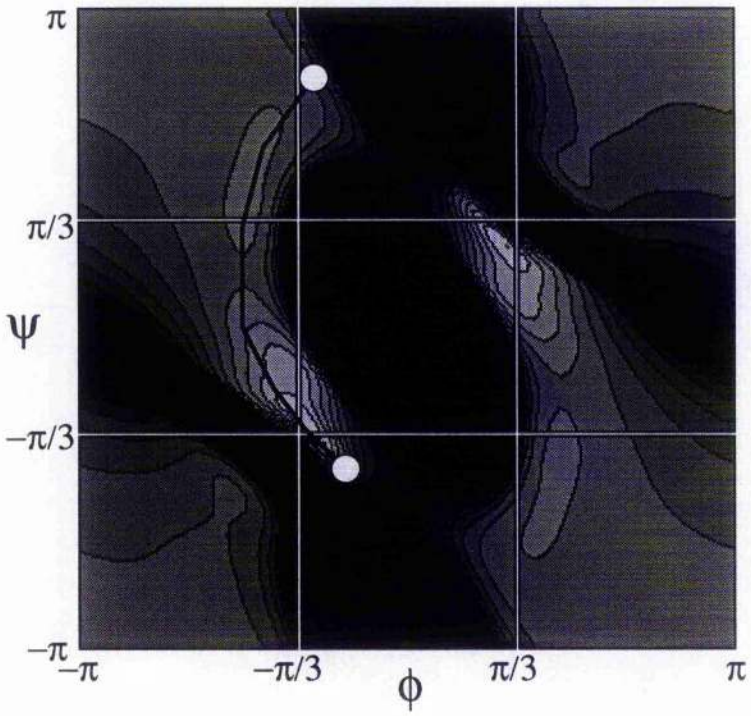

Figure 5.9: Overall Potential Energy for Concerted Helix Formation.
This shows the result of adding the 4 interactions i+2 to i+5 to give the energy per peptide of a homo-conformer. The line shows the folding route of figure 5.2. Notice the barrier to folding at $\phi = -90°, \psi = 30°$, the collisions preventing the formation of the π helix at $\phi = -60°, \psi = -60°$, and the absence of any potential energy barrier between the 3_{10} and α conformations.

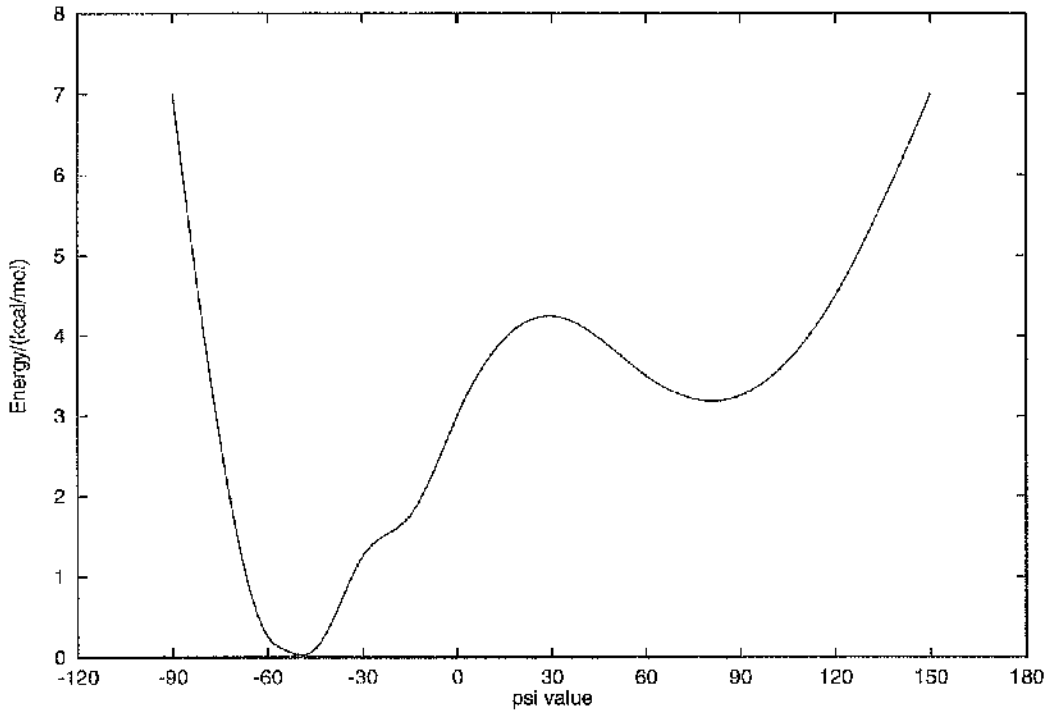

Figure 5.10: The Energy Profile of the Strand-Helix Transition.
This shows a 1D cross section of taken along the line marked in figure 5.9, emphasising the presence of a significant barrier to $\beta \rightarrow 3_{10}$ folding from $\psi = +70^{o} \rightarrow -30^{o}$, the absence of any barrier to $3_{10} \rightarrow \alpha$ folding from $\psi = -30^{o} \rightarrow -55^{o}$, and the impossibility of $\alpha \rightarrow \pi$ folding from $\psi = -60^{o} \rightarrow -80^{o}$.

Chapter 6

The Ends of α-Helices and the Dynamics of Helix Formation

Since concerted, one-step folding of strand to helix is unlikely in proteins, a model of a stepwise process is developed, based on the same conformational energy calculations as in the previous chapters.

In the first stage, a 30 residue polypeptide in twisted beta strand conformation has its centre changed to an eleven residue helix one residue at a time. Looking at the total and per-peptide energy of this system identifies a block to helix formation at the C-terminal end of the helix caused by O_C–O_C collisions. The effect of this is to destabilise the whole C-terminal loop of the helix. No comparable effect is seen for the related H_N–H_N interaction at the N-terminal end.

To provide a more realistic description of the helix ends, a further model is developed in which the end residues of the final $\beta - \alpha - \beta$ structure are allowed to rotate freely. Conformational energy calculations on this system give a clearer picture of the behaviour of the ends of helices while folding. The N-terminal residues can convert from beta to alpha with ease, and the reverse, but the C-terminal residues are constrained to adopt turn-like conformations.

Together, these results explain the observed hydrogen bond patterns (narrow ends, wide ends, and reverse ends) seen in most helices and further suggest that helix growth has a preferred C' to N' direction. Evidence for this preference from sidechain helix breaker/promoter positions, water insertion, host-guest experiments and protein engineer-

ing studies already exists.

6.1 Introduction

It has been shown that strand to helix transitions are unlikely to happen in a concerted manner, and that many of the properties of the polypeptide backbone can be described using the neutral-peptide three term potential of Lifson et al [23]. It remains to be seen what sort of cooperative processes can explain the ease of helix formation. If this is not happening through some repetitive structural intermediate, it must be happening through a set of random coil conformations, with neighbouring residues in different conformations at various stages throughout the process. This leaves a combinatorial problem, with a $2N$-dimensional space of possible conformations where N is the number of residues which form the helix or strand. One approach which can be taken is to release the constraints of the previous studies very carefully. Selecting the constraints to relax requires some insight into the system under study, and this is where structures from the Brookhaven database are invaluable.

Looking at helices from a wide range of proteins, it is possible to identify frozen elements of the helix forming pathway, snapshots of the process which for some reason have been trapped in intermediate conformations. Such distortions seem to be rare within helices themselves, but it has long been recognised that the *ends* of helices nearly always exhibit curious conformations (see for example, the 1981 review of Richardson [3] and the study of Baker & Hubbard [26]), and it is in these structures that suggestions about the helix formation pathway can be found.

6.2 Observations: Helix–strand transitions

Four main classes of distortion are defined at helix ends, with the definitions based on atypical hydrogen bonding patterns. Work by Rose et al [38] has separated these distortions, and ends of helices which do not have distorted hydrogen bonding patterns, into more precise clusters, but as a starting point for the models constructed here the broad definitions of the ends are all that is required.

Therefore, a set of helices from 10 proteins, all refined to a resolution of 2.0 Angstrom

or better, were taken and the peptide to peptide interactions as studied in chapter 5 were examined, using the same potential energy calculations as for the model structures. This gave a semi-quantitative measure of the significant interactions in these proteins, and allowed a more flexible interpretation of when a hydrogen bond could and could not be said to exist. This is a particularly useful approach in the cases where hydrogen bonds were *bifurcated*, and a single amide hydrogen or oxygen could be said to be participating in a hydrogen-bond-like interaction with two hydrogen bonding partners at any one time.

The use of these energy calculations revealed many more of these bifurcated bonds, which are often missed by geometric searches. The results for the helix distortions are described in the following sections: this was not an exhaustive survey, just an attempt to gather structural data in a similar format to the models which were to be developed. Important features of this data were:

- The commonest pattern, an $i+3, i+4$ bifurcation at the N-terminal end of a helix, is the least strongly distorted. The $i+3$ part of the bond is usually very weak, which means that the distortion is not significantly different from a normal helix.

- The distortions at the C-terminal end fall into three classes which together account for up to 80% of all helix C-termini. Most of the remainder are terminated by proline (which cannot make an N–H...O=C hydrogen bond) or aspartic acid (whose side chain carboxyl group disrupts the hydrogen bonding). It is these C-terminal structures which are concentrated on in the next sections.

The structures were identified by their hydrogen bond pattern, then further investigated by other techniques. The energy of each peptide to peptide interaction was calculated, and the individual contributions to overall potential energy plotted as matrices to give a breakdown of the i+3, i+4, and i+5 contributions. The phi/psi values were also calculated, and these were analysed to find clusters where a single residue position in a pattern always had phi/psi values within a standard deviation of 20° of the centre of the cluster. These clusters are plotted, and show how the end of the helix deviates from the classic α-helix conformation.

6.2.1 Narrow C-terminal ends of helices

The first class of end is characterised by a hydrogen bond pattern as shown in figure 6.2, with a bifurcated $i+3/i+4$ bond followed by a single $i+3$ bond. This pattern narrows the loop of the helix.

The existence of the $i+3$ bonds suggests that this type of end is closely related to the 3/10 helix (as is the common N-terminal distortion). This suggests that this pattern is an example of a "frozen" folding intermediate based on the results of chapter 5; the 3_{10} helix is on the pathway from extended strand to α helix, and short 3_{10} helices are stable in their own right. The structure may represent the beginning or the end of a helix folding event.

The phi/psi values shown in figure 6.2 show that this structure is indeed a 3_{10} helix, with the conformation getting closer to the neck region (around ψ=0°) towards the end.

Figure 6.3 shows the interaction between each pair of peptides in residues 75–95 of dihydrofolate reductase, which includes the narrow helix end in figure 6.1. The pattern of energies shows how the stabilisation mid-helix is given almost entirely by i+4 hydrogen bonds whereas at the C-terminus the stabilisation is partitioned between i+4 and i+3 along the last turn of the helix. A similar but less pronounced effect is seen at the N-terminal end.

Loops of this class account for around 30% of helix C-termini.

6.2.2 Wide C-terminal ends of helices

This class of end is characterised by the hydrogen bond pattern shown in figure 6.4. It exhibits a bifurcated $i+4/i+5$ hydrogen bond followed by a single $i+5$ bond. This pattern has the effect of widening the last loop of the helix.

The hydrogen bonds alone suggest that this should be identified as a pi helix by analogy with the narrow end described before. However, it has previously been shown that such a structure is virtually impossible, and in fact analysis of the phi/psi angles which define these ends in figure 6.5 shows that it is more closely related in conformation to the 3/10 helix, with conformations again on the direct folding pathway from strand to helix.

Figure 6.6 shows the energy interaction matrix between each pair of peptides in residues 93–113 of hemerythrin. Again, notice that while geometric criteria only identify two i+5 hydrogen bonds at the end of this helix, the real picture is more complex with three

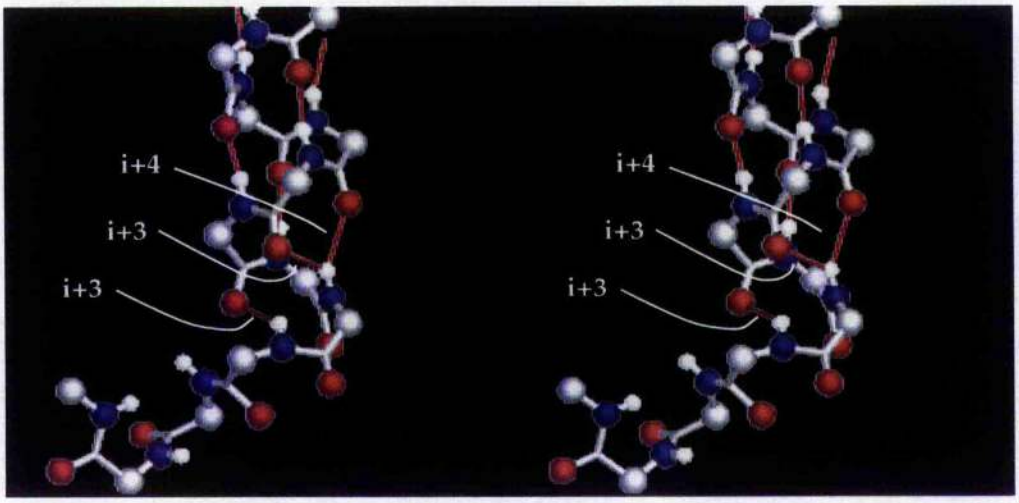

Figure 6.1: A Narrow Helix C-terminal End
Residues 82 to 91 of dihydrofolate reductase (3DFR) show a bifurcated $i \rightarrow i+3,4$ hydrogen bond then a single $i \rightarrow i+3$ hydrogen bond at the C-terminal end of helix E.

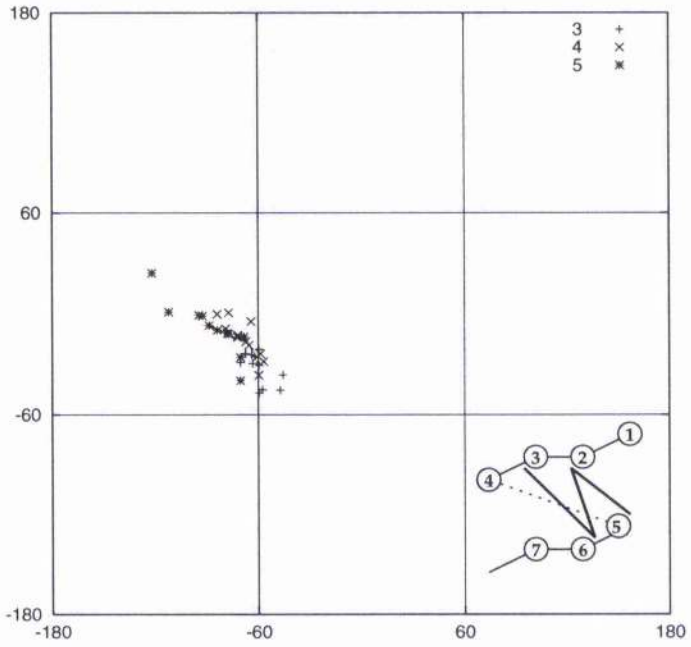

Figure 6.2: ϕ, ψ Values for Narrow Ends
This figure shows the values of ϕ and ψ for the residues relative to their position in the hydrogen bonding pattern, as labelled in the inset, of narrow ends found in 10 high resolution crystal structures. The values show a simple transition from $\alpha \rightarrow 3_{10} \rightarrow 3_{10}/neck$ at the helix ends.

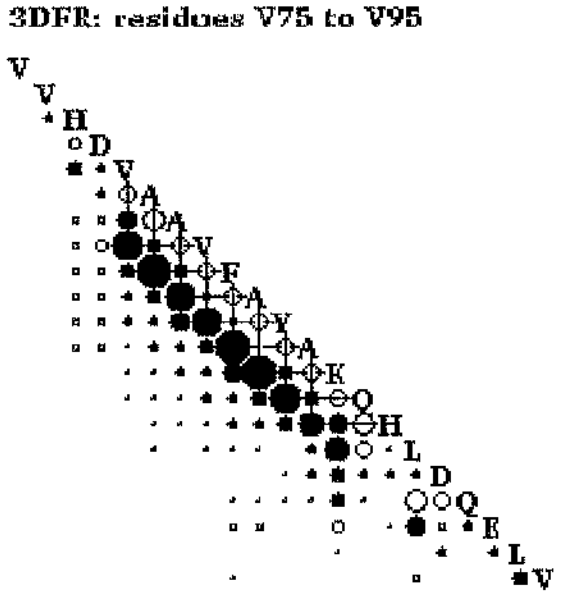

Figure 6.3: Peptide-peptide Interaction Energies at a Narrow End

This figure shows the interaction between each pair of peptides in residues 75–95 of dihydrofolate reductase, including the narrow helix end in figure 6.1. The radius of the circle is proportional to the potential energy of the interaction, black for negative (favourable), white for positive (unfavourable). Absolute values of the energy should not be taken as significant - the strongest here are around -5kcal/mol for a helical hydrogen bond, so the rest should be seen as relative to those values. Although on geometric criteria the helix end is an $i \rightarrow i+3, 4$ $i \rightarrow i+3$ pattern, looking at the energies gives a less clear distinction, with the i+4 interaction still strong at the end.

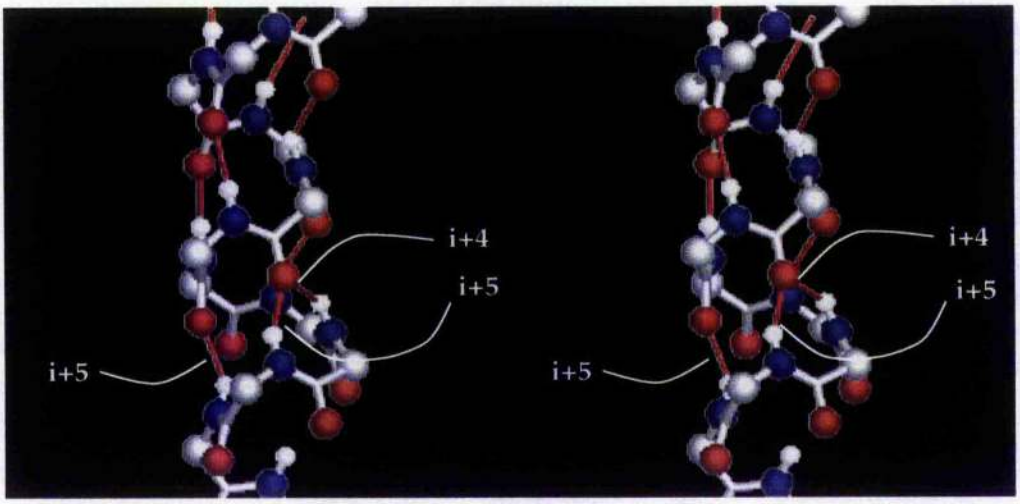

Figure 6.4: A Wide Helix C-terminal End
Residues 98 to 107 of hemerythrin show a bifurcated $i \rightarrow i+4,5$ hydrogen bond followed by a single $i \rightarrow i+5$ hydrogen bond at the C-terminal end of helix H8.

peptide pairs exhibiting strong i+5 interactions.

Loops like this account for a further 30% of helix C-termini.

6.2.3 Reverse C-terminal ends of helices

The third class of C-terminal pattern is characterised by a hydrogen bond pattern as shown in figure 6.7, with a bifurcated $i+4/i+5$ bond followed by a single $i+3$ bond. This pattern reverses the last loop of the helix. Figure 6.9 shows the pattern of peptide/peptide interaction energies associated with it, which in this case are a simple match for the hydrogen bond definition of the pattern.

This structure is unusual, since one of the residues adopts a conformation which is nearly the mirror image of the alpha helical conformation, the so callled α_L conformation. In fact, looking at the phi/psi values actually adopted (in figure 6.8) suggests that the conformation is closer to the mirror of the 3/10 conformation. Side chain atoms interfere with this conformation, so the residue in question must be either glycine or asparagine - in the example here it is glycine.

This hydrogen bonding pattern is referred to as a "paperclip loop" [11] and often appears on its own in the absence of other helical residues. This suggests that such loops

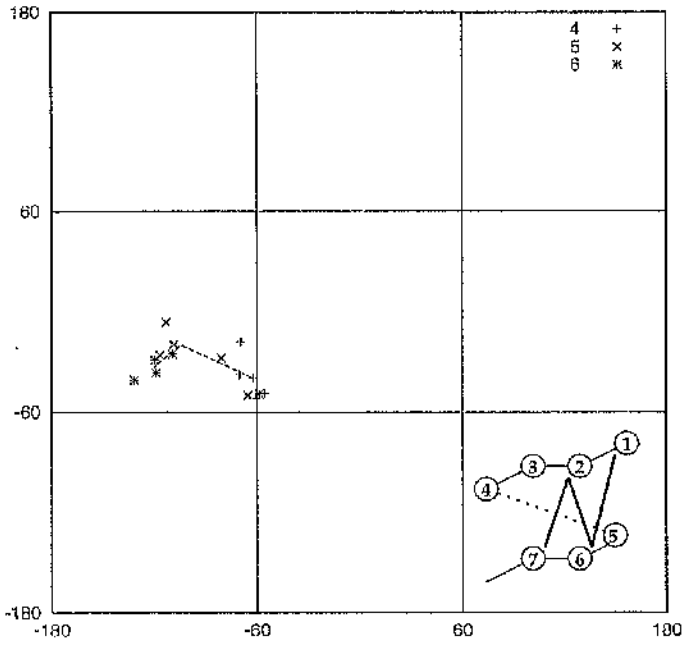

Figure 6.5: ϕ, ψ Values for Wide Ends

This figure shows the wide ends, as defined by the hydrogen bonding pattern defined in the inset, from the same sample of proteins as figure 6.4. The result is unexpected: the i+5 bond characterises the π helix, and might be expected to be formed by residues in the $\phi \approx -60°, \psi \approx -60°$ region of the Ramachandran plot. In fact the values fall around $\phi = -110°, \psi = -25°$ – the π region is only required to form an extended helix of these bonds. The dotted line connects the centres of the three well defined ϕ, ψ clusters.

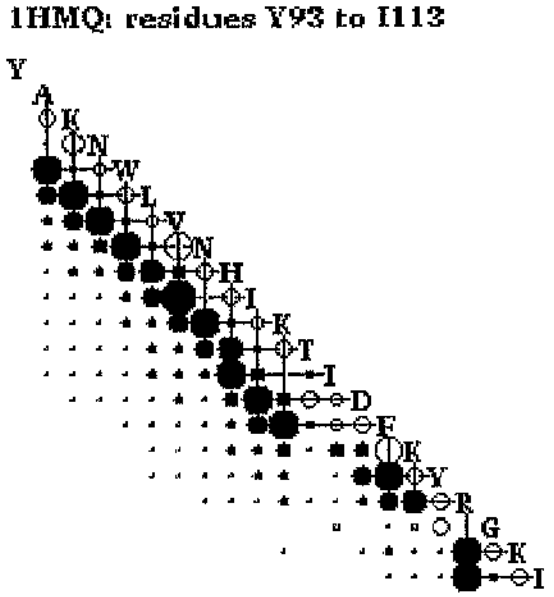

Figure 6.6: Peptide-peptide Interaction Energies at a Wide End
This shows the energy of interaction between each pair of peptides in residues 93–113 of hemerythrin using the technique described in figure 6.7.

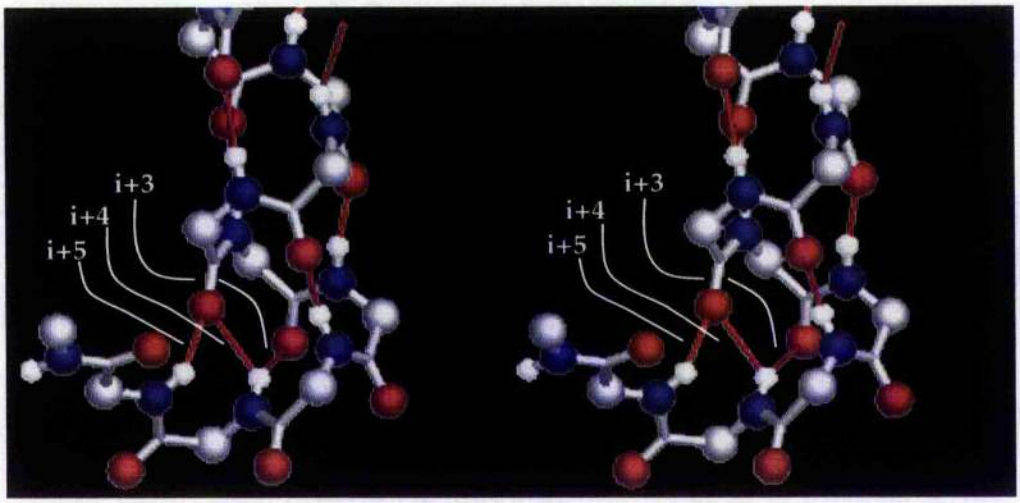

Figure 6.7: A Reverse Helix C-terminal End
Residues 74-82 of actinidin show a bifurcated $i \rightarrow i+3,4$ hydrogen bond and a single i+5 hydrogen bond at the C-terminal end of helix A3, a so-called "paperclip" loop.

could act as nucleation points for helices, while being stable in their own right if conditions for helix propagation are not favourable.

This class of loop accounts for 20% of observed helix C-termini from the sample.

6.2.4 Observed ϕ/ψ distributions

Looking at the phi/psi values for the C-terminal residues in figures 6.2, 6.5 and 6.8, it is possible to see that some of the residues show distinct clusters of values, while others are more widely distributed, mostly in extended conformations. The figure shows the residues which could be assigned to clusters with standard deviations of less than 20°.

The most interesting feature of these plots is that the residues adopt a set of values which are in the classic 3/10 region, even in the wide ends when the bonds in question are not $i \rightarrow i+3$ bonds. Some sort of cooperative effect must be keeping the residues in a conformation which, as shown in chapters 3 and 5, has no particular stabilisation associated with it. The paperclip loop has one residue in the α_L conformation, which constrains the residues which can be involved to be either glycine or asparagine. This is the only type of end (excluding the "dead ends" which are capped by proline, glutamate, or aspartate) which shows such a strong sequence preference.

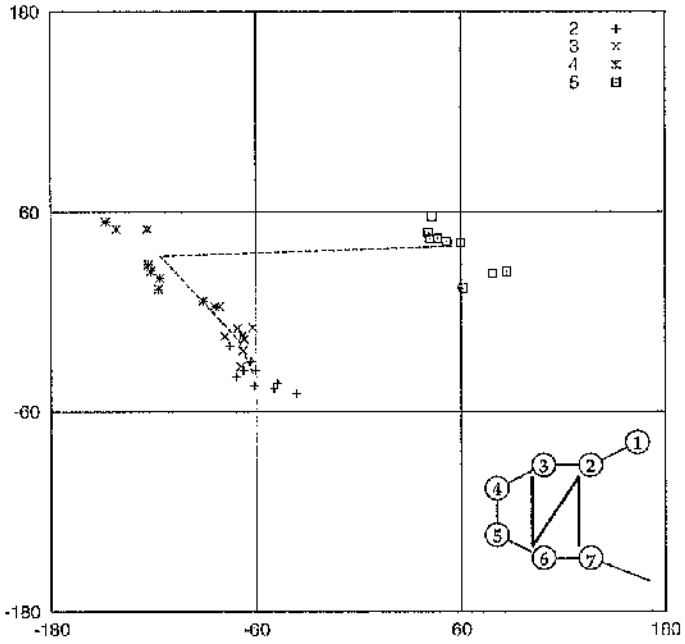

Figure 6.8: ϕ, ψ Values for Reverse Ends

Conformations of paper-clip loops at C-termini. In this case, there is a requirement for residue 5 to be glycine or asparagine, since in order for both i+3 and i+5 bonds to be formed this residue must adopt the α_L conformation.

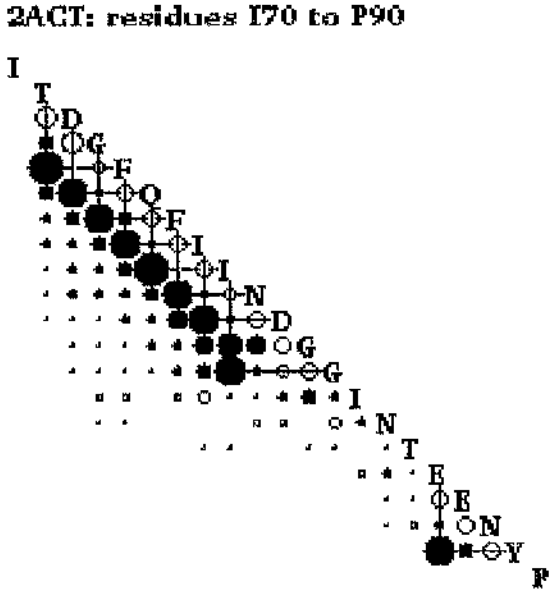

Figure 6.9: Peptide-peptide Interaction at a Reverse End
This shows the peptide interaction energies of residues 70-90 of actinidin, clearly showing the $i \rightarrow i+3,4$ *and* $i \rightarrow i+5$ *hydrogen bonds.*

An explanation for why these particular residues show such a strong tendency to adopt this otherwise unfavoured conformation is clearly needed, and the next parts of this chapter develop models of the ends of helices which give a very clear explanation of these patterns, provide a solution to the high energy barrier to concerted folding, and suggest a partial solution to one part of the protein folding pathway.

6.3 Methods: Analysing regions of the phase space of the polypeptide backbone

As explained before, the phase space for the polypeptide backbone even with bond lengths and angles fixed has a number of dimensions equal to twice the number of residues. Even for a ten reside stretch of backbone, this represents a twenty dimensional phase space, which is not easy to interpret through any display technique. Even if it were, calculating the potential energy surface at a resolution of 1.0° would require 360^{20} calculations.

Some way of restricting the phase space while still allowing enough freedom to find significant stable (and destabilising) conformations is required, and the observations reported in the previous sections suggest a way this can be examined. It is possible to carry out a conformational analysis of the peptide-peptide interactions in any stretch of polypeptide whatever its conformation. The observation that interesting patterns of interactions appear to be concentrated in the boundaries between regions of extended conformation and alpha helical conformations identify a partial restriction which can be applied.

Since the system we are interested in is the general case $\beta - \alpha - \beta$, it is possible to look at this system in isolation, generating the coordinates for a number of residues in a beta conformation, then investigating the effect of folding individual residues into a helical conformation. Since there is an interaction leading to distortions at both ends of helices in real proteins, and these seem to be dependent on distinct patterns of $i+3$, $i+4$, and $i+5$ interactions, it is likely that these effects are dependant on the number of residues which are in the helical conformation - the $i+4$ interaction requires three intervening residues to adopt a helical conformation, the $i+5$ requires four, so only when enough residues are in the alpha conformation can the interactions start to have a stabilising effect.

The calculations on the models constructed are based on the Lifson et al force field parameters, assuming net neutral peptides which can be treated as isolated units. Each

Residue	1-9	10	11	12	13	14	15	16	17	18	19	20	21	22-30
Peptide 1	β	β	β	β	β	β	β	β	β	β	β	β	β	β
Peptide 2	β	β	β	β	β	β	β	β	β	β	β	α	β	β
Peptide 3	β	β	β	β	β	β	β	β	β	β	α	α	β	β
Peptide 4	β	β	β	β	β	β	β	β	β	α	α	α	β	β
Peptide 5	β	β	β	β	β	β	β	β	α	α	α	α	β	β
Peptide 6	β	β	β	β	β	β	β	α	α	α	α	α	β	β
Peptide 7	β	β	β	β	β	β	α	α	α	α	α	α	β	β
Peptide 8	β	β	β	β	β	α	α	α	α	α	α	α	β	β
Peptide 9	β	β	β	β	α	α	α	α	α	α	α	α	β	β
Peptide 10	β	β	β	α	α	α	α	α	α	α	α	α	β	β
Peptide 11	β	β	α	α	α	α	α	α	α	α	α	α	β	β
Peptide 12	β	α	α	α	α	α	α	α	α	α	α	α	β	β

Table 6.1: Conformation of test peptides

peptide is labelled as belonging to the residue which provides the N-H group, so a peptide/peptide interaction's sequence difference does not necessarily map directly to the i,i+n nomenclature for hydrogen bonds. For the per-peptide energy, the contribution of all the other peptides to the potential energy of the peptide in question is simply summed: this is equivalent to adding together all of the values in one row or column of the matrix in figures 6.3, 6.6 or 6.9 and assigning them to the residue which labels that row or column. In this way a more detailed picture than the overall chain energy can be obtained.

6.4 Results: Measuring the energy associated with helix growth

Table 6.1 shows the conformations of the model helices used, and Figures 6.10, 6.11, and 6.12 show the results of carrying out the energy calculations for each conformation. The results show how the stabilisation per residue changes as a 30-residue beta strand has one residue at a time transformed into an alpha helical conformation, ending with an eleven-residue helix embedded in the strand. The results explain a number of features of helix formation.

The transformation proceeds one residue at a time, starting from residue twenty and working backwards through the chain until a ten-residue helix has been formed. Although this effectively means that the growth direction has been fixed, this is not meant to be implied by the simulation strategy: the overall energy of the beta strand conformation is constant at around -1.8 kcalmol^{-1}, and moving the position of the helix backwards or

forwards by one or more residues would have no effect on the overall potential energy of the system.

The important features of these results are best seen by looking at the residue-by-residue energy of the chain with a ten-residue helix, in figure 6.12. This has four different environments, and each has something significant to say about the helix. First, there are the "pure" strand (residues 3-8) and helical (residues 24-28) regions, where the potential energy per peptide is roughly constant. This shows that the strand conformation on its own, although stabilised by gamma-turn type interactions and the other forces described in chapter 3, has only a fraction of the stabilisation of the alpha helical residues. This is only to be expected, though, in the absence of sheet interactions which would be present to provide extra stability for the strand residues in a real protein. The significant features are the two ends, the interfaces between strand and helix.

The N-terminal end shows a distinct step, a region of low stabilisation, for the three residues at the end of the helix. These are residues which can only make one hydrogen bond, so their stabilisation is roughly half that of the other helical residues. This is not altogether unexpected, and illustrates quite clearly the driving force which distorts these ends and gives them their tendency to form bifurcated hydrogen bonds or to bind to side chains: the loose N–H groups represent "sticky ends" to the helix, and moreover ends which could have their hydrogen bonds satisfied by simply folding up a few more residues into the helical conformation: this would always provide greater stabilisation unless the residues in question already had other hydrogen bonding partners which had to be displaced.

The C-terminal end might be expected to be just the converse of that, with half the binding energy because of the free C=O hydrogen bonding groups. In fact the situation is entirely different. The plot shows a pair of peaks, one for residue 19 with a stabilisation of only -0.3 kcalmol^{-1}, which is actually a *destabilisation* relative to the beta strand "ground state" of 1.4 kcalmol^{-1} , and one for the peptide associated with residue 22 with a net destabilisation of 0.9 kcalmol^{-1}, a full 2.7 kcalmol^{-1} above the ground state beta conformation.

This is clearly an indication that some factor other than hydrogen bonding is affecting the C-terminal ends of helices, and suggests that the "sticky end" picture of easy incremental growth at the N-terminal end is not applicable at the C-terminal. There is a destabilising interaction between the peptides of residues 19 and 22: that of 19 is stabilised

in one direction by a hydrogen bond so has less net destabilisation.

The situation is even starker when fewer residues are in helical conformations. For short helices, the weakening of the N-terminal end of the helix and the unfavourable interaction at the C-terminal end combine to give little or no net stabilisation to short helices. One or two residues on their own adopting the helical conformation produce a region of high potential energy, a very strong destabilisation of the strand, even though analyses based on the Ramachandran plot energies of the alanyl and glycyl dipeptides suggest that the alpha conformation is intrinsically stable, a minimum on the peptide potential energy surface. Even helices of four residues are no more stable than the isolated strand (figure 6.13 shows the total energy of the peptides). Some concerted effect of including other elements of the chain must destabilise short alpha helices. This cannot be a side chain effect, because the polypeptide considered here is polyglycine.

6.4.1 This identifies an unstable conformation

The question then becomes, what is the destabilising interaction? The restricted Lennard-Jones method was developed to find stabilising electrostatic interactions, but has been able to identify an unfavourable interaction too. More detail is required, though, and the results are amenable to breakdown (similar to the study in chapter 4), because the energy calculations use a peptide as a basic unit rather than attempting to get a quantitatively accurate energy for the whole chain.

In fact for this case a reproduction of the interaction matrix is not required. The main contribution is a collision between the peptides associated with residues 19 and 22, and the collision in question can be clearly seen by visualising the system without breaking that interaction down into atomic components. The amide oxygen of the peptide associated with 19 has generated a clash with the amide oxygen associated with residue 22 (in fact, these are the carbonyl groups of residues 18 and 21 respectively). A clean strand to helix transition is simply not possible at the C-terminal end of an alpha helix, since it would generate a clash between two large atoms both with significant negative partial charges. The clash is shown in figure 6.14.

An explanation of why the C-termini of alpha helices distort has therefore been provided. The question which remains is why the particular distortions seen are stable, and what implications this has for he protein folding pathway.

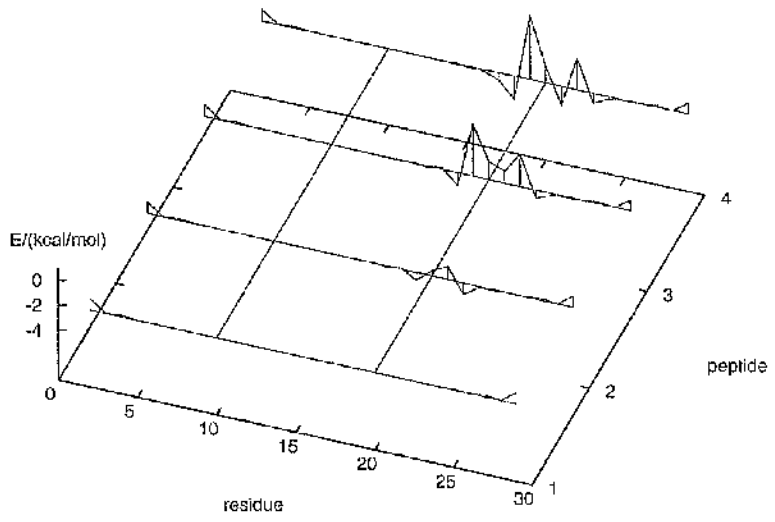

Figure 6.10: Per-peptide Stabilisation of Short Helices

This figure shows the sum of all peptide to peptide interaction energies for the first four peptides in table 6.1. The baseline -1.8 kcal/mol is drawn in for the all-beta conformation, peptide 1, and all changes in potential energy are marked relative to this. Peptides 2,3, and 4 represent 1, 2, and 3 residues adopting the alpha conformation respectively. The first two short helices are unstable relative to the strand, and even in peptide 4 the conformational energy is dominated by the unfavourable interaction between the peptides of residue 19 and residue 22.

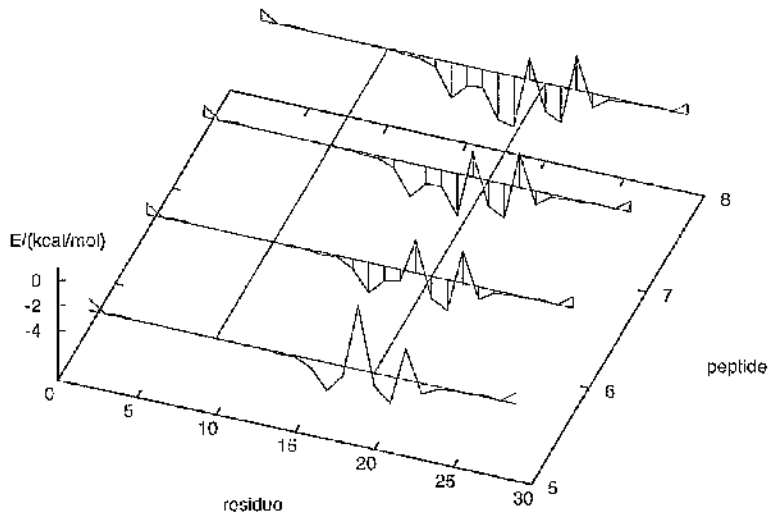

Figure 6.11: Per-peptide stabilisation of Mid-length Helices

Peptides 5 to 8 of table 6.1 have 4,5,6, and 7 hydrogen bonds respectively, and it would be expected that these should show considerable stabilisation. In fact, peptides 5 and 6 are still dominated by the destabilisation which showed its peak in peptide 4, and this does not disappear in the longer helices, even when the peptide of residue 19 is stabilised by hydrogen bonding to the peptide of residue 16 (the i+4 hydrogen bond, three residues apart in this model because the C=O of residue 15 has been assigned to residue 16 for this calculation.)

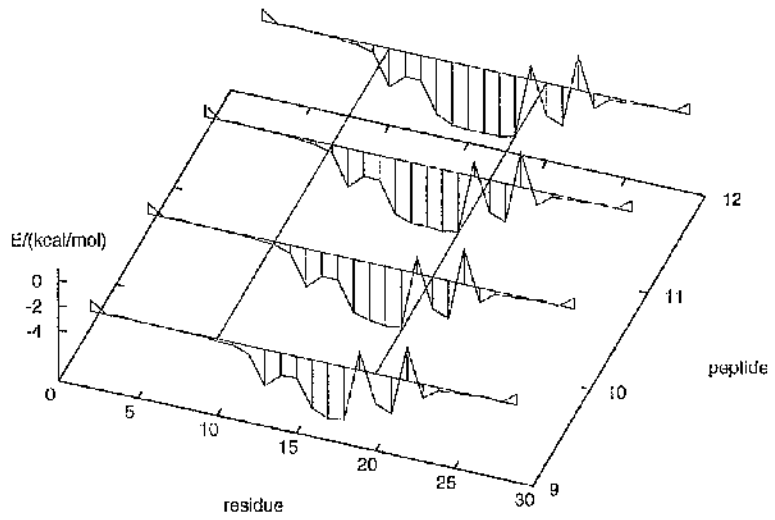

Figure 6.12: Per-peptide stabilisation of Longer Helices

Peptides 9-12 of table 6.1 are 8 to 11 residues long, and should be dominated by strong hydrogen bonds. However, two features are clear – first, that the N-terminal is less stable than the body of the helix, as expected since each peptide only participates in one out of the possible two hydrogen bonds, and second that the unfavourable interaction at the C-terminal end has been "frozen in" rather than weakened by chain lengthening.

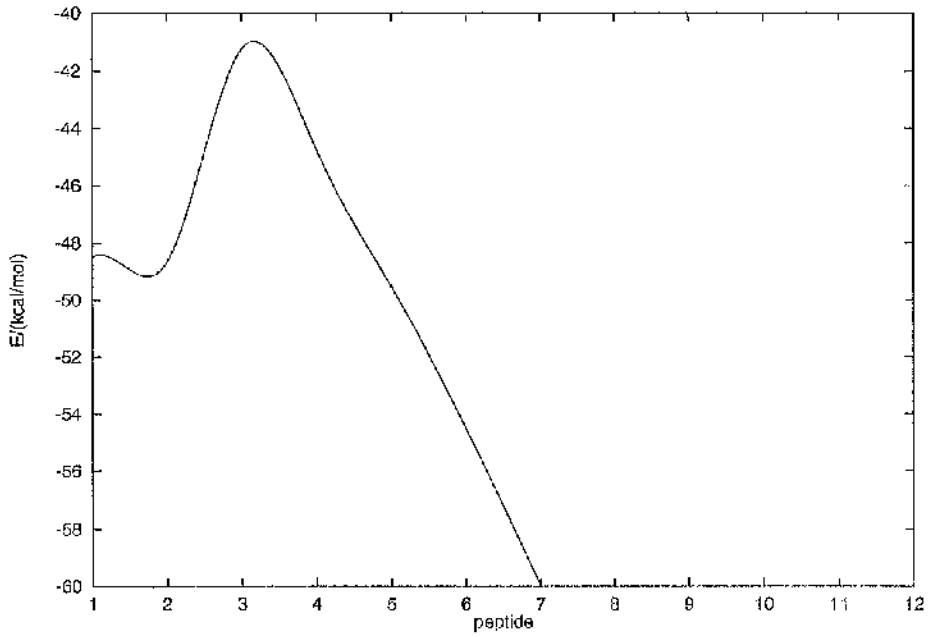

Figure 6.13: Total Stabilisation of Peptides 1 to 12

This figure shows the sum of all the peptide-peptide interactions as a single value for each of the peptides in table 6.1. The feature of note is the instability of 1, 2, 3, and 4 residue α helices: only when 5 or more residues adopt and α_R conformation does the system have significant enthalpic stabilisation relative to the strand conformation.

Residue	1-9	10	11	12	13	14	15	16	17	18	19	20	21	22-30
Peptide 1	β	$\star$	α	α	α	α	α	α	α	α	α	α	β	β
Peptide 2	β	α	α	α	α	α	α	α	α	α	α	$\star$	β	β
Peptide 3	β	α	α	α	α	α	α	α	α	α	$\star$	$\star$	β	β

Table 6.2: Conformation of peptides with free ends

6.4.2 And suggests ways it could be avoided

One strategy for overcoming the steric and electrostatic O–O repulsions has already been identified: simply building one of the three types of C-terminal loop into the model would remove the collision, which is never seen in real protein structures for this reason. A fuller explanation would be preferable, however. Peptide 12 from table 6.1 now acts as the starting point for a new set of constraints – a purely helical region of polypeptide changing to a region of pure strand – the challenge is to develop a system which fulfils these constraints but can still represent the distorted ends. The simplest thing to do is to free the phi/psi values of the interfacial residue and see how the potential energy then changes as this one residue changes conformation.

Table 6.2 shows three systems which model this interface. With peptide 12 from table 6.1 as starting point, one or more residues are relaxed and allowed the full 360° range of phi/psi values. The free residues are marked with stars.

Potential of Free Rotation for the N-terminal Residue.

Figure 6.16 shows the potential energy of peptide 1 from table 6.2, with the N-terminal residue free to rotate and the rest of the peptide conformation frozen. The result is a potential energy surface which shows no barrier along the pathway from beta strand to alpha helix, and indeed shows the stability relative to the (un-solvated) strand of forming an extra helical hydrogen bond. The minimum covers the area of α and 3_{10} residues equally, which is how the bifurcated hydrogen bonds at the N-caps of helices are allowed.

Potential of Free Rotation for the C-terminal Residue.

Figure 6.17 shows the same system but this time for peptide 2 in table 6.2, that is to say with the C-terminal residue free to rotate.

The resulting phi/psi plot shows the preferred conformation of a single interfacial residue. The minimum is very clearly in the "neck" region of the Ramachandran plot,

a region which was originally classed as excluded, then allowed by bond distortion and weakening of the original hard sphere constraint as it became clear that it was a consistently significant feature in the distribution of phi/psi values in crystal protein structures. It is closely related to the 3/10 region, although it has a slightly higher ψ value, and has some stabilisation through possible $i \rightarrow i+3$ hydrogen bonding, but its main stability comes from it simply lying roughly midway between extended and helical conformation. If the interfacial residue were in either of those two conformations, the interface would be "pure" again and O–O the repulsion would have full force. The minimum must lie elsewhere, a 3/10 type conformation could provide weak stabilisation effects which favour this region.

The effect of adopting the 3_{10} conformation is shown in figure 6.15: an i+3 hydrogen bond is formed and the collision is completely relieved.

Potential of Free Rotation for Paired C-terminal Residues.

C-terminal distortions usually involve 2 or more residues. However, setting two residues free gives a 4-dimensional phase space, impossible to visualise. Coupling the ϕ, ψ values of the two residues – in this case, keeping them everywhere equal – allows a 2D representation. This shows how the narrow end gives a clear single minimum for the pair of C-terminal residues in the 3_{10} conformation.

6.5 Conclusions: Helix growth has a preferred C' to N' direction

The observed strand helix transitions with irregular hydrogen bond patterns and the existence of short 3_{10} helices can be explained in terms of the interactions *destabilising* a standard helix embedded in a strand. Some distortions would be expected in any case, to perform a tidying operation on the loose ends of the helix: and it is possible that this is what is seen at the N-terminal end of the helix. The more extreme patterns at the C-terminal end arise from the need to relieve the O_i^C, O_{i+3}^C collision.

Chapter 5 suggested that the 3_{10} helix would always fold to α helix; the results of this chapter reinstate it somewhat, since α helices of 5 or fewer residues are destabilised by the collision identified here. This explains the observed partition between 3_{10} and α helix lengths, with $3_{10} \leq 5$ residues long and $\alpha \geq 5$ residues long in general.

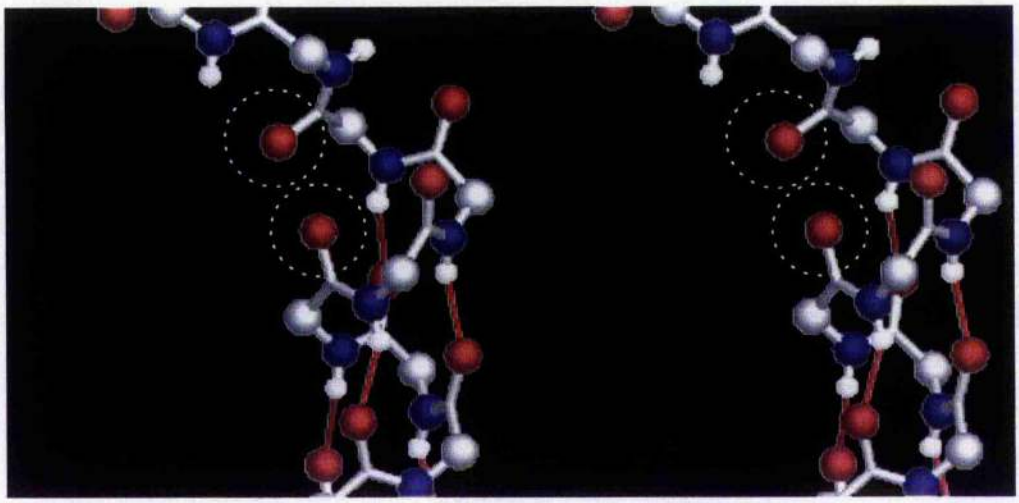

Figure 6.14: The Unfavourable Interaction at the C-terminus of an alpha Helix.
This figure shows part of peptide 12 from table 6.1. The source of the unfavourable interaction between peptides 19 and 22 is immediately obvious as a combined coulombic/steric repulsion between O^C_{18} and O^C_{21}.

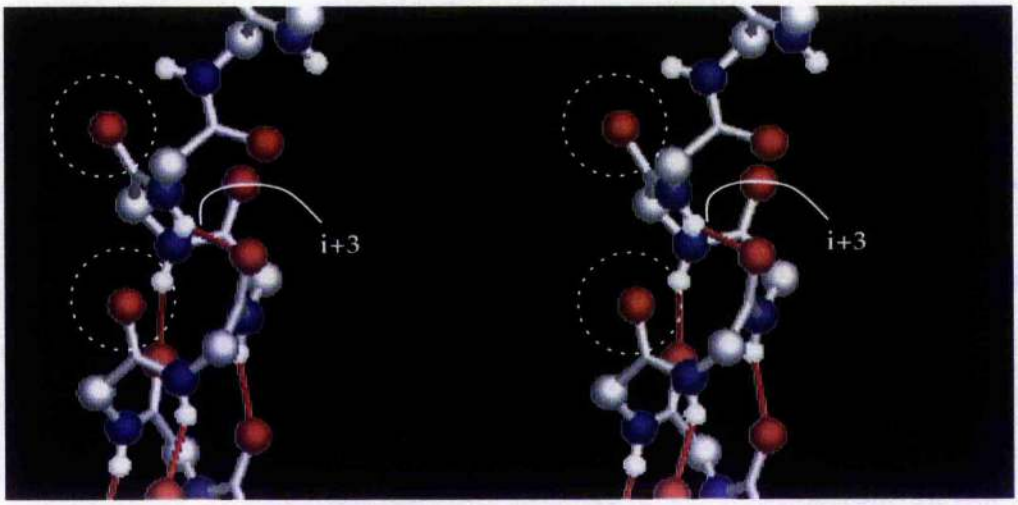

Figure 6.15: Relaxing the unfavourable interaction at the Helix C-terminus.
Figure 6.14 suggests an easy solution to the $O_i \rightarrow O_{i+3}$ clash: simply changing the conformation of residue 20 from α to a 3_{10} or neck conformation relieves the collision and can even provide a palliative i+3 hydrogen bond.

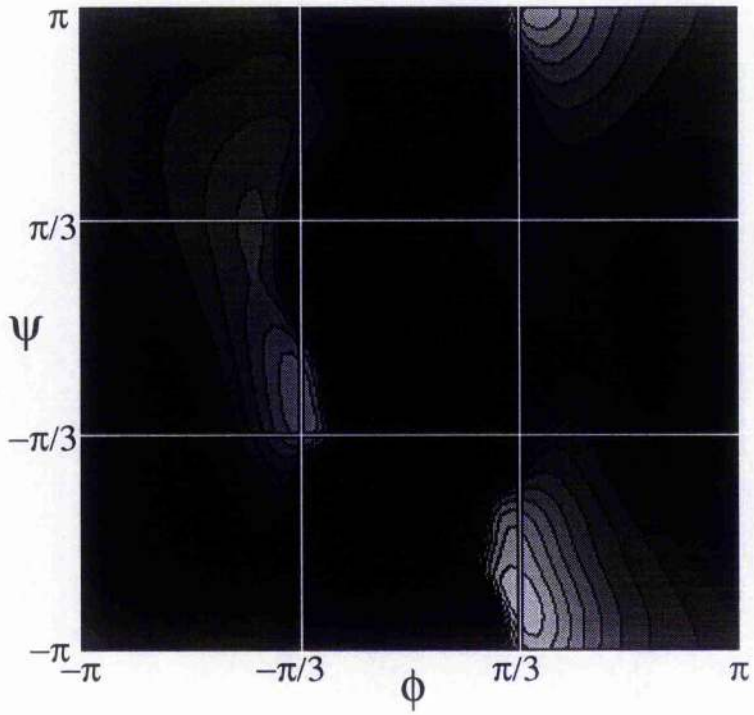

Figure 6.16: Potential of Free Rotation for the N-terminal Residue.
This plot shows the total potential energy as a function of ϕ, ψ for a peptide with the constraints of entry 1 in table 6.2. This is a peptide with backbone fixed in a $\beta - \alpha - \beta$ structure like peptide 12, but with residue 9 free to move. It clearly shows that (for non-glycine residues) there is no significant barrier to folding $\beta \rightarrow \alpha_R$ at the N-terminus of an α helix, and in fact for these residues the helical conformation is strongly favoured over extended conformations.

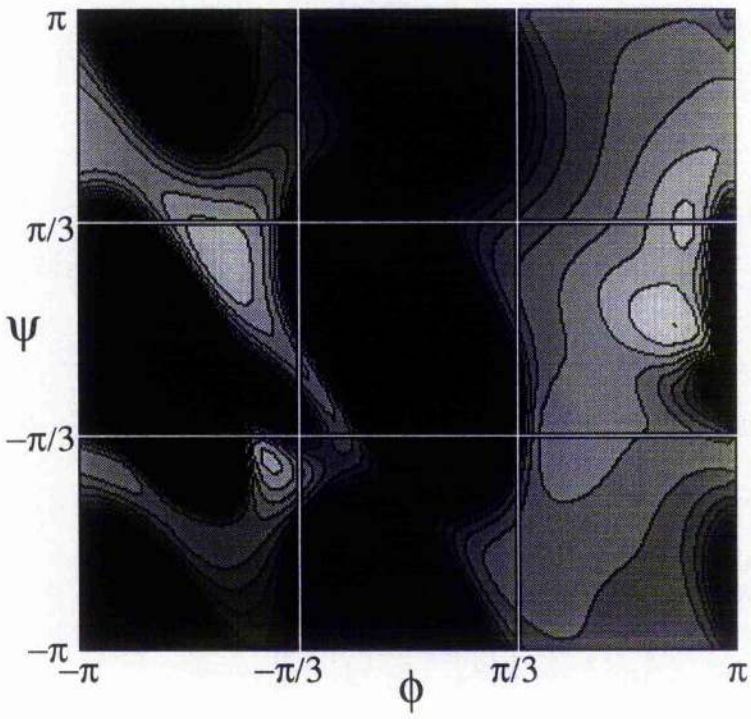

Figure 6.17: Potential of Free Rotation for the C-terminal Residue.
This figure shows the effect of keeping all parameters other than the conformation of residue 20 (the Helix C-terminal residue) fixed. The unfavourable interaction present when the pattern is $19\alpha - 20\alpha - 21\beta$ *is also present when the pattern becomes* $19\alpha - 20\beta - 21\beta$*, moved along one residue. It is expected that a value between* α *and* β*, somewhere in the "neck" region around* $\psi = 0°$ *would be favoured, and this is seen; but strikingly this simple model also predicts stable conformations with i+5 and reverse i+4,5 hydrogen bonds (although the predicted i+5 bonds are in a region of* ϕ, ψ *phase space rarely occupied.*

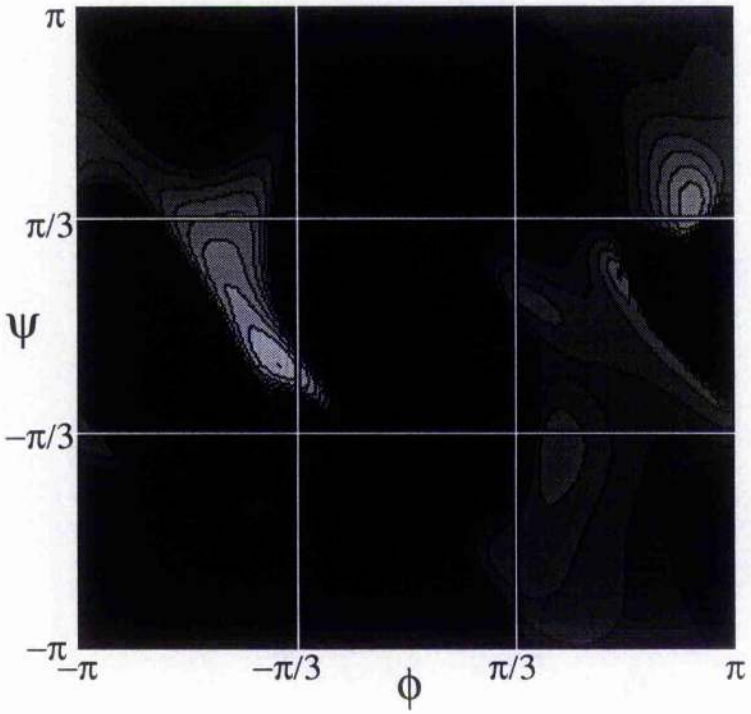

Figure 6.18: Potential of Free Rotation for Paired C-terminal Residues.
This figure shows the same system as in figure 6.17, but this time with two residues coupled but free to rotate at the end of a helix. This provides a single minimum in the 3_{10} region of the Ramachandran plot for the pair of residues.

The most significant conclusion which can be drawn from this chapter is that there is evidence for the possibility of, and also a preferred direction for, stepwise helix growth.

Growth at the N-terminal end of a helix is easy, provided no side chain or tertiary effects interfere with the hydrogen bonding pattern. There is no energy barrier, and the helical form is more stable than extended or random coil alternatives.

The C-terminus, on the other hand, requires cooperative distortion to overcome an energy barrier caused by the O_i^C, O_{i+3}^C collision.

These properties suggest that helix growth may have a preferred C' to N' direction. This is important enough to merit searching for experimental techniques which could verify whether it is true: and there is ample evidence supporting this interpretation in studies by other researchers.

How amino acid insertions are allowed in an α helix

Heinz et al [39] studied the structural effects of inserting residues into an alpha helix of phage T4 lysozyme. They found that insertions are allowed in two ways: in some cases inserted amino acids are accommodated within the helix leading to the translocation of wild-type residues to the preceeding loop, in others a looping out of the residues at the N-terminal end is seen.

Insertions of up to 3 Alanine residues in the middle of helix 39-50 of WT*(cysteine free) T4 lysozyme led to translocation of the N-terminal residues into the N-terminal loop. Only insertions right at the C-terminal led to distortions in the C terminal direction, and this was a result of disruption of a loop (looping out) rather than translocation. No translocations towards the C-terminus were seen in any of the mutants.

Insertions did not increase the length of the helix, instead distortions up to 1 nm away were seen. Helix length is determined by the packed interface with the rest of the protein rather than simply by effects at the helix ends.

Amino acid preferences for specific locations at the ends of alpha helices

Reviews by Richardson & Richardson [37] and Rose et al [38] clearly demonstrate that several types of residue cluster at specific points in relation to the ends of helices. To decide what role these residues play in folding it would be necessary to distinguish helix breakers from promoters, a difficult task from static structures alone.

It is possible to discriminate on the grounds that a capping helix breaker residue acts either on the N-terminal end, or the C-terminal end, and rarely both. This suggests that a helix which is growing in one direction but meets a cap residue for the "wrong" end will not be affected.

This effect would be observable in differing statistics for cap residues: helix breakers are residues which have an increased preference above mean frequency for the caps but below mean for the body of the helix, while promoters would have a near or above average preference for the helix body.

In fact, the pattern shows that the N-cap residues are breakers. Gly is a helix breaker with C-cap preference, but we have already seen that it forms an important loop which can exist on its own and could therefore be a nucleation point. Pro, Asn, Gln, Asp and Glu all have N-cap preference and below mean mid-helix preferences – so they can be classed as helix breakers. Lys and Arg and His are C-cap residues, and have higher than mean mid-helix preferences, and could therefore be classed as helix promoters. The position relative to the cap residue are much more clearly defined for the breaker residues than for the promoters, suggesting that their effect is either more well defined or that the C cap residues play their role early in folding, before final rearrangements to settle the structure are complete.

Water inserted alpha helical segments

Sundaralingam and Sekharudu [40] surveyed water inserted alpha helical segments and reached the conclusion that, as these insertions were seen in typical α-like reverse turn conformations, water could catalyse the folding of alpha helices by forming links between the $C{=}O_i$ and $N\text{-}H_{i+4}$ groups which initially stabilise three centre $C{=}O{\rightarrow}H_2O{\cdots}H\text{-}N$ systems and then be expelled as the i→i+4 bond is formed.

Of the 33 cases where water was inserted into an i→i+4 bond, 26/33 were N-terminal (between N-3 and N+3), 4/33 were within the body of the helix, and 3/33 were C terminal. If these are indeed frozen folding pathway intermediates, it means that the N-terminal of a helix is some 7 times more labile than the C terminal end, and more prone to fraying either during helix folding or unfolding.

(It would be interesting to construct a model of early stages of helix nucleation, even assuming a simple model such as the one developed here, and investigate the effects of

possible external water binding on the transition states and stable minima.)

Evidence from Protein Engineering

Fersht et al [41] have shown that in barnase, mutations TA6 and TA26 (using stopped flow fluorescence spectroscopy and folding kinetics) have a very small effect on intermediate and transition states of folding, but a significant effect on the stability of the folded state. These are helix N-cap residues. In contrast, mutations TS16 and HQ18 destabilise all three postulated states on the folding pathway. These are C cap residues.

Complementary nmr experiments [42] show that in helix T6-H18 the C-cap hydrogen atoms are protected against deuterium exchange early on in folding, and likewise two C cap H atoms in helix T26-G34. In neither helix were atoms between N and N+3 protected.

Together, these two pieces of evidence suggest that the C-cap forms its final structure long before the folding is complete, while the N-cap is either unfolded or remains free to fold and unfold repeatedly until the final structure of the protein is defined.

Part II

Some Factors Stabilising Protein Tertiary Structures, and Novel Techniques for Examining Them

Chapter 7

Tertiary Ring Structures Stabilising Proteins

Hydrogen bond formation is one of the strongest constraints determining the allowed states of folded proteins, so it is to be expected that many of the important stabilising interactions can be identified by identifying the hydrogen bonding patterns associated with them. This is already known for mainchain-mainchain interactions, and for sidechain-mainchain hydrogen bonds in loops and secondary structure elements, but extending such searches to tertiary sidechain-mainchain and mainchain-mainchain interactions has been a laborious manual task in the past.

This chapter presents a method for automated searches of hydrogen bond databases to find specified patterns of hydrogen bonds based on topology, atom types, and sequence separation. The first result of this search technique is the assessment of the significance and rate of occurrence of rings involving mainchain hydrogen bonding to both the sidechain O^C and syn-H^N of asparagine and glutamine residues. These are found to be very common (1-2 per protein) and typically occur in 9- or 11-membered rings which constrain the mainchain residues involved to adopt extended or strand conformations.

It is possible these rings represent a long range sidechain mediated effect of tertiary interactions on local backbone structure. They are also common in binding of peptide substrates, where they serve to constrain backbone rotations and define the conformation of the bound peptide.

7.1 Introduction

Although a detailed understanding of the conformation of the polypeptide is vital, and all of the features described in previous chapters have relevance to the pathway of protein folding, in the final analysis it is protein *sequence*, and hence the sidechains, which determine the final fold of the protein. There has been considerable work on the local effects of side chain to main chain hydrogen bonding, and many surveys of individual long range hydrogen bonds.

Clearly, hydrogen bonds are significant in determining folding patterns, but other factors such as hydrophobic effects may well be important too, and it is hard to assess the relative importance of different stabilising interactions. Clearly they need to be identified and classified, but if they are to be said to be having a significant directing effect on protein folding they must be shown to be conserved interactions. Many apparently strong interactions, such as salt bridges, have a surprisingly low degree of conservation as compared with main chain hydrogen bonds which show how well backbone structure is conserved. Any hydrogen bonds which have a strong effect on tertiary structure will both be conserved and have some distinctive properties which show they are influencing the backbone conformation, rather than just occurring opportunistically, binding externally to structures which already exist.

7.1.1 The stability of hydrogen bonded rings

While individual hydrogen bonds may or may not be exerting a significant structural influence, something which is hard to decide from a static structure, any hydrogen bonding pattern which forms a ring is likely to be significant because its formation requires several atoms to be constrained into a roughly fixed geometric arrangement, often constraining several bond rotations, with a concomitantly high entropic cost. Any ring which is present in a folded protein is likely to be exerting a strong directing effect on the local structure.

7.2 Methods: Analysing hydrogen bond databases

There is a huge amount of hydrogen bonding information implicit in the Brookhaven protein data bank [12], and this has been the subject of a number of exhaustive reviews

(see eg Baker & Hubbard 1984 [26] and Kabsch & Sander 1983 [20]): in the case of the survey of Kabsch and Sander, the results are developed into a database of hydrogen bonding and other structural parameters complementary to the Brookhaven database. None of these works represent an attempt to tabulate all the hydrogen bonds in proteins, although this is in practice easy even with relatively modest computing resources.

The work of Poet, Milner-White, and Belhadj-Mostafeda [10] focused on identifying hydrogen bonds and maintaining a list of them for each protein as an aid in visualising protein structure. This code already had all the data structures required to identify residues, generate hydrogen bonding hydrogen atom positions, and identify hydrogen bonds using the geometric criteria of Baker & Hubbard. This formed a program which was extended for this study to generate an exhaustive database of hydrogen bonds in 68 proteins, which then formed the basis for the analyses in this and subsequent chapters.

7.2.1 Lists of hydrogen bonds

Any database has to be constructed with a little care, in particular taking into account the way in which it will be used to find patterns. The DSSP files have a fixed number (2, in fact) of fields to represent main chain hydrogen bonds, and for example could never identify a three centre hydrogen bond in a protein - not a common structure, but nevertheless one which occurs from time to time. Each hydrogen bond needs to be labelled with its type (sidechain-sidechain, sidechain-mainchain, or mainchain-mainchain), the donor and acceptor residues, and the atoms involved. Ideally the hydrogen atoms involved should be labelled too, to distinguish which one is involved in a hydrogen bond when there is more than one possibility (for example, the case of *syn* and *anti* amide hydrogens examined in this chapter).

It is useful when using geometric criteria to present the hydrogen bond lengths and angles, because when investigating new structures this helps to identify patterns which only consist of borderline structures and can focus the researcher on those which show strong, distinctive hydrogen bonding patterns.

7.2.2 Exhaustive lists with energetics

The other side of this is that a simple geometric description often misses significant interactions, especially when the electrostatics of whole groups rather than just the hydrogen bonding partners is taken into account. Part I has shown several significant interactions which are only explained with a full energy calculation, and it is to be expected that a wide range of interactions as yet unidentified will become apparent when systems are treated more realistically. Chapter 4 shows how important taking into account the partial charges across all the atoms can be.

However, some restriction has to be made on the interactions which are recorded if any interpretation is to be possible. A full list of all the interactions between side chains and peptides would require $2N(2N-1)/2$ recordings. This is certainly possible (indeed it is done routinely during most molecular dynamics simulations), but means that any search for a given structure gives rise to a huge number of possible matches, each of which then has to be accepted or discarded using some criterion, usually the energy of the interaction; but if the lowest energy at which an interaction becomes significant is not clear – especially when other factors such as the entropic contribution of restricted flexibility become important – a simple measure of enthalpy may not be enough.

For this reason, it seems reasonable to base initial searches for structural patterns on the "best guess" for a stabilising structure - which often means a pattern based on hydrogen bonds, even when many other types of electrostatic effect are actually taken into account in the final analysis. Simply loosening the geometric criteria, for example allowing C=O...H and N–H...O angles of up to 180° and extended H...O distances of up to 5 Angstrom, catches nearly all hydrogen-bond-like electrostatic interactions, and gives a set which can then be analysed and then classified - possibly just with slightly wider acceptance criteria, possibly with a purely enthalpy based cutoff, or perhaps with phi/psi constraints on donor, acceptor and intermediate residues.

This is the approach which was used in chapter 6 to identify the patterns at the C-termini of alpha helices; the geometric rules of Baker & Hubbard and the electrostatic calculations of Kabsch & Sander both failed to identify a significant proportion of the hydrogen bonds which defined these structures. Using the 9-6-1 potential, it was possible to select an enthalpic criterion, simply excluding peptide-peptide interactions below

1.5 kcalmol^{-1}. As explained before, these energies are likely to be innacurate, by at least 10%, but they provide a very useful qualitative discriminant. In the case of the C-terminal patterns, it was possible to verify the significance of these structures by looking at the phi/psi angle distribution, and in most cases it must be hoped that some similar secondary verification of the structure can be found.

7.2.3 Searching for hydrogen bonding patterns

Once a wide range of hydrogen bonds has been collected in a database, it is possible to look at hydrogen bond patterns in a range of proteins very quickly. It is crucial that the hydrogen bond data be stored as text files (rather than interactively generated from within a protein visualisation package, for example), so that searches can take advantage of database software or a UNIX environment which has been optimised for fast file processing.

For example, in the work described in the next section and in the accompanying paper of Le Questel *et. al.* [17], a possibly significant structure was identified visually using the program Sybyl. A fast survey of a set of 68 proteins was then required to see if these structures were common, if related structures occurred with any frequency, and if the pattern could be said to be common enough to be more than just a statistical effect. Carrying out the full survey would have taken many weeks using interactive software and searching for each pattern visually. Each protein could be searched in a few seconds using text based hydrogen bond files, and the search of the whole database could be automated allowing more time for the structural analyses and searches for related structures in other biomolecules.

The search itself is easily carried out using software written in C. Each pattern can be defined in terms of a few hydrogen bonds. One hydrogen bond is defined as the **root**, and identified by the residue types acting as donors and acceptors, the residue atoms which can be involved as donor and acceptor, and the criteria used for accepting or rejecting a given choice. If the pattern is fully described by a single hydrogen bond – as, for example in the work of Baker & Hubbard – then any positive match constitutes a result. Otherwise, other hydrogen bonds must be present, and these are defined as **offsets** from the donor and acceptor residues, and again as the atoms and residue types which can participate. Offsets can be defined both by the sequence difference between the root donor and/or acceptor and the offset acceptor and/or donor residues.

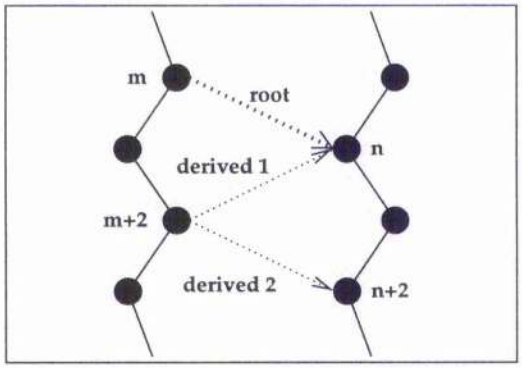

Figure 7.1: Hydrogen bond pattern searches
Searches are defined relative to an initial match, the **root** *hydrogen bond. In addition to matching the type of residue and atom specified for donor and acceptor, hydrogen bonds must also conform to specified sequence differences. The difference* **n-m** *can be specified for the root hydrogen bond, useful in finding helices for example, and the other hydrogen bonds can have their offsets from the root donor amino acid (***m** *here) and acceptor amino acid (***n** *here) specified, depending on which chain segment they belong to.*

Table 7.1 shows the generic search pattern, with the root defined by type (donor or acceptor), residue and donor or acceptor atom. This defines two residues, the source and the target, and offsets can be specified to have a given sequence difference from the source (specified by parameter x) and from the target (parameter y), as well as having specified atom and residue types.

These searches can be used to identify main chain structures, both repetitive secondary structure such as helices and types of sheet, and main chain loops such as paper-clips and beta turns, as well as patterns involving side chains.

root	type	residue	D/A	is donor?	atom 1	atom 2	m-n	
derived1	type	residue	D/A	is donor?	atom 1	atom 2	x	y
derived2	type	residue	D/A	is donor?	atom 1	atom 2	x	y

Table 7.1: Format for a generic hydrogen bond pattern search.

root	main/main	$\star$	N/O	donor	N^H	O^C	m-n=4	
derived1	main/main	$\star$	N/O	donor	N^H	O^C	x=1	y=1
derived2	main/main	$\star$	N/O	donor	N^H	O^C	x=1	y=1

Table 7.2: Search pattern for a section of alpha helix

7.3 Results: Rings involving amide sidechains

In 1992, Jean-Yves Le Questel, working in the Chemistry department at Glasgow University, visually identified a hydrogen bonded structure which appeared to occur several times in each protein. It quickly became clear that two related ring structures, with the amide side chains of asparagine or glutamine doubly hydrogen bonded to distant part s of the main chain, were quite common: questions remained as to how common they were, whether there were there any related structures which also occurred, and if the rings have any significance beyond the simple mopping up of side chain and main chain hydrogen bonding opportunities. A search of the database of 68 proteins whose hydrogen bonds had already been found soon showed that the patterns are indeed a common, and probably significant, feature.

7.3.1 Amide/mainchain rings

Figure 7.1 shows the two types of ring which were found. Both asparagine and glutamine sidechains form rings of these types, where both the amide CO and the *syn* hydrogen of the amide NH_2 bind to the peptides on either side of a single C^{α} atom, either to the NH and CO of a single residue i, giving a nine-membered ring structure with seven covalent bonds and two hydrogen bonds, or those of the two adjacent residues $i-1$ and $i+1$, an 11-membered ring structure with nine covalent bonds and two hydrogen bonds.

the survey found 827 asparagine and glutamine residues in total, of which 89 had both NH and CO hydrogen bonded to the main chain. Of these, 54 involved th e syn hydrogen only, and 33/54 of these were the 9- or 11-rings under investigation.

7.3.2 9 member rings

A search for these was based on the instructions listed in Table 7.3.

21 9-member rings were found, all bound to residues with phi/psi values in the β strand or extended conformations.

The variability of the Asn or Gln residues as defined by Schneider & Sander [43] fell in the range 0-46, with most values in the range 10-20 (indicative of significant conservation during protein evolution.

7.3.3 11 member rings

A search for glutamine and asparagine 11-member rings was based on the instructions listed in Table 7.4.

12 of these rings were found, again binding to regions of β or extended strand – with a single exception, Q232 in penicillopepsin, where the mirror phi/psi conformation is adopted.

The variability value for these rings was higher than for the 9-rings, with a range of 0 to 58 but most of the values clustered around 35. This means that the 11-rings are less well conserved: in this case the conservation is not significantly greater than for this type of side chain in general.

7.3.4 Other possible rings

While the search for 9 and 11 member rings was based on the rational idea that ring structures are stabilising in proteins and should therefore be favoured, the question still remained as to whether these rings were more than just a statistical side effect of the number of hydrogen bonds asparagine and glutamine can make to the main chain. In order to test the opposite hypothesis, that mainchain amide rings were so common that the rings were insignificant, it was necessary to search for a range of structures where both *cis* hydrogen and carbonyl oxygen formed mainchain hydrogen bonds, even in cases where the "rings" formed contained tens of atoms and were impossible to find visually.

The search pattern was extended to cover all cases where the value y (the sequence difference between the two target residues) in table 7.1 was in the range -5 to +5, and also the cases where y>5 and where y< −5. Thanks to the automated search machinery, this only involved changing two lines in a file 10 times, rather than the time consuming effort of searching along the main chains of 68 proteins visually.

The results of the search are given in table 7.5. While there were a number of cases where the larger rings were found, they were not significant relative to the frequency of 9- and 11-ring occurrence. The apparently high number of "rings" where y> 5 and y<5 is an artefact of the algorithm finding any case where both the N-H and C=O were hydrogen bonded, not necessarily to the same structural feature.

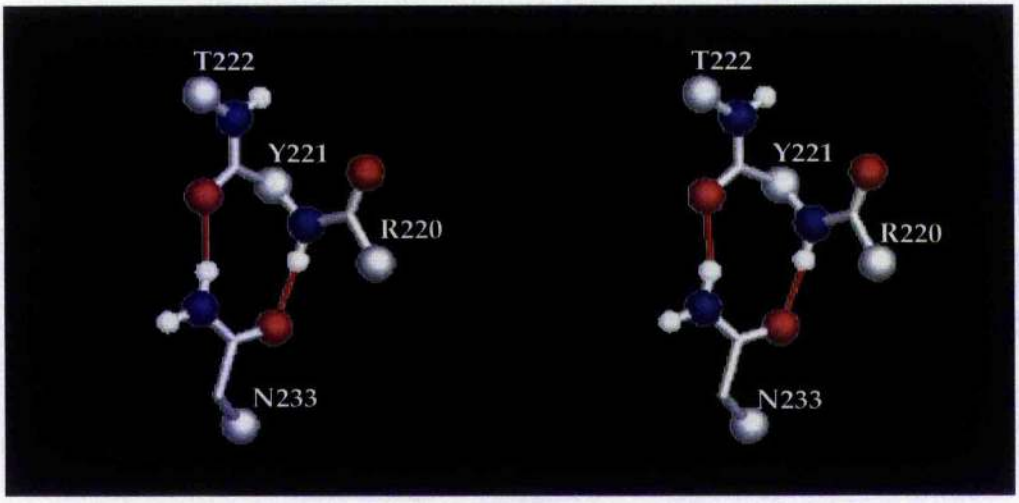

Figure 7.2: A Nine-membered Carboxamide-Mainchain Ring
This figure shows the hydrogen bonded ring formed between N233 and R220–Y221–T222 of thermolysin, with 9 members. The "target" mainchain is in a slightly twisted β conformation, the usual conformation for this type of ring.

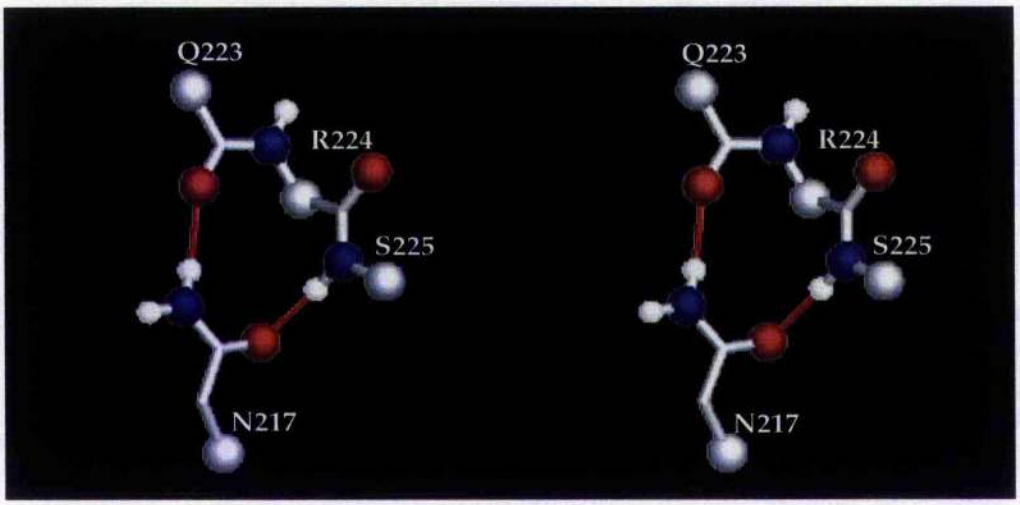

Figure 7.3: An Eleven-Membered Carboxamide-Mainchain Ring
This is the hydrogen bonded 11-ring between N217 and Q233–R224–S225 of α-lytic protease. Again, the "target" mainchain is in a typical β conformation.

7.4 Conclusions: amide sidechains act as mainchain conformational locks

These rings are not unexpected, and they might have been predicted as an effective way to mop up both sidechain and mainchain hydrogen bonding potential, but they have the special feature that they are binding to a distant region of the polypeptide backbone, and that when they do bond they provide a significant constraint on that part of the polypeptide chain. In cases where they can be seen to bind to the fraying edge of a beta sheet the importance of mopping up hydrogen bond potential is obvious.

If the rings are indeed strong tertiary determinants of secondary structure, they should be particularly prevalent in positions where local backbone conformations are important, such as substrate binding sites. The paper of Le Questel et al contains several examples, binding to flexible regions of substrates as if they were patches of backbone and presumably locking them in position.

Recently, more proteins binding peptide transmitters, substrates and recognition factors have had their structures solved. A feature of many of these is that they involve asn or gln residues in their binding sites, and the role of these amide sidechains is to hold the polypeptide backbone.

The most striking example of this is the human MHC class II receptor solved by Wiley et al [44], which shows three of these rings in the binding of a single 20 residue peptide, as shown in figure 9.3. This is an unusually large number of a single class of structural motif to find in a single binding site, and may be related to the unusual conformation in which the influenza peptide is bound. The peptide is actually in an extended polyproline II helix conformation, although the example crystallised was proline-free. Typical polyproline stretches in globular proteins are three to five residues long [33], so this is a unique feature of this particular binding site.

The peptide is bound for presentation to other recognition molecules, and it is clearly important that it is held in an extended conformation to expose as many residues as possible to determine a distinctive 3D/electrostatic profile for each peptide bound. The 9- and 11-rings must play an important role in this. It is even possible that the polyproline conformation used is significant, as folded protein structures will never present the residue pattern of this bound peptide, since they do not have extensive polyproline II helices as a

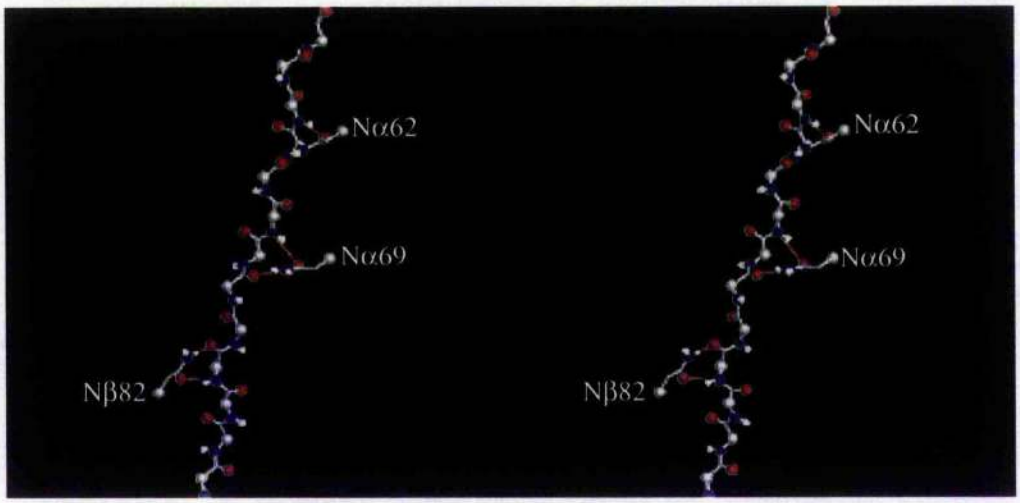

Figure 7.4: 3 Carboxamide Rings as Conformational Locks
Shown here are 3 carboxamide rings in the binding of an influenza virus peptide by human class II MHC protein. Nβ82 and Nα62 form 9-membered rings with the virus peptide mainchain, and Nα69 an 11-member ring. The effect is to hold the virus peptide in a polyproline (overtwisted strand) conformation. The tertiary hydrogen bonding network, dominated by these three rings, is locking the local secondary structure of the peptide.

secondary structure element.

The evidence from protein structures alone is enough to show that 9- and 11-member rings are significant features in protein structure, and their positioning and conservation suggest that they are important structure determinants. The natural extension of this is to polypeptide substrate binding, where the expected role in backbone binding is actually seen. The hypothesis that these are effective conformational locks suggests that many more of these rings will be found in sites analogous to the MHC case.

Asparagine:

root	side/main	Asn	N/O	donor	N^{γ}	O^{C}	$\star$	
derived1	side/main	Asn	N/O	acceptor	O^{γ}	N^{H}	x=0	y=0

Glutamine:

root	side/main	Gln	N/O	donor	N^{δ}	O^{C}	$\star$	
derived1	side/main	Gln	N/O	acceptor	O^{δ}	N^{H}	x=0	y=0

Table 7.3: Format for a asparagine and glutamine 9-membered ring searches.

Asparagine:

root	side/main	Asn	N/O	donor	N^{γ}	O^{C}	$\star$	
derived1	side/main	Asn	N/O	acceptor	O^{γ}	N^{H}	x=0	y=2

Glutamine:

root	side/main	Gln	N/O	donor	N^{δ}	O^{C}	$\star$	
derived1	side/main	Gln	N/O	acceptor	O^{δ}	N^{H}	x=0	y=2

Table 7.4: Format for a asparagine and glutamine 11-membered ring searches.

NH of	i	i	i	i	i	i	i	i	i	i	i	i	i
CO of	i-n	i-5	i-4	i-3	i-2	i-1	i	i+1	i+2	i+3	i+4	i+5	i+n
size	5+3n	20	17	14	11	8	9	12	15	18	21	24	9+3n
syn	11	0	0	1	12	0	21	1	0	0	0	0	8
anti	17	1	0	0	0	0	0	0	0	1	0	0	15
both	3	0	0	0	2	0	4	0	0	0	0	0	2

size – number of atoms in ring
both – number of occurrences with both syn and anti hydrogen bonded

Table 7.5: Numbers of possible amide-main chain rings observed

Chapter 8

Showing hydrogen bonds in relation to protein backbones

The size and complexity of protein molecules has long put biomolecular visualisation at the forefront of computer graphics technology. Searching databases of protein structures creates a need for visually simple, automatically generated pictures which retain a high information content.

In this chapter, several ways of viewing hydrogen bonds in relation to protein structure are compared, and a technique based on drawing the chain through the midpoint of the peptides is presented. Combining this technique with automatic sheet recognition based on hydrogen bonds allows patterns in side chain distributions on each side of a sheet to be easily picked out.

8.1 Introduction

There are a range of algorithms available for abstracting protein structures into cartoons to observe the overall fold with or without regions of higher detail. Rather than simply provide yet another piece of software which has one key display type, in this chapter the emphasis is on how much these different display types have in common, and how easy to specify they actually are. In chapter 10 some conclusions are drawn from the effort of implementing some of these display types, and some practical suggestions are made on what could be done to make the job of the biochemist easier and the job of the programmer

more rewarding. In this chapter, rather than reproducing C code I will assume that one of the key recommendations of chapter 10 has already been adopted, namely that all that is required to specify a display or search algorithm for single protein structure files is a descriptive definition, akin to computer scientists' Pascal-like pseudo-code but based on a language allowing object structures, where a hierarchical protein/residue/atom system is defined, from which it is possible to access atom coordinates, scalar parameters, residue types and even complex constructs such as hydrogen bonds, provided generation rules have been defined (see the end of chapter 10 for some of the significant definitions).

Once a protein object and its constituent atoms can be handled, the algorithms in this chapter work by displaying the structures they define. This assumes that some software exists for this function, but again does not define what that software is. This is in keeping with current developments in computer graphics, where drawing functions are handled by specialised hardware or optimised software libraries: for example on Silicon Graphics computers it is now usual to specify objects in Open Inventor format, which can then be handled by a variety of software libraries and even included in documents to be redisplayed when read by suitably configured viewers. In this work, two different renderers were used, one an X-windows based interactive display based on 2D primitives such as lines, circles and text and the other a ray-marker (a reflection-free ray-tracer) for higher quality output - but both worked from the same object definition algorithms. The figures in this chapter are from the X-based renderer, those elsewhere in this work are from the ray-marker. Chapter 10 contains a discussion about appropriate renderer technology.

The basic operations needed for this chapter are to draw cylinders connecting two points (thick lines for 2D primitives), spheres (circles in 2D), shaded polygons, and to write text. The colours, sizes and drawing styles of these objects can be specified. It is assumed that the order in which these objects are presented to the rendering tool is unimportant as drawing order, depth clipping and buffering will all be handled by the renderer.

8.2 Methods. Displaying Polypeptide Backbones

Polypeptide backbones play a key structural role, and nearly every important interaction is either observed in relation to the backbone or itself involves the backbone. Having a

technique to simplify a protein down to the direction of the backbone chain alone is a prerequisite for interpreting a protein structure, and having a variety of such techniques is useful when the objects or patterns under study are not known in advance.

8.2.1 Detailed visual information, and plots with hydrogen bonds

As a baseline, the all-atom representation of an object is important to consider. Such pictures are far to complicated to be used for study of proteins, but are important nonetheless. The *reason* they cannot be used is significant – proteins are densely packed masses of atoms, a fact which it is easy to lose sight of when working with smoothed, tidied tertiary structure representations.

The all atom representation is easy to define, it is simply a line drawn for each bond, with a sphere of appropriate radius placed at each atom coordinate. Hydrogen bonds can be represented by coloured lines thinner than the bonds themselves, or by dotted lines. For the backbone it is enough to draw the alpha carbon atoms and the peptide atoms, where the hydrogen positions can be defined by a technique such as that described in appendix A. All residues then look like glycine (the alpha carbon hydrogen atoms can safely be left out as they are rarely of any structural significance), except proline which is a special case. Because of its disruptive role in hydrogen bonding, chain conformation and even hydrophobic interactions, it is worth drawing all the proline atoms.

Figure 8.1 shows four strands from the sheet of dihydrofolate reductase. This system is used throughout this chapter to contrast the different sheet display techniques.

ALL-ATOM BALL AND STICK BACKBONE

for chain segment C $\in$ **segment list:**

  for residue $R_i \in$ C**:**

    line { N_i^h, C_i^α, *bond* }

    line { C_i^α, C_i^o, *bond* }

    line { C_i^o, O_i^e, *bond* }

    sphere { N_i^h, *nitrogen* }

    sphere { C_i^α, *carbon* }

    sphere { C_i^o, *carbon* }

    sphere { O_i^e, *oxygen* }

    where R_i =**Proline:**

      line { N_i^h, C_i^α, *bond* }

      line { C_i^α, C_i^β, *bond* }

      line { C_i^β, C_i^γ, *bond* }

      line { C_i^γ, C_i^δ, *bond* }

      line { C_i^δ, N_i^h, *bond* }

      sphere { C_i^β, *carbon* }

      sphere { C_i^γ, *carbon* }

      sphere { C_i^δ, *carbon* }

    elsewhere:

      line { H_i^n, N_i^h, *bond* }

      sphere { H_i^n, *hydrogen* }

    if $R_i < R_{max}$(C)**:**

      line { C_i^o, N_{i+1}^h, *bond* }

for $h_i \in$ **hydrogen bond list**

  if $h_i \rightarrow$ **type** = mc/mc

  and $h_i \rightarrow$ **donor residue** $D_j \in$ **segment list**

  and $h_i \rightarrow$ **acceptor residue** $A_k \in$ **segment list:**

    line { H_j^n, O_k^e, *hydrogen bond* }

pass { **lines, spheres** } **to visualiser**

8.2.2 Ribbon diagrams and protein taxonomy

A turning point in the understanding of protein structure was the development of the ribbon diagrams of Jane Richardson [3]: these were the first effective attempt to show what was common to protein structure while retaining the true shape of the molecules and hence giving a feel for their diversity.

Excellent tools exist for drawing ribbon diagrams, including MolScript [15] and its interpreter raster3D, and techniques have been developed for including side chains and substrates in these diagrams. The algorithm given here for generating "quick and dirty" ribbon diagrams is included to provide a baseline for the other techniques presented in this chapter.

The algorithm takes the ribbon to be a set of lines or a polygon which is everywhere tangent to the peptides making up the backbone. Practically, taking a vector between the hydrogen and oxygen atom of each peptide and using these as ribs perpendicular to the ribbon allows an almost trivial implementation of the algorithm. The difference between extended and helical conformations is handled by always connecting the edges of the ribbon along the shortest edges, ie $\min\{(O_{i-1},O_i), (H_i,O_i)\}$, to prevent the ribbon from being twisted once per residue.

This type of display represents the compromise all protein visualisation must make between the insight into overall fold and the local detail which can be presented. While there is no better way to assign protein secondary structure class and overall fold, it is very hard to include the interactions of single residues into a ribbon diagram consistently, much of the detail of the main chain is lost, and novel properties based on anything other than known secondary structure motifs are very hard to integrate.

There is no obvious place to put hydrogen bond information, as the ribbon itself does not contain markers for the various atoms involved, so there is no absolute way of knowing which atoms are involved in any connecting lines which are drawn.

Figure 8.2 shows the result of using this technique. Although it is not as clear as a true ribbon diagram, some of the problems with the technique are common to any implementation. The system here contains a β–bulge, for example: the bulge cannot be

identified unambiguously by the algorithm.

SIMPLE RIBBON DIAGRAMS

for chain segment $C \in$ **segment list:**
  for residue $R_i \in C$,
  $R_i \neq R_{min}(C)$ **and** $R_i \neq R_{max}(C)$:
    $V_1 = H_i^n O_{i-1}^c$
    $V_2 = H_{i+1}^n O_i^c$

    if $|O_{i-1}^c O_i^c| < |O_{i-1}^c H_{i+1}^n|$:
      for a in { 0.2, ..., 0.8 } **step** 0.2:
        $A = O_{i-1}^c + a.V_1$
        $B = O_i^c + a.V_2$
        line { A, B, *dash* }

    else:
      for a in { 0.2, ..., 0.8 } **step** 0.2:
        $A = H_i^n + a.V_1$
        $B = O_i^c + a.V_2$
        line { A, B, *dash* }

pass { **lines** } **to visualiser**

8.2.3 Alpha-carbon plots

The alpha carbon plot is perhaps the oldest of the cartoon style representations of proteins to be used. Each residue has an alpha carbon atom, and it is of course part of the backbone. It is easy to relate the position of the backbone points to sidechain points if the alpha carbons are used, and the backbone drawn is considerably simpler than the all atom case.

The disadvantage lies in the puckered nature of protein three dimensional structures. Strands are not straight or gently twisted lines when only alpha carbon positions are drawn, but are distractingly puckered. Alpha helices become jagged and hard to distinguish from random coil regions.

Hydrogen bonds are easy to draw, as the residues involved in each can simply have their alpha carbon atoms joined up by a line of appropriate colour and appearance. This gives rise to cases where ambiguities arise: where the donor NH and acceptor CO of one residue are involved in bonds to one other residue, two hydrogen bonds become drawn with the same start and end points. Unfortunately, this case is very common since it is the arrangement in antiparallel sheet. Milner-White et al [46] have developed a technique for drawing these double lines in a different thickness and colour, but the effect is still far less intuitive than for the all-atom representation.

Figure 8.3 shows the application of this technique to the test strands.

BASIC ALPHA CARBON PLOTS

for chain segment $\mathbf{C} \in$ **segment list:**
  for residue $\mathbf{R}_i \in \mathbf{C}$**:**
    if $\mathbf{R}_i \neq \mathbf{R}_{max}(\mathbf{C})$**:**
      line { $\mathbf{C}^{\alpha}_i$, $\mathbf{C}^{\alpha}_{i+1}$, **backbone** }

for $\mathbf{h}_i \in$ **hydrogen bond list**
  if $\mathbf{h}_i \rightarrow$ **type = mc/mc**
  and $\mathbf{h}_i \rightarrow$ **donor residue** $\mathbf{D}_j \in$ **segment list**
  and $\mathbf{h}_i \rightarrow$ **acceptor residue** $\mathbf{A}_k \in$ **segment list:**
    line { $\mathbf{C}^{\alpha}_j$, $\mathbf{C}^{\alpha}_k$, *hydrogen bond* }

pass { **lines** } **to visualiser**

8.2.4 Smoothed Alpha-carbon plots

The jagged appearance can be relieved by smoothing the line. Several software packages allow backbone plots based on alpha carbon positions, but smoothed out by using these as control vertices for a spline function.

There is a simpler way to get much of the same effect: simply average the coordinates with a three residue kernel and the effect is to remove the excessive curvature seen. This averaging can be carried out with a larger kernel, or can be repeatedly applied for the

same effect. Small curves are removed, while the overall path of the backbone is clearly shown.

This approach also has a problem associated with it. The averaging has one stark immediate effect: it hides small twists in the backbone path. Here "small" means involving three or four residues, and unfortunately this includes all the frequently occurring classes of turn and helices. In addition, there is some widening of the structure on repeated application, so parallel and antiparallel strands become divergent and less representative of real atomic positions. Hydrogen bonds can be included, and are useful for keeping information about the types of turns and positions of helices.

This is an approach ideally suited for quickly generating a smoothed backbone plot to see the overall 3D direction of the strand, and is useful, for example, in conjunction with surface plots, where it can be overlaid using transparency to give the relationship between the surface features and the underlying fold.

Figure 8.4 gives an example of this technique. Not that although effective for beta strands and larger loops, smoothed C^{α} plots are no better than pure C^{α} plots for helices: indeed they contain *less* information and are less clear in these regions.

SMOOTHED ALPHA CARBON PLOTS

for chain segment $\mathbf{C} \in$ **segment list:**
  for residue $\mathbf{R}_i \in \mathbf{C}$**:**
    if $\mathbf{R}_i = \mathbf{R}_{min}(\mathbf{C})$ **or** $\mathbf{R}_i = \mathbf{R}_{max}(\mathbf{C})$**:**
      $\mathbf{P}_i = \mathbf{C}_i^{\alpha}$
    else:
      $\mathbf{P}_i = 1/3(\mathbf{C}_{i-1}^{\alpha} + \mathbf{C}_i^{\alpha} + \mathbf{C}_{i+1}^{\alpha})$

  for position $\mathbf{P}_i \in \mathbf{C}$**:**
    if $\mathbf{P}_i \neq \mathbf{P}_{max}(\mathbf{C})$**:**
      line { $\mathbf{P}_i$, $\mathbf{P}_{i+1}$, *backbone* }

for $\mathbf{h}_i \in$ **hydrogen bond list**
  if $\mathbf{h}_i \rightarrow$ **type = mc/mc**
  and $\mathbf{h}_i \rightarrow$ **donor residue** $\mathbf{D}_j \in$ **segment list**
  and $\mathbf{h}_i \rightarrow$ **acceptor residue** $\mathbf{A}_k \in$ **segment list:**
    line { $\mathbf{P}_j$, $\mathbf{P}_k$, *hydrogen bond* }

pass { **lines** } **to visualiser**

8.2.5 Midpeptide plots

Finally, a technique which developed during this study takes a slightly less conservative approach than alpha carbon backbones, but still keeps enough reference to the original coordinate set that other information can be attached to it.

A protein backbone is a sequence of linked peptides, and it is through forces directed more or less towards the centre of gravity of the peptide that it must be stabilised. The alpha carbon is the attachment point of the sidechain, which in most cases is trying to adopt a quite different environment to that of the main chain, and so might be expected to adopt positions well away from the axis of the backbone (and indeed explicit puckering effects add to this distinction). As a result, it might be expected that some averaged position for the peptide would be a better visual guide for the backbone, and this is what is found.

Simply assigning a peptide to a single point midway between the alpha carbons of the bonded residues gives a versatile reference point. A backbone chain can be drawn by connecting the points up: each residue then has a single line segment associated with it, which can be coloured or otherwise marked according to the properties or type of the sidechain. Even the N and C terminus can be included, simply by assigning the missing peptide midpoints to the N^h_{term} or C^o_{term} atoms.

The result is a smooth chain for most of the protein, markedly less jagged than the simple alpha carbon plot everywhere except the alpha helix. In fact, for sheet regions the picture actually adds information over the all atom representation: the sheets can be seen to consist of well aligned strands, with no puckering along the length and very little lateral deviation except at the ends. Parallel and antiparallel look very similar. This suggests that, because sheets represent a mesh of peptides connected in two directions by hydrogen bonds and in two directions by chemical bonds, the resulting structure is a minimised balance over its whole area with very little residual strain.

Hydrogen bonds can be included, with the peptide point as the connection. Since donor and acceptor bonds invariably point in different directions, there is never any confusion between hydrogen bonds which share a single point: each hydrogen bond in the backbone will be represented separately. The appearance of sheets as regular structures is enhanced: they can now be seen to consist of a cluster of rectangular cells, each comprising four residues and two hydrogen bonds. The proportions of these cells is the same throughout the sheet, whether parallel or antiparallel strands are represented.

Figure 8.5 shows the application of this technique to the test case. This is the only example other than the all-atom representation which draws the β-bulge as a distinguishable item and displays each hydrogen bond separately.

SMOOTHED ALPHA CARBON PLOTS

for chain segment $C \in$ **segment list:**

  for residue $R_i \in C$**:**

    if $R_i = R_{min}(C)$**:**

      $P_i^1 = N_i^h$

      $P_i^2 = 1/2(C_i^\alpha + C_{i+1}^\alpha)$

    if $R_i = R_{max}(C)$**:**

      $P_i^1 = 1/2(C_{i-1}^\alpha + C_i^\alpha)$

      $P_i^2 = C_i^\alpha$

    if $R_i \neq R_{min}(C)$ **and** $R_i \neq R_{max}(C)$**:**

      $P_i^1 = 1/2(C_{i-1}^\alpha + C_i^\alpha)$

      $P_i^2 = 1/2(C_i^\alpha + C_{i+1}^\alpha)$

  for position $P_i^1, P_i^2 \in C$**:**

    line { P_i^1, P_i^2, *backbone* }

for $h_i \in$ **hydrogen bond list**

  if $h_i \rightarrow$ **type = mc/mc**

  and $h_i \rightarrow$ **donor residue** $D_j \in$ **segment list**

  and $h_i \rightarrow$ **acceptor residue** $A_k \in$ **segment list:**

    line { P_j^1, P_k^2, *hydrogen bond* }

pass { **lines** } **to visualiser**

Midpeptide representation of the β Sheet

This figure shows the effect of displaying hydrogen bonds in conjunction with a midpeptide plot. Each hydrogen bond is guaranteed to be represented, although parallel and antiparallel sheet is not distinguished. In fact, all sheets take on a uniform and flat appearance.

The Basis of the Representations Compared

Figure 8.6 gives a schematic comparison of the representations in figures 8.1 to 8.5, showing how their features correspond to the underlying arrangement of mainchain atoms and hydrogen bonds.

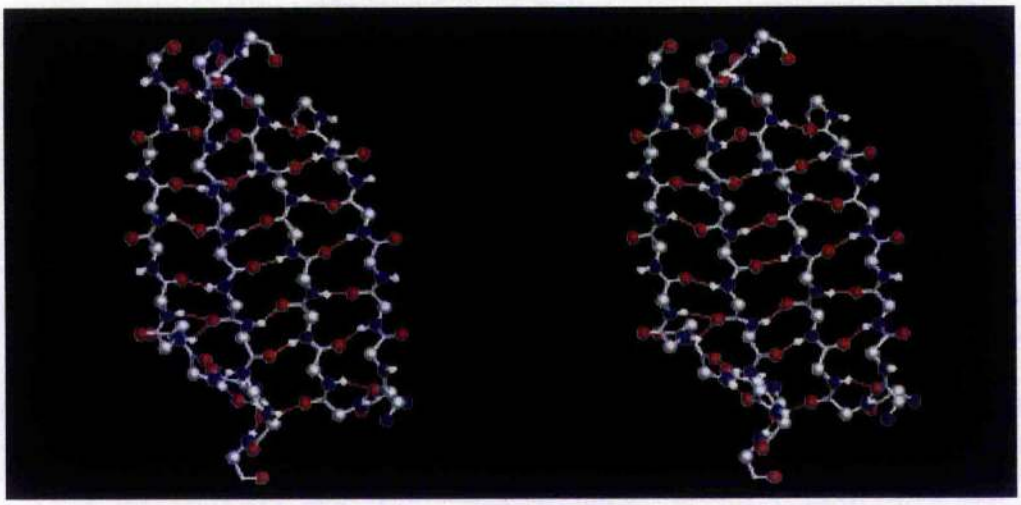

Figure 8.1: Ball and Stick Representation of β Sheet
This shows four strands from the sheet of dihydrofolate reductase: from the left of the picture strands 135–145, 152–161, 1–8 and 110–120, showing two antiparallel β ladders, one parallel β ladder, and one example of a β bulge in strand 135–145, at the bottom left of the picture.

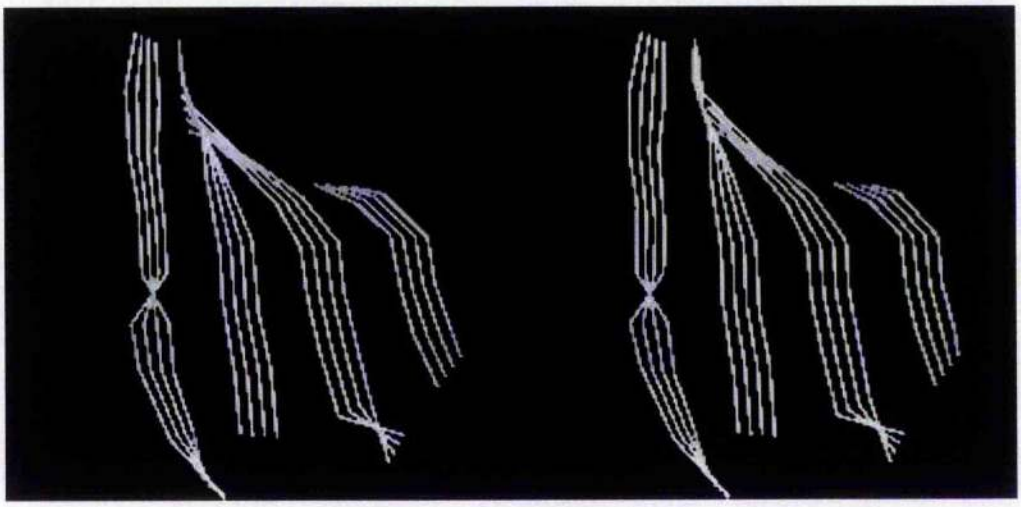

Figure 8.2: A Ribbon Representation of the β Sheet
This shows the four strands of figure 8.1, with guidelines roughly corresponding to the classic ribbon representation of J. Richardson. Note the loss of detail, particularly the β bulge.

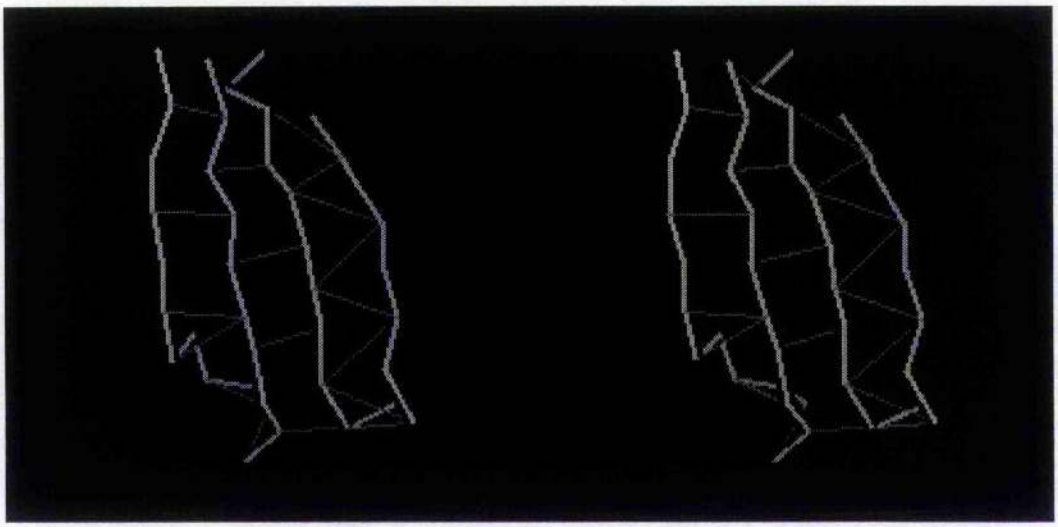

Figure 8.3: A C^{α} Representation of the β Sheet
The four strands of figure 8.1 are shown, with the C^{α} atoms joined to form a schematic backbone representation. Hydrogen bonds are drawn, connecting the C^{α} atoms of the donor and acceptor residues as appropriate. Note that the parallel and antiparallel regions are distinguishable by their hydrogen bonding patterns.

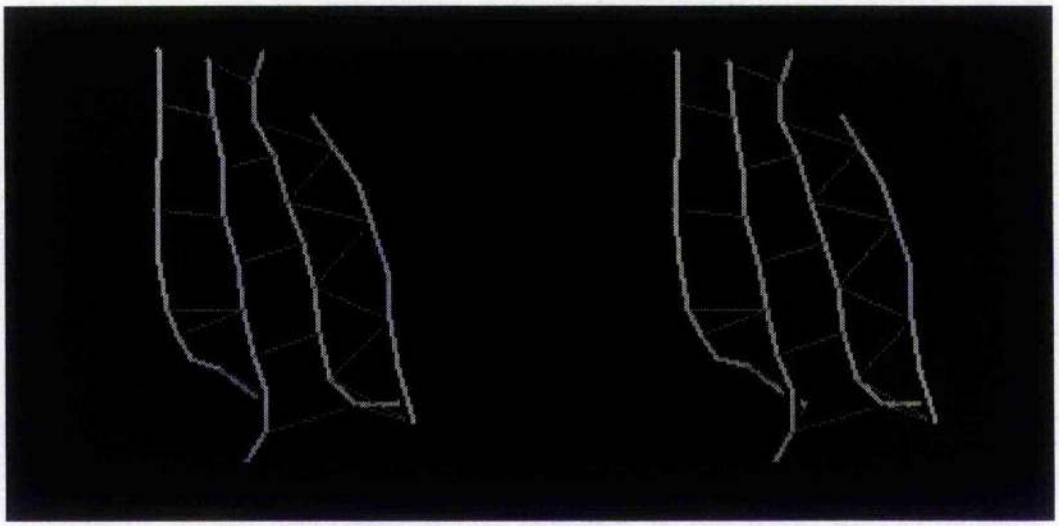

Figure 8.4: A Smoothed C^{α} Representation of the β Sheet
This is the same system as figure 8.3, except that the C^{α} positions have been smoothed by averaging with a 3 residue kernel as described in the text.

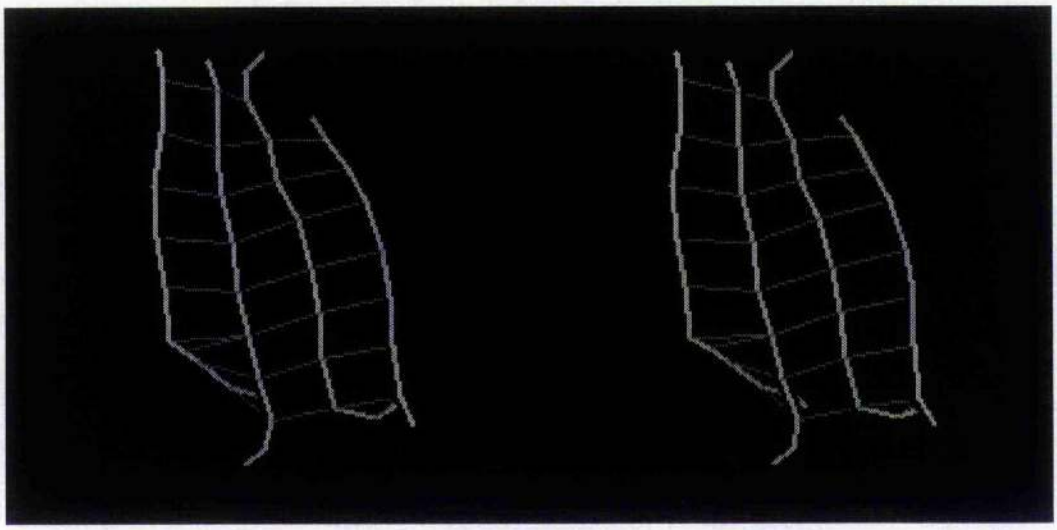

Figure 8.5: Midpeptide representation of the β Sheet
This figure shows the effect of displaying hydrogen bonds in conjunction with a midpeptide plot. Each hydrogen bond is guaranteed to be represented, although parallel and antiparallel sheet is not distinguished. In fact, all sheets take on a uniform and flat appearance.

8.3 Results. The appearance of beta sheets in proteins

8.3.1 Picking Sheets out of Midpeptide Representations

The picture of the beta sheet given by midpeptide plots is remarkable in the way that it picks out a smooth, rectilinear pattern for even the most irregular of sheets, even sweeping away the differences between parallel and antiparallel structures. It can only be assumed that this is a feature grounded in the physical stabilisation of the sheet structure: the midpoint of a peptide is the approximate centre of mass where the forces of hydrogen bonding, chemical bonding and bond torsion act, and the sheet structure represents a balance between the internal energy of the dipeptide and tri-peptide interactions along the chain length and the hydrogen bonding and steric effects perpendicular to each strand.

The resulting flat structure lends itself rather well to being shaded, perhaps better than the ribbon representation, because any distortion in the sheet or break in the hydrogen bonding pattern is easily seen and any lines drawn are rooted firmly within one of the units of protein structure, so other features such as sidechains and substrates can be drawn as usual without serious risk of unwanted intersection.

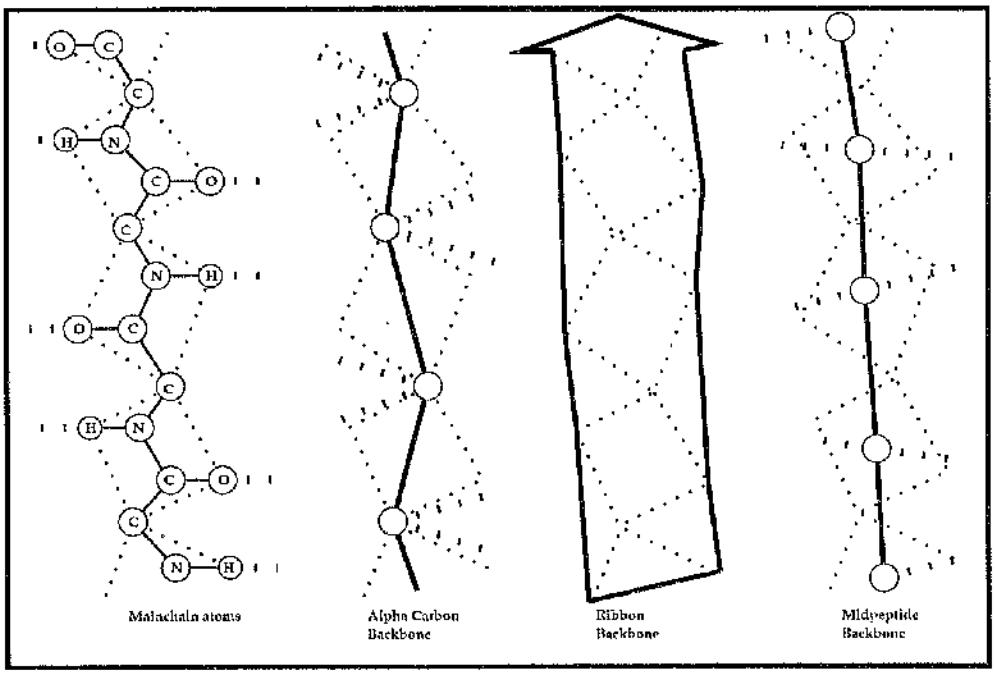

Figure 8.6: The Basis of the Representations Compared

This figure gives a schematic comparison of the representations in figures 8.1 to 8.5, showing how their features correspond to the underlying arrangement of mainchain atoms and hydrogen bonds.

Each of the rectilinear units can be picked out using a search on a list of hydrogen bonds similar to those described in chapter 7. Search details are given in tables 8.1 and 8.2. Basically, a unit is a quadrilateral made up of four midpeptide points, the two chain segments and the two hydrogen bond representations. This can be divided into two or four triangles and each one rendered as a filled, double-sided polygon. Distinguishing the two possible directions that the quadrilateral can face results in a smooth representation for parallel sheets and a checkerboard pattern for antiparallel sheets, which may be all the discrimination needed in many cases: otherwise, arrows must be introduced along the chain length.

root	main/main	*	N/O	donor	N^H	O^C	*	
derived1	main/main	*	N/O	acceptor	O^C	N^H	x=0	y=2
derived2	main/main	*	N/O	donor	N^H	O^C	x=2	y=2
derived1	main/main	*	N/O	acceptor	O^C	N^H	x=2	y=4

Table 8.1: Search pattern for a section of parallel sheet

root	main/main	*	N/O	donor	N^H	O^C	*	
derived1	main/main	*	N/O	acceptor	O^C	N^H	x=0	y=0
derived2	main/main	*	N/O	donor	N^H	O^C	x=2	y=-2
derived1	main/main	*	N/O	acceptor	O^C	N^H	x=2	y=-2

Table 8.2: Search pattern for a section of antiparallel sheet

Figure 8.7 shows the effect of shading on the sheet of dihydrofolate reductase. Even features such as β bulges can be retained, and the shading helps to identify areas where the sheet is most strongly twisted.

8.3.2 Showing Sidechains in Schematic Representations

For a whole protein, displaying all sidechains leads to a confusing overall effect: but backbone plots provide no sequence information, the defining property of the protein under investigation. Putting a mark at the C_β position of each residue allows the direction of sidechains to be seen. Figure 8.8 shows this for the dihydrofolate reductase strands of figure 8.1.

The midpeptide plot is particularly amenable to sidechain representations. in figure 8.9 the sheet of figure 8.1 is shown: lines of residues can be seen running alternately above and below the sheet perpendicular to the strands, crossing regions of both parallel and

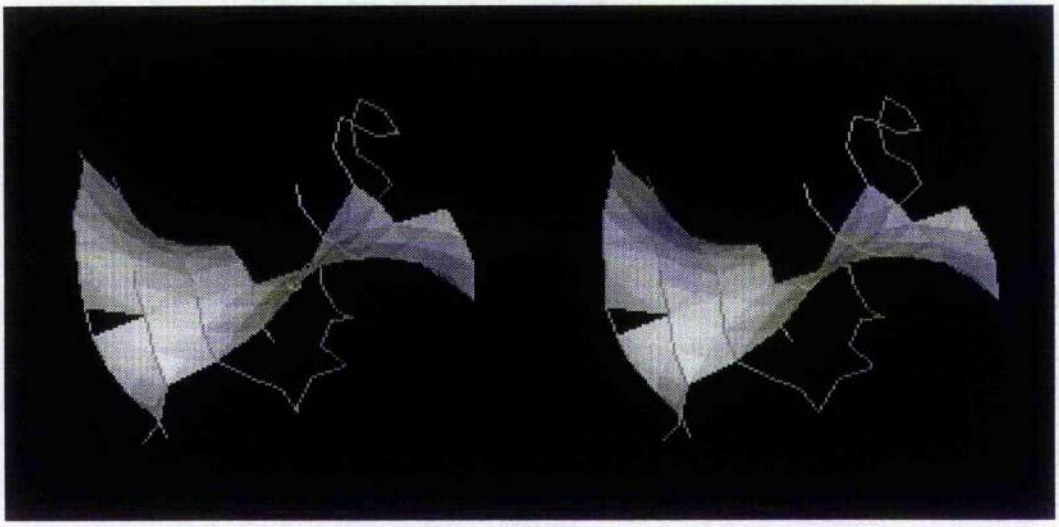

Figure 8.7: Picking Sheets out of Midpeptide Representations
This figure shows the whole sheet of dihydrofolate reductase, and two helices important for substrate binding, as a midpeptide plot with the sheet picked out and shaded. The β-bulge appears as a tear in the sheet.

antiparallel sheet. The direction of the β-bulge can also be seen.

Full ball-and stick sidechain representations are possible, if the place of the C^{α} atom for each residue is taken by the midpoint of the line joining the two peptide midpoints it shares. The resulting C^{α}-C^{β} bond is stretched slightly, but all other side chain atoms can be treated as normal. It is even possible to represent proline, with C and C^{α} both taken by points on the midpeptide chain: the result is considerably distorted, but it is still possible to see proline positions and pucker directions easily, which is usually all that is needed.

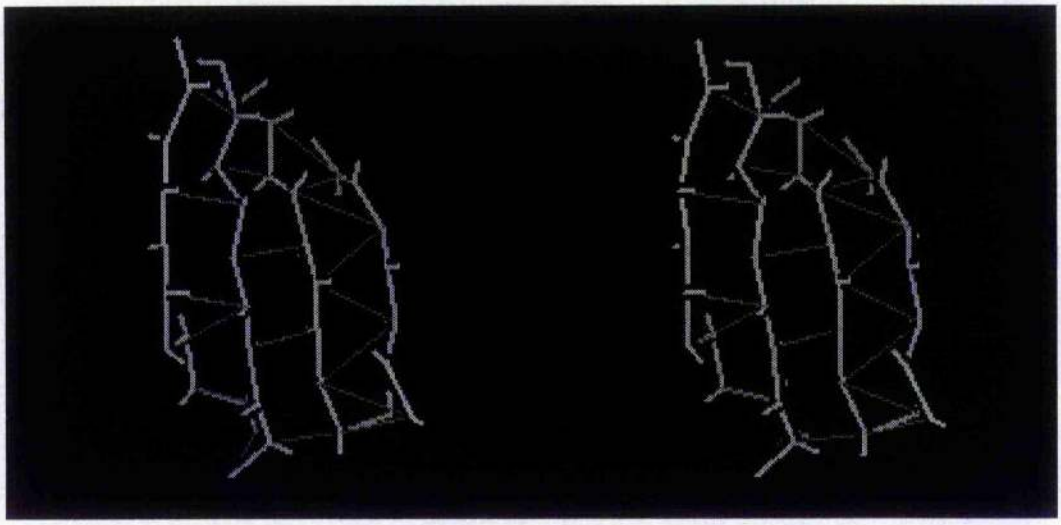

Figure 8.8: Showing Sidechains in Schematic Representations
This figure shows the effect of putting a mark at the C_β position of each residue for the alpha carbon plot of figure 8.3, allowing the directions of the sidechains to be seen.

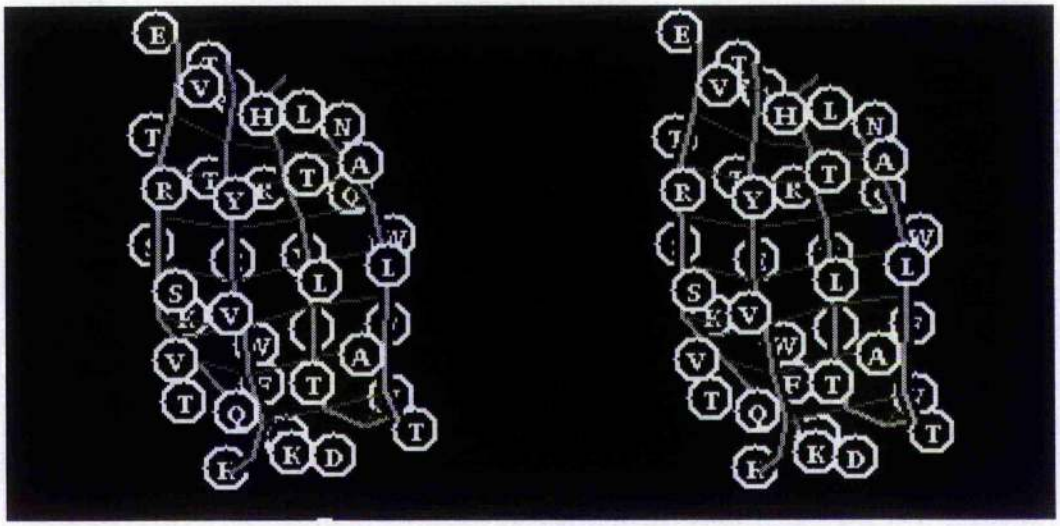

Figure 8.9: Showing Sidechains in Midpeptide Plots
This shows the effect of placing a marker for the side chain at the C^β position for each residue. Ridges of residues running perpendicular to the strand direction can be seen, as can the +direction of the β-bulge, in this case with the extra residue, a Valine, on the near side of the sheet.

Chapter 9

Hydrophobic ridges in beta sheets

It is known that strand-strand recognition in proteins is guided by a preference for certain residues to pair up in adjacent sites on the same side of the sheet in neighbouring strands. Using the approach from the previous chapter it is shown that this preference extends across the whole width of a sheet, giving rise to stabilising hydrophobic ridges which appear to be a feature of all beta sheets, large and small.

The ridges observed in a range of proteins are examined, and their importance is tested by showing how their conservation in proteins related in sequence is directly related to the structural similarity of those proteins.

9.1 Introduction

The tertiary structures of beta sheets in proteins have been subjected to extensive statistical analysis and modelling, but have remained as mysterious as any other feature of protein structure. While some patterns are clearly apparent in amino acid preferences of beta strands and their overall relationship to protein architecture, beta sheets are no more predictable than any other structural element in proteins.

In the face of this, it is interesting to see whether there are any new insights which can be gained from new techniques of visualising the structural patterns defined by hydrogen bonding in proteins containing extensive sheet structures. The method described in the previous chapter, plotting hydrogen bond position as a function of midpeptide position, has the particular virtue that it identifies sheet structures in proteins without requiring any input from the researcher and without abstracting the sheets so that they can no

longer be combined with other data such as side chain position.

This chapter shows how the pictures given by this method provide a particular extension to work by Lifson and Sander [47, 48] on inter-strand amino acid preferences, and in particular suggest a concept of *hydrophobic ridges* as structural elements of globular proteins with a possibly strong tertiary-directing effect. The technique is tested on a small number of systems, and raises some interesting questions about how these elements could be used to help in a more general study of protein architecture.

9.1.1 Amino acid preferences for beta sheet residues

Beta sheets have been analysed for length, direction, type and handedness of crossover connections, and a number of other statistical properties which could be classified as *architectural*. More relevant for the results presented here, studies have highlighted the importance of inter-strand nearest neighbour interactions.

The first study to show that the inter-strand nearest neighbour interactions were specific, with certain pairs of residues or pairs of residue types being significantly more common than would be expected if they were purely a function of random chance, was the work of Lifson and Sander [48].

The pairwise groupings they found reflect both a preference for some types of non-polar residue to become nearest neighbours, with preferences representing stereotypical stacking interactions and also for certain pairs of polar residues to match up, either to form salt bridges or some stabilising hydrogen bonding pattern. The most specific non-polar interactions are between Ile and Val and Ile/Leu, with differences in preference for antiparallel and parallel sheets which were significant enough to not be an artifact of the small size of the database used in the original study.

9.2 Methods. Midpeptide plots of beta sheets

The method described in the previous chapter is used here to pick out the beta sheets in proteins based on their hydrogen bonding patterns. As explained previously, taking the midpoint of the peptide bond as the representative point at which to draw the end of a hydrogen bond has the effect of regularising the observed structure: the protein backbone is a set of rigid units with bonding and non-bonding interactions pulling them in different

ways, but the net effect is a stable structure, with the basic elements (strands, loops and helices) pulled into a local energy minimum. The shape this describes for a sheet is surprisingly uniform.

The slightly twisted, flat sheets defined in this way can clearly be seen without needing to pick out the strand residues with any other marker. This means that the images are simple enough to allow extra information to be provided. An added advantage of the technique of midpeptide smoothing is that the C^{α} and C^{β} positions are not being used or distorted in any way for the backbone model, and as a result information can be added at these points in the images. This makes it easy to display sequence information along with the structural information already being represented.

Placing all the sequence information creates a problem, though. The large number of residues in a typical protein quickly fill the image with superimposed illegible characters. However, any sheets present will have a smooth, flat representation, and this when viewed in isolation can have sequence data superimposed and still be understood, particularly if seen in three dimensions.

One of the advantages is that the sheets could be identified without having to name the beta strands in advance, and it would be nice to keep this feature. This is possible to a large extent by looking at the way the structures are drawn. Sheets are characterised by a tiling of slightly twisted rectangles, with two sides made up of the spars connecting adjacent peptides and the other two formed by the representations of hydrogen bonds. These can represent a number of different ring structures, as shown in figure 9.1, but all can be identified using the same technique described in chapter 7 for identifying hydrogen bonded rings from the hydrogen bond file used to generate the image in the first place. If only residues participating in these rings are displayed, the result is a summary of the super-secondary structure of the protein. It turns out that α helices can be defined as rings in a similar way, and also shown.

Two additional features of sheets and helices also have to be taken into account. One is that the residues at the ends of strands or helices are also participating - since although a peptide (and hence a midpeptide point) can be regarded as belonging to either of the amino acids contributing to it, the actual hydrogen bond is regarded as belonging to the residue which provided the atoms involved - in other words, bonds involving C=O "belong" to the previous residue, those involving N-H belong to the "next" residue. The

other feature is specific to sheets - the phenomenon of β bulges, individual residues within strands which act as insertions, offsetting the hydrogen bond pattern by one residue. Both of these classes of residue are of interest, and form annoying omissions if neglected. Both can easily be incorporated by simply adding in residues on either side of the ones detected as participating in the rings defined above.

9.2.1 Midpeptide plots highlight similarities between parallel and antiparallel sheets

One striking feature of the midpeptide plots is that they do not distinguish parallel from antiparallel sheets. This is in one sense frustrating, as it means that extra cues have to be provided to allow the researcher to classify the sheets which can be seen - but this could be done by simply adding a small arrow on the end of each strand detected (i.e. on the trailing residues discussed above). In another sense it is useful, as it means that common features between different classes of sheet can be detected.

As figure 9.1 shows, the tiles on the sheet represent three different types of ring structure. But what they all have in common is that, *perpendicular* to the directions of the strands, the sidechains of residues are roughly co-linear, and their C^{α} and C^{β} atoms project above the plane of the sheet the same amount and *in that same direction*. What the grids show, when viewed parallel to the lines defined by the hydrogen bonds, are columns of residues which are forced to pack together.

Adjacent columns point in opposite directions, alternately above and below the plane of the sheet. This is the case both for parallel and antiparallel sections of the sheet, and can be clearly seen by placing a marker for each residue at the C^{β} position.

9.3 Results: ridges across strands

What these columns represent are ridges of residues forced into contact across the width of a sheet. The individual pairs which make up this structure are governed by the statistics of Lifson and Sander, but looking at the whole sheet at one gives a new insight. Individual pairs may be chosen on the basis of hydrogen bonding between sidechains, the formation of salt bridges involving charged residues, or specific packing interactions. But looking at the whole sheet shows a pattern which goes beyond pairwise.

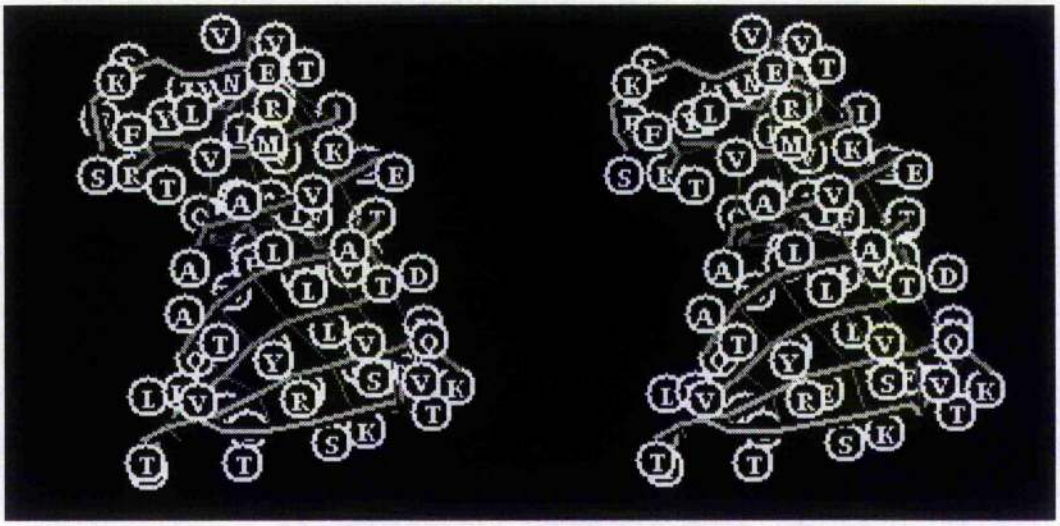

Figure 9.1: Ridges of Sidechains in β Sheets

This figure shows the sheet of dihydrofolate reductase as a midpeptide plot with residues drawn at C_β positions. The alternating ridges of sidechains can clearly be seen. Note in particular the long hydrophobic ridge V-V-L-L-A-V-V-V extending the full width of the sheet.

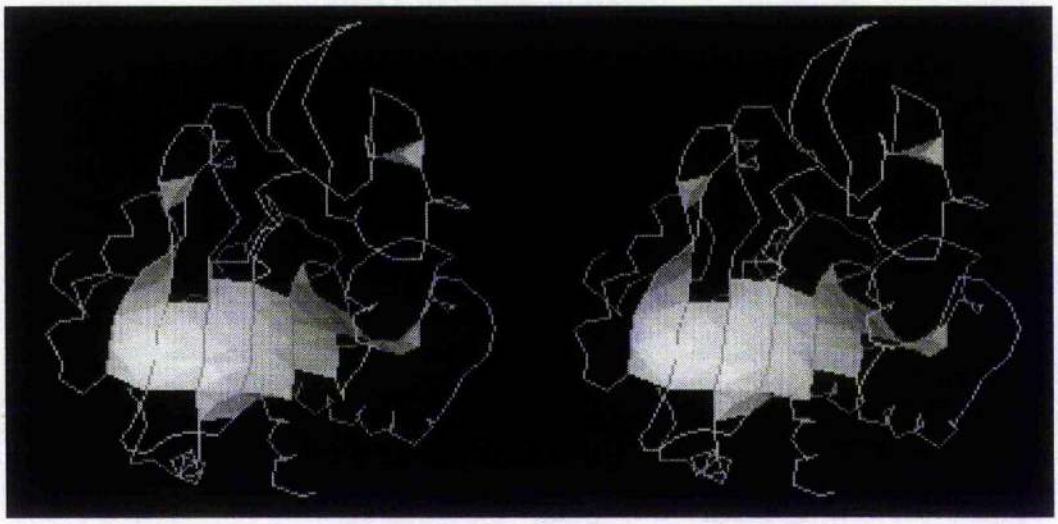

Figure 9.2: The β Sheet from Carboxypeptidase

This figure shows the carboxypeptidase sheet, defined using the hydrogen bond criteria described in chapter 8, and represented as a midpeptide plot with the sheet shaded. The whole protein is shown, but the sheet structure dominates.

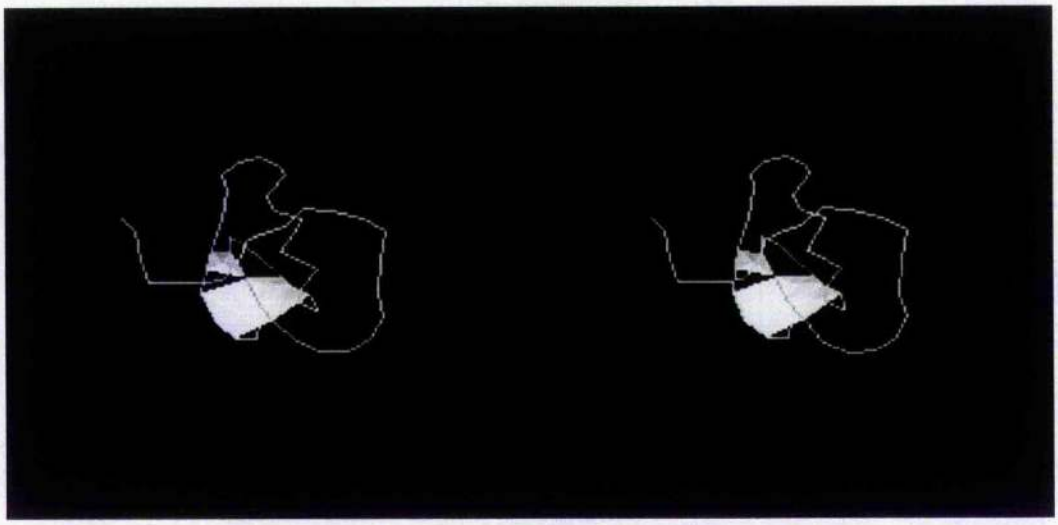

Figure 9.3: The β Sheet from Ovomucoid, Domain 3
Even small sheets can be picked out by midpeptide plots - this shows the whole of ovomucoid domain 3 with the sheet picked out using the same technique as figure 9.2

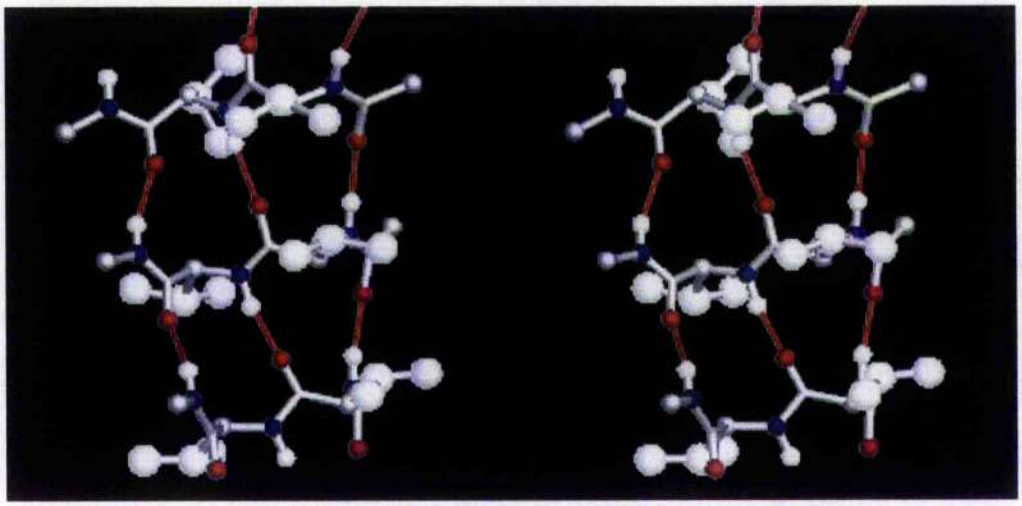

Figure 9.4: Hydrophobic Stacking along Ridges
This figure shows the two V-V-V ridges at the edge of the dihydrofolate reductase sheet, with sidechains drawn in to demonstrate the close hydrophobic packing which can stabilise a sheet.

Among the most obvious, and strongest, pair correlations were those between the non-polar sidechains. It is to be expected, therefore, that there will be some tendency for these contacts to add up to provide ridges of hydrophobic residues stretching over several strands. In fact, what is observed for proteins with a distinct beta sheet kernel is one or more distinct ridges of hydrophobic residues stretching the full width of the sheet. This chapter does not contain a full review of all available protein structures, unfortunately, but the patterns detected here provide some very suggestive ideas on the stabilisation of beta sheet proteins.

9.3.1 Wide sheets: carboxypeptidase and dihydrofolate reductase

The diagrams for the sheets of carboxypeptidase and dihydrofolate reductase, figures 9.5 and 9.6 respectively, both show the same principle: wide sheets are anchored by hydrophobic ridges stretching the full length of the sheet.

Carboxypeptidase shows two interesting properties of the sheet when viewed as a flat grid. The first is that there are three clear ridges of hydrophobic residues, 191A-I-I-F-33V, 204Y-F-L-I-L-50L, and 240Y-L-F-I-L-I-48Y which together can be seen to splice together the eight strands identified by the hydrogen bonded ring classification. The second is that all three of these ridges are on the same side of the sheet, and represent a single contiguous "bed" of hydrophobic residues. This feature is not ubiquitous; other proteins in this small sample show that ridges commonly occur on both sides of the sheet. In carboxypeptidase the pattern is intimately related to the evolution and function of the enzyme. Several of the active site residues of this protein are part of the sheet itself, and the hydrophobic face can be seen as an anchor for the necessarily less stable active, hydrophilic side of the sheet.

Dihydrofolate reductase shows a different strategy for a similarly wide sheet. Here there are two hydrophobic ridges at the centre of the sheet, 157V-L-L-A-V-V-75V and (139V)-W-L-F-I-V-V-74V (that first Valine residue in the second ridge is part of a beta bulge), adjacent but on opposite sides of the sheet. The sheet itself is strongly twisted at one end and appears to have a weak spot traversed by these two ridges. This sheet does not have a clear hydrophobic side, and if there is any significant stabilisation through hydrophobic effects it must be involving these two hydrophobic ridges, since together they account for more than half of the hydrophobic residues in this sheet.

9.3.2 Narrow sheets: ovomucoid domain III, human prealbumin

The pattern is not only found in proteins with wide beta sheets. Narrow sheets appear to be stabilised by one or more short hydrophobic ridges, again running the full width of the sheet.

For example, ovomucoid domain 3 displays a tiny kernel of beta structure shown in figure 9.3 and schematically in 9.7, with only 8 residues classified as sheet by the hydrogen bonded ring criteria. However, even in this case there is a distinct hydrophobic ridge, 53F-L-31Y.

More significantly, each monomer of the human prealbumin dimer is formed from a sandwich of two beta sheets, shown in figure 9.8. Each of these is four strands wide and shows a four residue hydrophobic ridge - 93V-V-V-43A and 55I-V-I-120A. There are other hydrophobic pairings which do not form part of longer ridges, as might be expected, but the two ridges are quite distinct and this does seem to suggest that their role is more than just coincidental.

9.3.3 Conservation: trypsinogen/tonin and actinidin/papain

Of even more interest is whether or not these ridge structures are conserved between evolutionarily distant proteins of similar structure. If they are then it is reasonable to assume that they are playing some significant role, although it may also be assumed that any hydrophobic residue may be hard to replace for other reasons, such as packing against the hydrophobic side of a helix or some crucial role in a folding intermediate.

First, trypsinogen and tonin can be considered. Both of these can be classed as beta barrel proteins, although looking at the midpeptide plots it is easy to see that they are actually constructed as single sheets twisted and then linked in such a way that there is no one line of hydrogen bonds forming a ring around the whole barrel. It is better to see these two proteins as being sheets whose diagonally opposite corners have been joined together.

The forces stabilising beta barrels might be expected to be quite different, but there are still distinctive ridges in these two proteins. The sheet is not rectangular, so no ridge can extend the full length, but in trypsinogen (figure 9.9) there are two ridges which span two of the widest parts of the sheet, 29Y-L-V-104I and 83I-V-L-41F. There is also a triple

of residues next to the first ridge, 53V-L-89I.

In tonin (figure 9.10), there have been significant changes to the upper right hand quadrant of the sheet (as drawn here), and the second ridge has not been conserved - in its place are some individual pairwise interactions presumably performing a similar role. The missing ridge is associated with a significant change in sheet structure, with a β bulge in trypsinogen completely absent in tonin. However, the first ridge is still present, now 29W-L-V-106L, and the triple adjacent to it is now a four member ridge, 45V-I-L-89F. As in the papain/actinidin case, the residues themselves are not strongly conserved, but the hydrogen bond pattern and the hydrophobic ridges are.

The second pair considered are actinidin and papain. Both of these proteins have related, complicated sheet structures, and it is in cases like this where the midpeptide plots really show their worth. The sheet of each is forked, but by plotting the hydrogen bonded grid it is possible to flatten the sheets out and compare the two. As expected by now, there are hydrophobic ridges, one serving each branch of the fork.

In actinidin (figure 9.11) the ridges are 5V-Y-W-I-152F and 214Y-V-I-V-194M, while in papain (figure 9.12) they are 5V-Y-I-I-149F and 208Y-V-A-I-187I. These mutations are not what would be classed as conservative in every case either through the Dayhoff matrix or in terms of the Lifson and Sander statistics (particularly the I→A conversion in the second ridge), but when seen as a ridge of hydrophobic residues, the pattern of hydrogen bonds and of hydrophobic residues is conserved. This clearly implies that the conservation of these ridges has some structural significance, even when the surface defined by these residues (in other words, packing effects against the face of the sheet) can be significantly altered.

9.4 Conclusions: hydrophobic ridges stabilise beta sheets

Although this work was not exhaustive, it does seem clear that hydrophobic ridges are a significant feature of beta sheets. Determining whether they are simply a statistical effect of pairwise preferences extending in both directions (or conversely, pairwise preferences are simply consequences of the need to form hydrophobic ridges) would require a statistical analysis of a large sample. Sadly, the search software in its current form cannot find features like the ridges, as that would need a recursive definition in which offset hydrogen

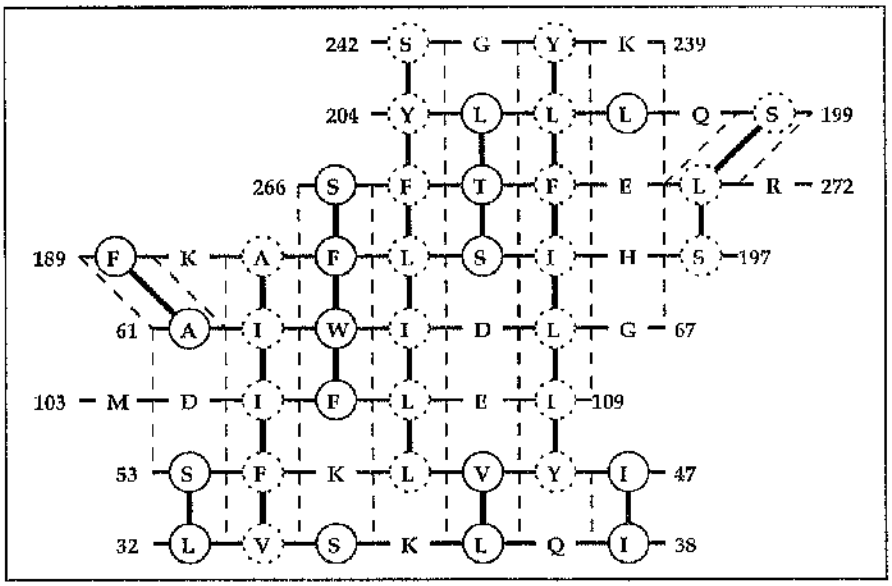

Figure 9.5: Hydrophobic Ridges in the β Sheet of Carboxypeptidase

This and the following figures show schematic representations of midpeptide plots like that in figure 9.1. To better illustrate the regularity of the patterns of residues seen, the sheets have been flattened out, but the pattern of hydrogen bonds has been kept intact. Dotted circles represent sidechains on the near side, solid on the far side of the sheet. Hydrophobic residues [AFILVWY] are highlighted, as are [T] (mixed hydrophobic/polar) and [S] (polar, but with a particular propensity for the edges of beta sheets, so probably a sheet-edge stabiliser). Where pairs of these residues are found, a thick line joins them. In this and the other proteins in this sample, this reveals prominent hydrophobic ridges defining the sheet.

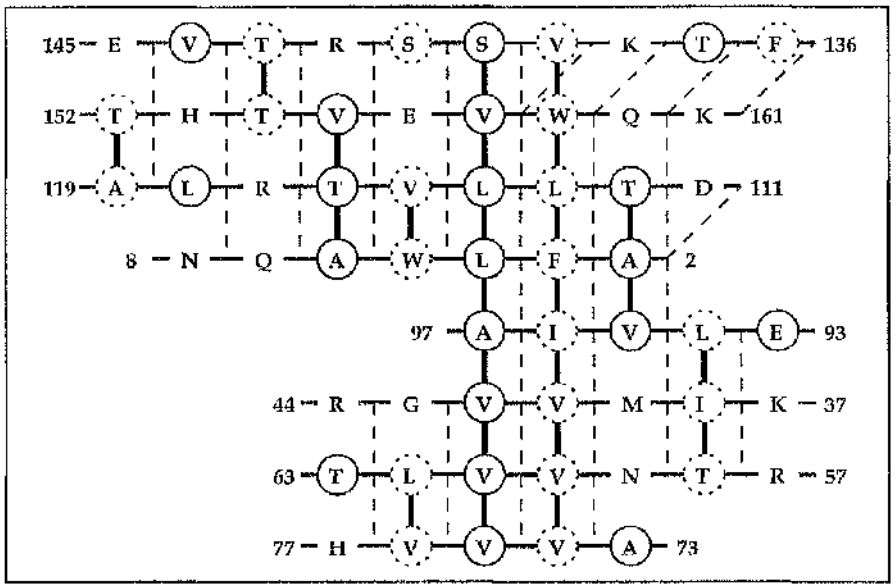

Figure 9.6: Hydrophobic Ridges in the Sheet of Dihydrofolate Reductase
This is a very striking case, with two hydrophobic ridges stretching across two distinct β sheet domains.

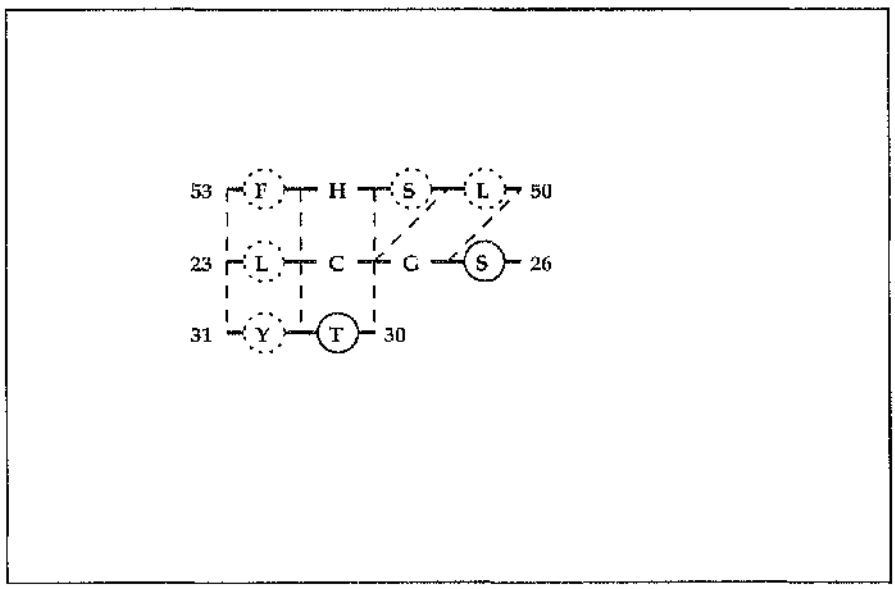

Figure 9.7: A Hydrophobic Ridge in Ovomucoid Domain 3
This tiny protein fragment has 9 residues in what could be termed a sheet structure. Even in this case, a ridge of hydrophobic residues is found.

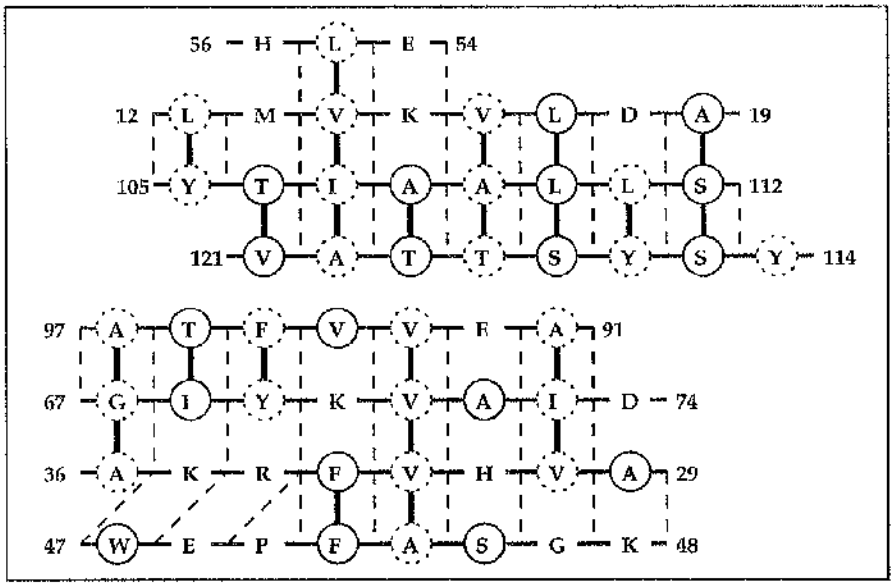

Figure 9.8: Hydrophobic Ridges in Human Pre-albumin

This protein forms a dimer; only the sheet from the monomer is shown here. The four-strand sheets both have four-residue hydrophobic ridges stabilising them. The dimer has strands 114-121 of each monomer hydrogen bonded antiparallel to each other to form one large sheet. The junction is at S117, giving ridges L-L-S-S-L-L, V-A-T-Y-L, and A-T-S-S-A, unusually serine and threonine rich.

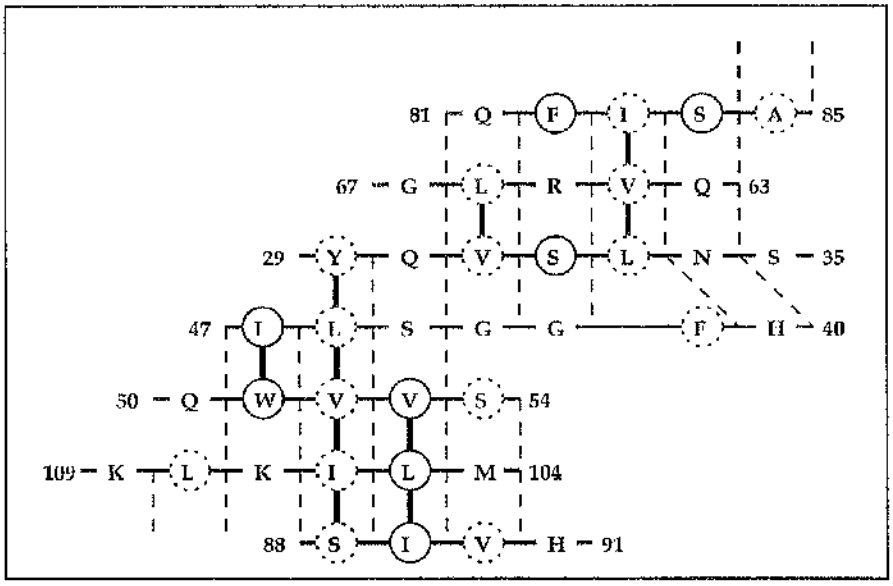

Figure 9.9: Hydrophobic Ridges in Trypsinogen

This figure shows the first β sheet in trypsinogen, which forms a β-sandwich: the "hanging" hydrogen bonds from residue A85 match those of L108 to complete the sandwich.

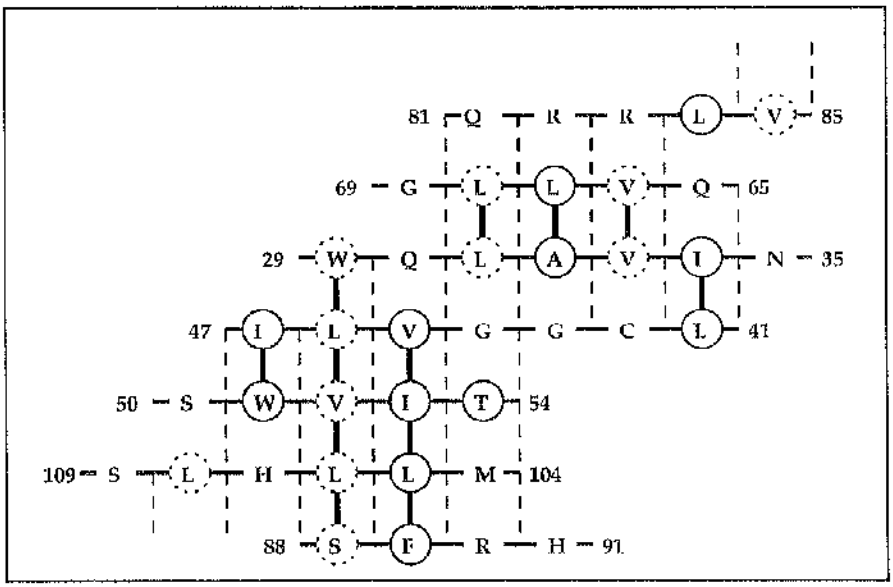

Figure 9.10: Hydrophobic Ridges in Tonin

As in figure 9.9, the sheet shown here is a sandwich, so residues V85 and L108 connect to complete the hydrogen bonding pattern.

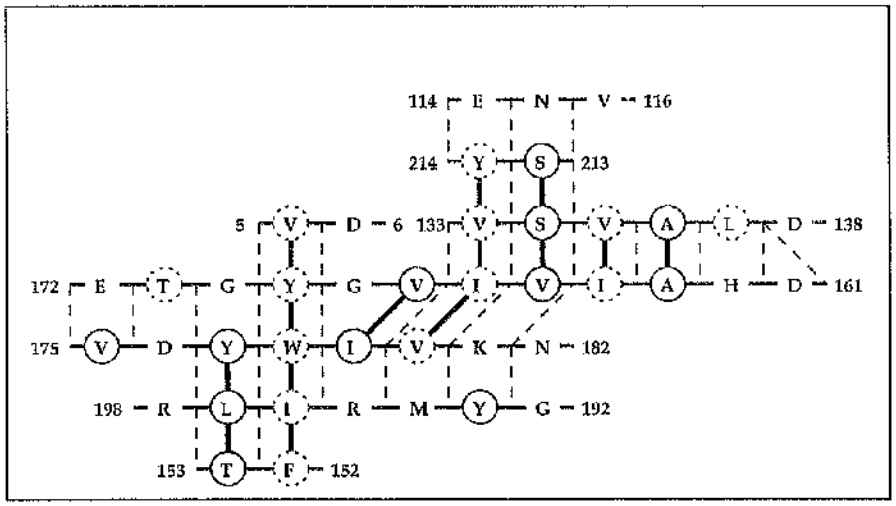

Figure 9.11: Hydrophobic Ridges in Actinidin

The sheet in actinidin appears "torn" in a midpeptide plot, with strand 159-167 binding to both strands 5-6 and 133-138 on the same side. It is still very easy to follow the path of hydrogen bonding and sidechain direction by switching between the representation styles of figure 9.1 and 9.2.

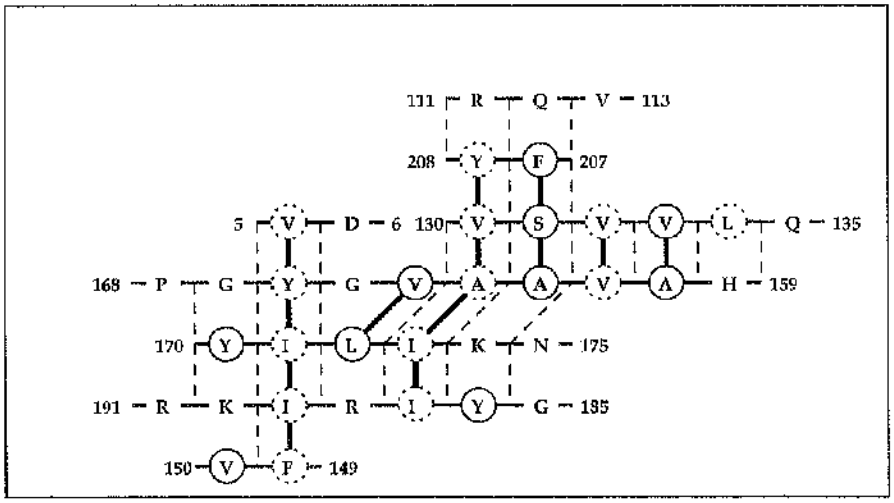

Figure 9.12: Hydrophobic Ridges in Papain

This sheet shows the same distortion as actinidin: using a flat midpeptide representation helps to emphasise the similarity of the two folds.

bonds could be taken as new roots, to allow data on ridge properties to be tabulated automatically.

Results from other sources do suggest that

> Unlike α-helix formation, β sheet formation is determined in large part by tertiary context...and not by intrinsic secondary structure preferences.[49]

according to Minor and Kim, in the discussion of experiments which showed free energy changes for moving residues from the edges to the centre of strands correlated strongly with the residues water/octanol partition function and not at all with statistical β sheet frequencies for those residue types.

An effective model of the beta sheet will have to go beyond sequence data and pairwise strand matching and deal with the full tertiary structure of the sheet, with effects parallel and perpendicular to the strands and distinct hydrophobic and hydrophilic environments to be accounted for.

Chapter 10

Problems in Protein Visualisation

Interpreting protein 3-Dimensional structure is a many-faceted activity, shaped by the features which are of current research interest but also by the databases and software which happen to be available. Most software is either weighted towards detailed representations used in determination and analysis of single structures or cartoon simplifications for general classification of related folds. Work towards the production of software which can easily handle comparison of structures across a family of related structures, using new visualisation techniques where these become available, representing information from the wide variety of sequence related databases which are produced, and switching easily between 1D,2D and 3D representations has been hampered by the different agendas of structural chemists, molecular biologists, and computer scientists.

This work brought the author into a close working relationship with all three subject areas, and this chapter presents an interpretation of the current state of the art, the missing pieces of software and interface, and a specification for a system which could be implemented with existing or easily developed technology which would bridge the gaps between the different classes of knowledge required at different levels of interpretation.

10.1 Introduction

10.1.1 The requirements of structural biochemists

Structural biochemists and crystallographers require tools which can show the detailed structure of parts of a protein, in conjunction with the forces and physical constraints

which affect shape, stability and interactions with other molecules. They often need to visualise electrostatic potential fields, solvent exposure, and individual hydrogen bonds.

Analysis of possible new types of contribution to protein stability will often require the development of new techniques for visualising interactions.

10.1.2 The requirements of experimental biochemists

Molecular biologists work with protein sequence, and are interested in the effects of single residues or sets of residues which can be identified as significant by genetic manipulation. They also work with families of proteins on the same basis. Their needs from a structural database are quick identification of the overall fold of a protein, identifying the relative position of loops and residues of interest, and comparing the structures of sequence related molecules. Data from software for 1D structure analysis must be imported to the 3D model, often by a researcher too busy to learn a graphics scripting language for one brief session in front of a graphics machine.

Enzymologists need tools to identify active site pockets, place active site residues within them, and recognise the local chemical environment of the transition state binding.

10.2 Methods. Tools for protein visualisation

10.2.1 Simple graphics tools for prototyping

Providing a simple way to visualise a 3 dimensional object and annotate it with information is part of a more general class of problems in scientific visualisation which is unresolved. However, there are some parts of the problem which are easily handled, and these have yet to be compiled into a system which is easy enough to use and still flexible.

Ribbon diagrams, for example, are hard to specify complex 3D objects, but the basic principle, a ribbon tangent to the plane of the peptide bonds, is quite easy to describe and draw. Likewise, the backbone displays described in chapter 8 all have a definition in terms of the molecular model, but are then drawn as geometric primitives not centred on those molecular positions but on derived points.

Writing a visualiser is a time consuming and often repeated activity. Developments in computer graphics mean that soon there will be standards in scene description (VRML, OpenInventor) and in primitive display (PHIGS, OpenGL) which will render much of this

work obsolete. It is important that chemists are free to focus on the description of the visualisation in abstract terms, while computer scientists and hardware developers provide the current state of the art in scene visualisation and manipulation.

To encourage this way of thinking, even in the absence of a set of tools to provide the functionality, I would suggest that an approach be adopted whereby chemists communicate their ideas on molecular properties in a standardised format, independent of computer language or visualisation system, which can then be used as the basic input system to viewers and interpreters. The system I suggest is based on the idea of a hierarchical class, since this fits naturally into the idea of proteins as sequences of residues, but this is merely a useful shorthand: trends in computer science change, and Object Oriented programming is a current useful paradigm, but the descriptions should also be interpretable by languages such as C and Fortran, and by systems which have no hierarchical model such as AVS.

10.2.2 Relating sequence and biological activity to structure

A large part of the work at Glasgow has been directed towards tools for relating sequence to three dimensional structure. Past successes in this approach have included an interpretation of turn types based on observed hydrogen bonding patterns and software for quick comparison of the hydrogen bonding patterns and hence the topology of related proteins. It seems clear that a system which maps 1D properties on to the three dimensional chain of the protein, in particular allowing interactive switching between the two types of representation, would be of great value. A prototype system for this has been developed in collaboration between Glasgow computing science and the University of North Carolina at Chapel Hill, but no package currently available combines enough flexibility with existing usable code.

10.2.3 Fast realistic image generation

Drawing chains constructed from thick lines is a fast and reliable way of getting an interpretation of the three dimensional shape of a whole protein or investigating a single side chain, but to fully understand the forces in the core of a protein it is necessary to have a system for generating space-filling models. Unfortunately, most space filling models are constructed out of spheres, and picturing spheres has a computational overhead either in

terms of square root calculation (see for example Appendix C) or in the large number of polygons required to approximate a sphere if it is to be constructed from flat surfaces.

The advantages of using accurate space filling models even extend beyond looking at packing effects. As the figures in previous chapters show, drawing even simple stick diagrams as composites of three dimensional objects enhances the perception of depth, and helps to promote stereo fusion in 3D viewing. As a result, there have been many attempts to render spheres quickly for molecular visualisation. One of the most promising approaches has been developed by John Patterson, [50] using a corrected parabolic approximation to generate spheres in integer arithmetic (in much the same manner as polygons are rendered).

Figures 10.1 and 10.2 show the results of the two techniques, one the ray-marker used throughout this work, and the other the fast spheres implementation written for this study. The two images are virtually indistinguishable, thanks to corrections to the depth calculations suggested by me and incorporated by John Patterson as third and fourth order differences (details can be found in appendix D). The significant factor is time of calculation: around 100 seconds for the ray-marked version, less than 1 second for the fast spheres implementation. Other fast renderers exist, such as Rasmol by Roger Sayle [51], but they rely on precomputed spheres and hence a lot of memory access and offset calculation. The fast spheres algorithm is at its best when calculating each sphere individually, and it can be combined with perspective effects and variable lighting conditions making it suitable for interactive use. A small image (400x400 pixels) can be manipulated at several frames per second, giving a very convincing 3D effect.

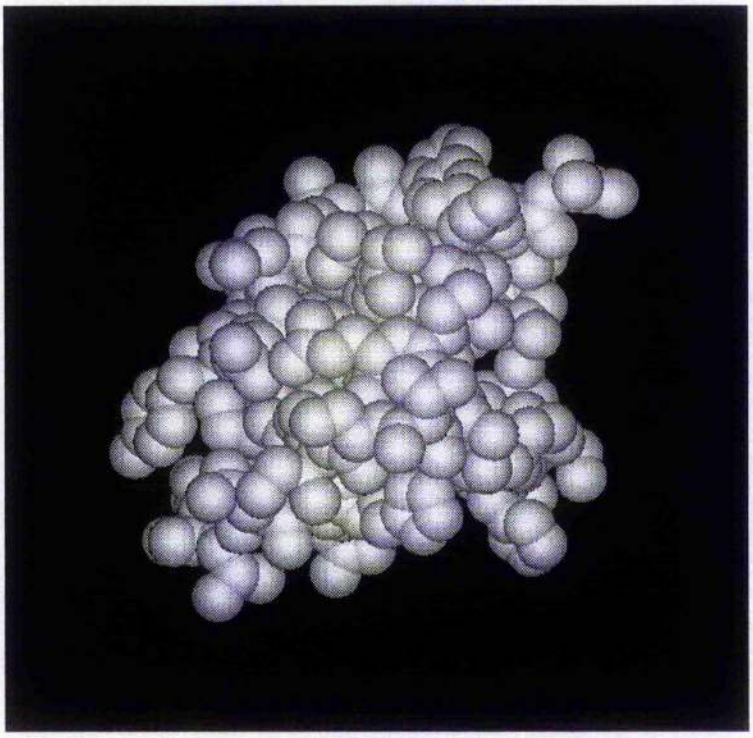

Figure 10.1: Ray-traced space-filling model

This is a space filling model of an insulin monomer, visualised using a ray-marker as detailed in appendix C. Colour and shadows have been excluded for better comparison with the next figure. Calculating an image in this way took around 100 seconds on a SPARC-10 based workstation.

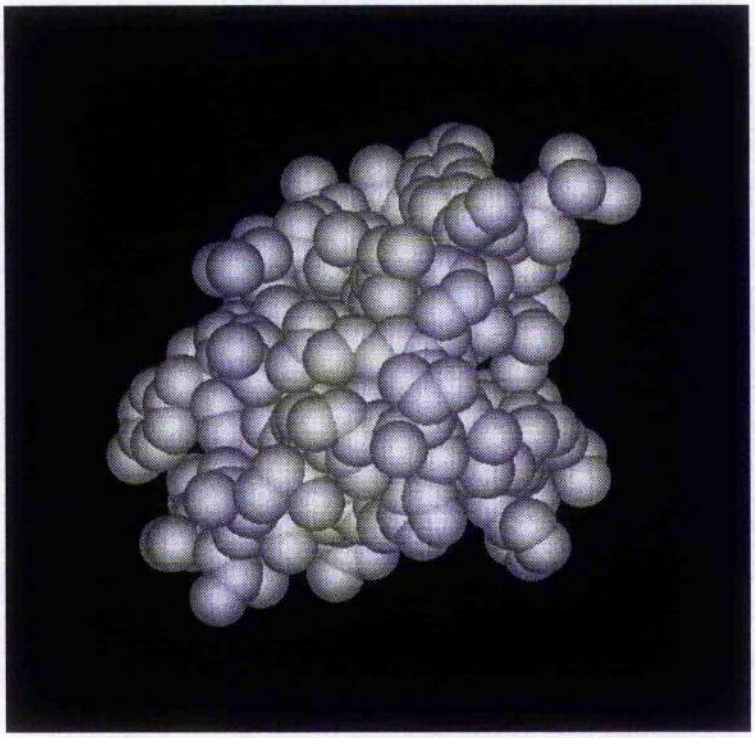

Figure 10.2: Rendered space-filling model

This is the same coordinate set as in the previous figure, this time with shading and depth calculated according to the corrected fast spheres algorithm of Patterson, as detailed in appendix D. The calculation of the image took under 1 second: it is also possible to include perspective and a movable light-source using this method, and under most circumstances the results are indistinguishable from simple ray-marked images.

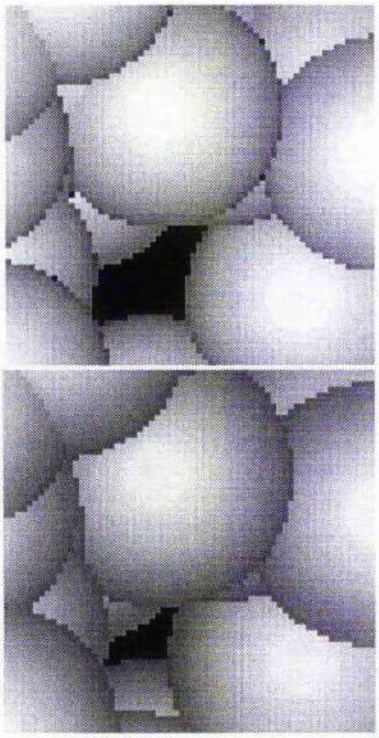

Figure 10.3: Differences between ray-marking and rendering
This figure shows details from the two images in figures 10.1 and 10.2. The original implementation of the fast spheres algorithm showed minor deviations in the intersections of spheres, which were tolerable in static images but led to visible "popping" in interactive visualisers or animations. Adding third and fourth order correction terms to the difference equations corrected these effects.

10.3 Conclusions. Compromising between software development and research

Having worked both as a structure investigator and a software engineer, and also having listened to some of the requirements of the diverse community of scientists interested in aspects of biomolecular structure, there are a set of common underlying problems which seem by their very nature intractable. Laboratory researchers have little time to learn complex visualisation systems, and certainly no time to program systems to produce the particular specialised representations they often require. Computational biochemists spend the bulk of their time re-implementing data structures for handling various molecular description formats and porting visualisation code to new graphics libraries or windowing systems, and keeping up with language and hardware developments. Computer scientists are happy to provide software tools which meet their understanding of the requirements of the biochemistry community, but find it hard to keep up with the changing specifications and requirements and tend to provide inflexible packaged solutions because the wider problem, what *in general* biomolecular software should provide, has not been well enough defined.

Figure 10.4 shows the range of problems visually. Problems for the development of an ideal molecular interpretation system come from three different areas: the complex and often contradictory nature of the researcher's requirements, the expanding data sets on which the software must now be able to work, and the fast changing nature of the computer technology which makes structural chemistry in its present form possible.

10.3.1 User requirements

One of the key problems in the development of molecular software is the wide range of different ways researchers have of referring to the same things. For example, a single atom in a protein may be referenced by its type, its mass, the side chain type it belongs to, the particular side chain it is part of, a hydrogen bond it participates in, its distance from some other non-bonded atom, its membership of some polar or hydrophobic subset of the mainchain or sidechain, or any one of a number of different definitions. One solution is to have a data structure which is arranged in the way a biochemist interprets the protein - as a set of backbone chains with sidechains, referenced by the amino acid residue they represent. But in fact this is just a shorthand - as earlier chapters have shown, it is often

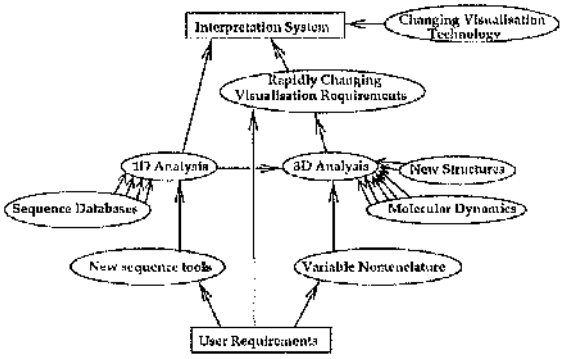

Figure 10.4: Problems for Biomolecular Structure Analysis
This diagram shows the range of problems which conspire to prevent a single elegant solution to the study of biomolecular structure being found. There are three main sources of change which make the ideal software a moving target: change in the requirements of the user both in terms of properties to study and in ways of referencing those properties, changing definitions of the amount of data which is of interest (always upwards), and changes in the hardware and software technology on which the tools must be implemented. In the text I suggest a solution of a system which is defined in the links between the pieces of software, not in the software itself.

valuable to cross the residue boundaries when defining an interaction: and in the work of Adzhubei and Sternberg [35], it was shown that taking the phi/psi values per *peptide* rather than per residue identified an important structural pattern much more clearly. Perhaps more useful would be a single convention on molecular data structures, based on IUPAC nomenclature, and then a set of tools to translate from the particular description being used to the underlying atoms which would then be used directly.

Programmers are also guilty of adding to the confusion, by imposing structures on data for computational convenience - for example, of the range of molecular description formats available, none have the flexibility to handle new molecular properties or hierarchical properties of sets of atoms such as hydrogen bonds without being redefined: none has a mechanism for nesting an object definition as an object itself, in other words. The biomolecular visualisation system VIEW developed by Larry Bergman [19] has within it a command scripting language which allows the user to define new objects to visualise based on molecular geometry, but the system is quite specialised, being purely directed towards interactive graphical manipulation, and the language used is a compromise between the underlying C implementation and a high level description. It shows what is possible

though, and suggests a rigid system which enforces a single nomenclature is not obligatory in robust software.

On the side of sequence analysis, there are a wide range of different techniques which would apparently need to be accommodated in any sufficiently flexible analysis package, such as conservation properties of loops, regions and single residues, intron boundaries, active site motifs and so forth. There are many many sequence analysis packages, implemented on a range of different machines, often laboratory Macintosh or PC systems, and it would be a mammoth task to implement even the most common algorithms as part of a new software system. However, all of these systems have one saving grace: their output is nearly always a linear stream of data. This linear data can be mapped on to the three dimensional chain in a simple mapping, and aside from a definition of the relevant ranges of the properties to be assigned, each different set of properties is essentially an identical scalar mapping. The problem of how to switch between 1D and 3D data then becomes a computer interaction design problem rather than an internal file or software structure problem – *but this assumes the underlying data structure is flexible enough to allow arbitrary properties to be assigned to residues,* otherwise the software would need to be rewritten each time a new analysis package was published.

10.3.2 Data Volume

The sheer volume of potentially relevant biochemical data is becoming a problem. There are three classes of data which are currently growing exponentially: sequence, structure, and simulation.

Sequence data is again the easiest for a visualisation system to handle, as there is a limit to how many sequences can meaningfully be displayed together at one time (around 10) and in any case most such displays are simply streams of characters, perhaps with connecting marks. If the results of a sequence analysis are to be combined with a 3D image, then it is likely to be only one or a few final linear data sets which are mapped onto the chain. The problem of keeping databases up to date over the Internet and analysing the relatedness of proteins is huge, but has only indirect consequences for the study of structure.

There is an increase in the number of protein structures published each year, and although the rate of growth is a little gentler than that for sequences it still leads to

administrative and searching headaches. Obviously it is important that any tools for searching for pattern in protein structure should be automated, but it should also be noted that 3D comparisons of structures can be very time consuming, and often slower when no matches can be found. For this reason, it would be useful to be able to take a set of relevant descriptors - a clearly defined set of properties such as alpha/beta content, size of molecule, refinement quality, and perhaps even the existence of certain structures such as loops or metal binding sites, and put them into a header on which fast searches could be carried out. Again it is no use having to reinstall software and databases whenever a new structure feature is defined, so new object descriptions would have to be part of the system's interpreter and also allowed in the files themselves. Searching for a given motif in...

```
copper binding proteins of less than 100 residues with a beta alpha beta
crossover of ten residues or fewer
```

...would then save a huge amount of computing power, and ease up on database access where this is done through a busy server or over the internet.

One final source of large volumes of data is from simulation experiments. As molecular dynamics becomes a common tool in structural analysis, new tools for identifying different conformers of a moving protein or for measuring changes in tension in loops through time will be developed. Any tool will need to be able to handle many thousands of related structures simultaneously, and abstract relevant information on differences and similarities from them for display. It cannot be overemphasised that the tools for this analysis are not yet complete: so to succeed in the future, molecular software must be re-programmable at short notice.

10.3.3 Interpretation of structure

The analysis of data within the software is hampered by another problem provided by the user of the system. Many structural chemists develop their own ways of representing interactions during their research (see for example chapter 8), and a considerable time is spent implementing these as part of an existing visualisation system or from scratch. It is also not uncommon to find that a version of an old visualisation method is needed - for example, a smoothed backbone or a ribbon representation - but it must be integrated into some new piece of software and so the algorithm must be rewritten from scratch. The

problem arises not just from the requirement for new visualisation techniques but from the range of ways they may need to be combined.

The technology of visualisation changes almost as rapidly as the requirements of the users. The simplest solution would be to find an accepted graphics standard and write software which links to a library of relevant graphics routines, but history shows that no "standard" in graphics description has remained the standard for more than 4 years (between 1985 and 1995 the *de facto* standards changed from PHIGS to Xlib to GL to OpenGL), so some caution needs to be adopted before accepting industry advice. Even the development and support of high level languages such as C and Fortran can not always be relied upon to follow a set course, and although both C and Fortran77 are both subsets of their next generations C++ and Fortran90 it is not clear that all hardware vendors are committed to supporting all types of compiler in the future.

Any solution to the display of molecular information must be above the graphics library level if it is to have any useful lifespan. Emerging standards for high level scene description such as OpenInventor and VRML (Virtual Reality Modelling Language) look promising – but notice that at the time of writing there are at least two such formats, related but not identical.

Finally, there is the problem of the interpretation system itself. In its simplest form this would simply be a set of tools for drawing lines or 3D primitives such as spheres and cylinders on a screen, but this is only a partial solution to a range of problems encountered by structural biochemists. While crystallographers are used to the constraints of working with a 3D graphics system, many researchers who need access to 3D data need it as part of a wider project and do not have time to come to grips with a fussy visualiser. A full solution would allow delicate interactive manipulation of a scene, but would also be capable of generating automated "executive summaries" of a protein or a complex for previewing. It is as important that quality hardcopy can be produced, and that diagrams can be automatically or at least easily annotated, as it is that the object on the screen is clear. In the near future the ability to output animations directly will become compulsory, and it may be that portable "snapshots", coordinates of one particular set of lines and objects, will need to be outputable for inclusion in hypertext documents.

10.3.4 Possible Solutions

There are a range of possible solutions, but all of them involve the same principle: the chemistry community must take a position on what is acceptable from molecular software, and should lay down a set of standards for commercial and academic software developers to use as targets.

In the short term this could mean the definition of basic data structures and at least one common nomenclature by which data should be accessible. Work on a common data interchange format and a set of structures for software to be based on will be valuable to this end, and it is to be hoped that the range of different approaches currently being followed can be resolved into a single policy.

In the longer term, however, I would recommend a more radical policy: it seems reasonable to believe that techniques currently under development will still be useful in 50 years' time, in the same way that the basic definitions of structure laid down by Pauling, Corey, Crick and Watson earlier this century are still widely referenced today. More useful than any amount of up-to-the minute software development would be a definition, in algorithmic terms but still largely comprehensible in a spoken language, of the basic structures and operations on those structures that a particular piece of software represents. In this work I hope to have shown how a specialised variant of the Pascal-based pseudo-code computer scientists use can be extended to give a clear and reproducible description of operations on three dimensional data sets, using reference systems (by atom, by residue, by temporary hybrid construct) which seem intuitive to a chemist.

It would not be impossible for the community to lay down a set of formal guidelines for "chemical pseudo-code" descriptions - and as the system became accepted, references to existing algorithms could be used to make complex definitions possible in a compact form. The system could be used in one of three ways: it could simply be a shorthand used to communicate code structure in scientific papers, it could be stored in a repository along with whatever implementations of the code existed in C, Fortran, or any other language which will be developed as a legible comment header, or it could be formalised into a compilable system as in the VIEW software. The key would be to make the naming of parts acceptable to a biochemist – with the ability to include Greek letters to make the printed form of the algorithm couched in the same terms as the atoms would be described

by in plain text.

A set of definitions would be laid down as a basis, and a central system for accrediting extensions would be added. Central to this effort would be the idea that it is algorithms and descriptions, not software and file formats, which will remain constant over the next 50 years of biomolecular research. In an ideal scenario, a file sent back from the future would contain within itself a description of the accumulated improvements and extensions to the description to allow the files to be converted into software: and if this is too unlikely, at least the converse, that any interpretation system will be able to handle any previously accepted algorithm and description, will be a minimum requirement of the files.

The advantage would be clear: the reinvention of the wheel at the start of every research project would effectively be eliminated. If the researcher could simply purchase a computer and a copy of the object display and data handling software, and then pick and choose which from the existing corpus of algorithms were needed, they could be compiled together into a single pice of tailored software. The system would grow in the way high level computer languages have grown, with occasional work on library development and clearing out old structures; but throughout, the agenda for data and algorithm management would be being set by the chemistry community rather than the computer scientists.

MODEL HIERARCHY FOR BIOPOLYMERS

molecule:

- **chain segment:**
  - **chain segment→Residue$_{min}$, Residue$_{max}$**
  - **residue:**
    - **atom:**
      - **atom→position**
      - **atom→{ Lennard-Jones parameters }**
      - **atom→partial charge**
    - **bond:**
      - **bond→atom$_1$, atom$_2$**
- **hydrogen bond list:**
  - **hydrogen bond:**
    - **hydrogen bond→type (mc/mc, mc/sc, sc/sc)**
    - **hydrogen bond→donor residue**
    - **hydrogen bond→acceptor residue**
    - **hydrogen bond→donor atom**
    - **hydrogen bond→acceptor atom**

GRAPHICS OBJECTS

line:

- **line→position$_{start}$, position$_{end}$**
- **line→type**

sphere:

- **sphere→centre**
- **sphere→type**

polygon:

- *...etc*

ribbon:

- **polygon list:**
  - *...etc*

surface model:

- **polygon list:**
  - *...etc*

Part III

Appendices

Appendix A

Energy calculations

A.1 Lennard-Jones parameters

A Lennard-Jones potential has the form

$$V_{ij} = 4\epsilon \left(\frac{\sigma^{12}}{r_{ij}^{12}} - \frac{\sigma^{6}}{r_{ij}^{6}} \right)$$

or

$$V_{ij} = 4\epsilon \left(\frac{\sigma^{9}}{r_{ij}^{9}} - \frac{\sigma^{6}}{r_{ij}^{6}} \right)$$

where ϵ is the depth of the potential well for the interaction between particles i and j, and σ is related to the position of the minimum of the interaction between the two particles in question. It is usual to treat the ϵ_{ij} and σ_{ij} terms for different atom types as simple products of the terms for interactions between atoms of the same type, and also to expand the equations to have a single parameter for each power term, giving

$$V_{ij} = \frac{A_{ii}^{1/2} A_{jj}^{1/2}}{r_{ij}^{9}} - \frac{C_{ii}^{1/2} C_{jj}^{1/2}}{r_{ij}^{6}}$$

for the 9-6 potential, where A_{ii} is $4\epsilon_{ii}.\sigma_{ii}^{9}$ and C_{ii} is $4\epsilon_{ii}.\sigma_{ii}^{6}$.

The parameters for these potential forms are chosen by refining the energy of a molecular model (for example, the coordinates of a known crystal), found by assuming a pairwise additive description of the energy, against some measurable property of the real system

		H^N	N^H	C^O	O^C	C^α	C^β	H^C
A	9-6-1	0.0	86.9 x10^3	12.5 x10^3	45.8 x10^3	38.9 x10^3	38.9 x10^3	445
C	9-6-1	0.0	2020	355	1410	1230	1230	15
q	9-6-1	+0.26	-0.26	+0.46	-0.46	0.00	-0.33	+0.11
A	12-6-1	0.0	2271 x10^3	3022 x10^3	275 x10^3	1981 x10^3	1811 x10^3	7150
C	12-6-1	0.0	1230	1340	502	1125	532	32.9
q	12-6-1	+0.28	-0.28	+0.38	-0.38	0.00	-0.30	+0.10
q	cm ab initio	+0.28	-0.50	+0.38	-0.38	+0.22	-0.30	+0.10

Table A.1: Lennard-Jones parameters and peptide partial charges used in chapters 2 to 6.

being modelled. In the work of Lifson, Hagler, Dauber *et. al.* [23], this was done by matching a set of parameters to the observed properties of a set of crystals of small organic molecules, assuming that the observed crystal forms for each molecule were fixed (ie ignoring bond stretching or bending) and attempting to obtain a single set of parameters which would be transferable between molecular crystals. This work was fairly successful, and has been used in one form or another as the basis for many of the biomolecular force fields in use today.

A.2 Partial charges

Table A.2 gives the partial charges and coordinates used for the calculations in chapter 4. The partial charges were estimated from ab initio calculations using a 6-31 G* basis set on the isolated sidechains in vacuo, as detailed in the original AMBER force field [6].

A.3 Modelling polypeptide backbones

Modelling polypeptide backbones from scratch is a useful first stage in full modelling of proteins ab initio. Many useful features can be found in simple polyglycine or polyalanine models which do not require elaborate energy minimisation or conformational space searching techniques, and in most cases these are fastest if implemented in a medium level computer language such as C or Fortran and recompiled for each case being studied. Other packages specifically for molecular modelling tend to be angled towards optimising a single model with a specified starting point, while specialist packages for algorithm im-

residue	atom	x coord	y coord	q
Arg	C^{μ}	0.0	0.0	0.813
	N^{ϵ}	-1.3	0.0	-0.493
	$N^{\eta 1}$	0.65	1.13	-0.634
	$N^{\eta 2}$	0.65	-1.13	-0.634
	$H(N^{\epsilon})$	-1.70	0.86	0.294
	$H^{syn}(N^{\eta 1})$	1.65	1.13	0.362
	$H^{anti}(N^{\eta 1})$	0.15	1.99	0.362
	$H^{syn}(N^{\eta 2})$	1.65	-1.13	0.362
	$H^{anti}(N^{\eta 2})$	0.15	-1.99	0.362
Asn/Gln	$H^{anti}(N^{\delta})/H(N^{\epsilon})$	-0.79	1.31	0.344
	$H^{syn}(N^{\delta})/H(N^{\epsilon})$	0.79	1.31	0.344
	N^{δ}/N^{ϵ}	0.0	0.7	-0.867
	C^{γ}/C^{δ}	0.0	-0.77	0.675
	O^{δ}/O^{ϵ}	1.19	-1.46	-0.470
	C^{β}/C^{γ}	-1.19	-1.46	0.00
Asp/Glu	C^{γ}/C^{δ}	0.0	0.0	0.620
	$O^{\delta 1}/O^{\epsilon 1}$	1.09	0.59	-0.706
	$O^{\delta 2}/O^{\epsilon 2}$	-1.09	0.59	-0.706
	C^{β}/C^{γ}	0.0	1.52	-0.208
His^{+}	C^{ϵ}	0.0	0.0	0.719
	N^{ϵ}	1.12	-0.81	-0.613
	C^{δ}	0.69	-2.12	0.103
	C^{γ}	-0.69	-2.12	0.353
	N^{δ}	-1.12	-0.81	-0.686
	$H(N^{\epsilon})$	2.01	-0.35	0.478
	$H(N^{\delta})$	-2.01	-0.35	0.486
His	C^{ϵ}	0.0	0.0	0.384
	N^{ϵ}	1.12	-0.81	-0.527
	C^{δ}	0.69	-2.12	0.122
	C^{γ}	-0.69	-2.12	0.122
	N^{δ}	-1.12	-0.81	-0.444
	$H(N^{\delta})$	-2.01	-0.35	0.320
Peptide	C^{α}	1.782	-1.3	0.000
	Hn	0.5	-3.04	0.260
	Nh	0.5	-2.04	-0.260
	Co	-0.66	-1.37	0.460
	Oc	-0.66	0.0	-0.460
	C^{α}	-1.84	-2.11	0.000
Ser/Thr	$H(O^{\gamma})$	1.0	0.0	0.310
	O^{γ}	0.0	0.0	-0.550
	C^{β}	-0.39	1.36	0.194

Table A.2: Partial Charges and Coordinates for hydrogen bonding groups in chapter 4

plementation such as mathematica have their own syntax and modus operandi which are unfamiliar to most computational chemists.

The backbone $\{N, C^{\alpha}, C\}_N$ is defined by a set of N backbone bond angles, $\{ \alpha, \beta, \gamma \}$, and a set of bond torsion angles $\{\phi, \psi, \omega\}$ where ω is usually within a few degrees of 0° or occasionally 180° (for *cis* peptide bonds). Producing these backbone atoms is then simply a case of proceeding down the chain, defining a local coordinate system based on the plane of the three preceding atoms, and using the known bond length, bond angle, and torsion angle to place the next atom: three atoms must be placed initially.

The atoms which do not define the chain, $\{H, O\}_N$, can be placed in the same way, by defining one torsion angle, one bond angle, and one bond length. There is also a simpler way, which is to assume that the bond is in the plane of the nearest three atoms ($\{C_{i-1}, N_i, C_i^{\alpha}\}$ for hydrogen, $\{C_i^{\alpha}, C_i, N_{i+1}\}$ for carbonyl oxygen, and that the bond bisects the angle between those three atoms. Only a bond length is then supplied: this is the technique used for hydrogen placement in any case, as hydrogen coordinates are not found by crystallography.

There is a simple test for a correct implementation of this algorithm: if the $\{\alpha, \beta, \gamma : \phi, \psi, \omega\}$ and bond distance values for a real chain are used, with the first three real coordinates as starting condition, the original coordinates must be reproduced to within machine precision. (Note that this will only be the case for all atoms if the $NC^{\alpha}CO$ torsion, $C^{\alpha}CO$ bond angle, and CO bond length are taken from the real coordinates, not using the heuristic placement given here.)

The general system for generating a backbone atom position based on three previously given positions is given here. A coordinate system relative to the plane of the three atoms is defined: the projection of the new bond along the I direction is simply $\cos(\gamma)$, and the projection in the J,K directions is $\sin(\gamma)$. The J and K components of this are then functions of $\sin(\phi)$ and $\cos(\phi)$ respectively.

Finding the other atomic positions simply involves the same procedure with a new set of angles and the relevant 3 atoms from the chain as it is generated.

BACKBONE GENERATION

$N_1=\{\ 0,\ 0,\ 0\ \}$

$C_\uparrow^\alpha=N_1 + r_{NC^\alpha}.i$

$C_1=C_\uparrow^\alpha + r_{C^\alpha C}.\cos\beta.i + r_{C^\alpha C}.\sin\beta.j$

for each residue m>1 $\in$ chain:

$I=C^\alpha C/|C^\alpha C|$

$J=I\times C^\alpha N/|C^\alpha N|$

$K=-I\times J$

$R=\cos\gamma.I + \sin\phi.\sin\gamma.J + \cos\phi.\sin\gamma.K$

$N=C + r_{CN}.R$

$I=CN/|CN|$

$J=I\times CC^\alpha/|CC^\alpha|$

$K=-I\times J$

$R=\cos\alpha.I + \sin\omega.\sin\alpha.J + \cos\omega.\sin\alpha.K$

$C^\alpha=N + r_{NC^\alpha}.R$

$I=NC^\alpha/|NC^\alpha|$

$J=I\times NC/|NC|$

$K=-I\times J$

$R=\cos\beta.I + \sin\psi.\sin\beta.J + \cos\psi.\sin\beta.K$

$C=C^\alpha + r_{C^\alpha C}.R$

$O=f_{bisect}(C^\alpha,\ C,\ N,\ r_{CO})$

$H=f_{bisect}(C,\ N,\ C^\alpha,\ r_{NH})$

$f_{bisect}(A,\ B,\ C,\ r)$:

$D=0.5(A+C)$

$E=B + r.DB/|DB|$

return{E}

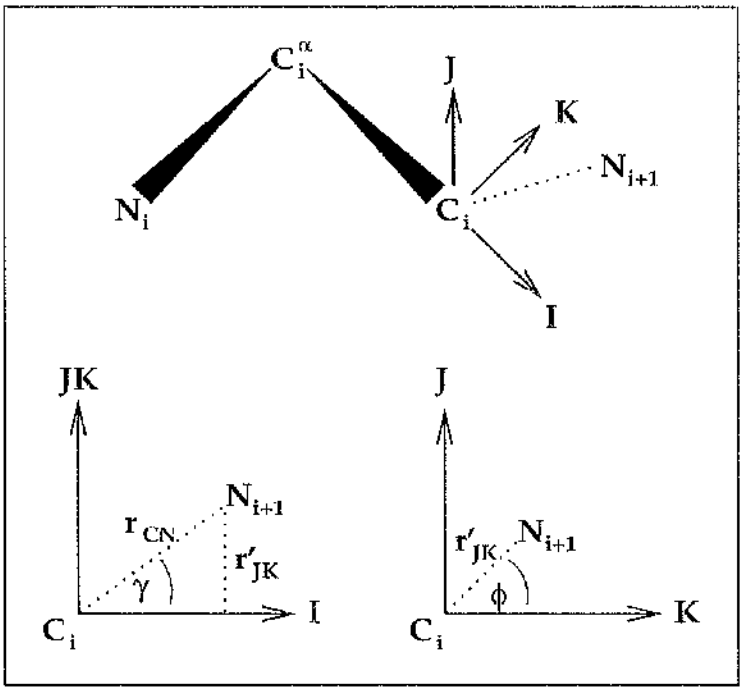

Figure A.1: Generating atomic coordinates from backbone angles

The general system for generating a backbone atom position based on three previously given positions is given here. A coordinate system relative to the plane of the three atoms is defined: the projection of the new bond along the **I** *direction is simply cosγ, and the projection in the* **J**.**K** *directions is sinγ. The J and K components of this are then functions of sinϕ and cosϕ respectively. Finding the other atomic positions simply involves the same procedure with a new set of angles and the relevant 3 atoms from the chain as it is generated.*

Appendix B

Hydrogen Bond Data

B.1 Generating and maintaining sets of hydrogen bonds

In the following pseudo-code showing how the hydrogen bond lists were calculated, a "donor" or "acceptor" means any user-defined hydrogen bond donor or acceptor group - a pair of atoms in the Baker and Hubbard definition, a more complicated structure in the cases where extra features such as out of plane or in plane angles are required (for example in chapter 4 where a review of angles relating to pairs of interacting main chain peptides was required).

The user-supplied function $f_{hbond}()$ is likewise considered to be flexible,returning a final object which is the set of properties associated with a hydrogen bond definition - whether it is mainchain(mc) to sidechain(sc), mc to mc, or sc to sc, the bond lengths and angles which can make up its definition, and so on.

HYDROGEN BOND TABLE GENERATION

for each amino acid i∈chain:
  for each donor d_i:
    for each amino acid j∈chain:
      for each acceptor a_j:
        where j≠i:
          hydrogen bond h_n=$f_{hbond}(d_i,a_j)$
          n=n+1
        where j=i
        and either (d_i →type=mc and a_j →type=sc)
        or (d_i →type=sc and a_j →type=mc):
          hydrogen bond h_n=$f_{hbond}(d_i,a_j)$
          n=n+1

$f_{hbond}(d_i,a_j)$:

atom D=d_i →D, H=d_i →H, A=a_j →A, X=a_j →X
angle α=∠D,H,A
angle β=∠H,A,X
distance r_{HA}=|H,A|
  where α <60°
  and β <90°
  and r_{HA} <minr(A→atomtype)
    return { h_n } (={i, j, d_i →type, a_j →type, D, A, α, β, r_{HA}})

B.2 Searching for regular patterns in hydrogen bond lists

This is the algorithm implemented to carry out the searches of the hydrogen bond lists. Note that as it is written here and implemented for this works it has an $O(N^n)$ execution time, where n is the number of hydrogen bonds defined in the search pattern. This was not a problem for the cases here on proteins of size up to 600 residues, but could be significant for large database query programs. There are several schemes possible to avoid this, with an approach based on knowledge of the evaluation order of the hydrogen bond list and

keeping searches for derived bonds within the limits specifies by the values of x and y would reduce this to approximately $O(NlogN)$.

HYDROGEN BOND SEARCHES

```
given search pattern S={root, derived1, derived2, ...}
for h_i∈hydrogen bond list:
    if match(h_i, S→root):
        N_matches=1
        for S_h∈{S→derived1, S→derived2, ...}:
            for h_j∈hydrogen bond list:
                if match(h_j, S_h):
                    N_matches=N_matches+1
                    next h_j
        if N_matches=number in S:
            success
```

Appendix C

Raymarked Images

C.1 A simple, realistic, image generator

Molecules have been modelled as balls connected as sticks for as long as structural chemistry has been a science. With the advent of computers capable of realistic computer graphics, representations of molecules as ball-and-stick have been common -- and in fact, the principles are very similar for some types of cartoon representation and for space filling models. Basically, all that is required is a way of drawing spheres and cylinders, shaded for the imagined lighting conditions and with some way of detecting and handling intersections between objects.

Fast – particularly real time interactive – systems rely on being able to render flat polygons, so models have to be built up from many faceted approximations to spheres and cylinders. For most applications this is acceptable (indeed, for many purposes of study, a simple model built of lines and circles is adequate if combined with motion or suitable 3D visualisation equipment). There are some cases where it falls down - particularly around the intersections of sticks with other sticks, and when the scene becomes enlarged to focus on a single detail.

To produce a mathematically correct model of a typical scene is actually fairly easy, and a simple technique for doing this is given here. With a full description of the surfaces of the spheres and cylinders it is possible to use a technique such as ray-tracing to produce a model of anything up to photo-realistic quality. Here ray-tracing with reflectivity and transparency set to zero is described, technically "ray marking".

Figure C.1 shows how the calculation proceeds for spheres. Figure C.2 shows the slightly more complicated situation for cylinders.

RAY-MARKING SPHERES

```
given lighting vector l
given ambient lighting intensity a
given diffuse lighting intensity range R.
for each pixel pix_xy ∈view screen:
    for each atom A∈visible atoms:
        if x and y∈A.bounding box:
            r²_xy=(A_x-u)² + (A_y-y)²
            r_A=A→radius
            if r²_xy<r_A:
                z=A_z + (r²_A - r²_xy)^1/2
                P={x, y, z}
                N={A_x, A_y, A_z}
                U=PN/r_A
                if z>zbuffer_xy:
                    pix_xy →intensity=max(U.l, 0).R + a
```

RAYMARKING CYLINDERS

given lighting vector I
given ambient lighting intensity a
given diffuse lighting intensity range R
for each pixel $\mathbf{pix}_{xy} \in$**view screen:**
  for each bond AB where A or B$\in$**visible atoms:**
    if x and y$\in$**AB.bounding box:**

$$\mathbf{r}_1^2=(\mathbf{A}_x\text{-}\mathbf{x})^2 + (\mathbf{A}_y\text{-}\mathbf{y})^2$$
$$\mathbf{r}_2^2=(\mathbf{B}_x\text{-}\mathbf{x})^2 + (\mathbf{B}_y\text{-}\mathbf{y})^2$$
$$\mathbf{r}_3^2=(\mathbf{A}_x\text{-}\mathbf{B}_x)^2 + (\mathbf{A}_y\text{-}\mathbf{B}_y)^2$$
$$\mathbf{r}_3=(\mathbf{r}_3^2)^{1/2}$$
$$\mathbf{a}=(\mathbf{r}_1^2+\mathbf{r}_3^2\text{-}\mathbf{r}_2^2)/2\mathbf{r}_3$$
$$\mathbf{C}=\mathbf{A} + \mathbf{AB}.\mathbf{a}/\mathbf{r}_3$$

$$\mathbf{r}_{xy}^2=(\mathbf{Cx}\text{-}\mathbf{x})^2 + (\mathbf{Cy}\text{-}\mathbf{y})^2$$
$$\mathbf{r'}_{xy}^2= (\mathbf{1}\text{-}(\mathbf{r}_{xy}/\mathbf{r}_C^2)).(\mathbf{r}_C.\mathbf{AB}_z/|\mathbf{AB}|)^2$$
$$\mathbf{r''}_{xy}^2=\mathbf{r}_{xy}^2 + \mathbf{r'}_{xy}^2$$
$$\mathbf{C'}=\mathbf{C}+\mathbf{AB}(\mathbf{r'}_{xy}^2)^{1/2}/\mathbf{r}_3$$

$$\mathbf{z}=\mathbf{C}_z' + (\mathbf{r}_C\text{-}\mathbf{r''}_{xy})^{1/2}$$
$$\mathbf{P}=\{\mathbf{x},\ \mathbf{y},\ \mathbf{z}\}$$
$$\mathbf{N}=\{\mathbf{C}_x,\ \mathbf{C}_y,\ \mathbf{C}_z\}$$
$$\mathbf{U}=\mathbf{PN}/\mathbf{r}_C$$

    if z>zbuffer$_{xy}$**:**
      $\mathbf{pix}_{xy}$ **→intensity=max(U.I, 0).R + a**

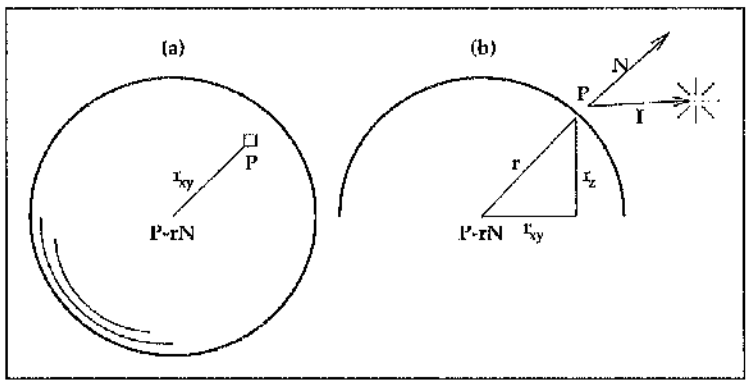

Figure C.1: Raymarking spheres

Calculating the shading for a given pixel proceeds by finding the distance of the pixel from the sphere centre in the xy plane, then, knowing the radius of the sphere, calculating the z coordinate of the sphere's intersection with the line of sight, $\mathbf{P}$. *The normal at that point is simply the reverse of the vector joining P to the atom centre, and the illumination of the pixel can then be calculated as a function of the angle between illumination vector and surface normal. (a) shows the relevant paramters and positions as seen from the viewpoint. (b) shows the scene perpendicular to the view direction.*

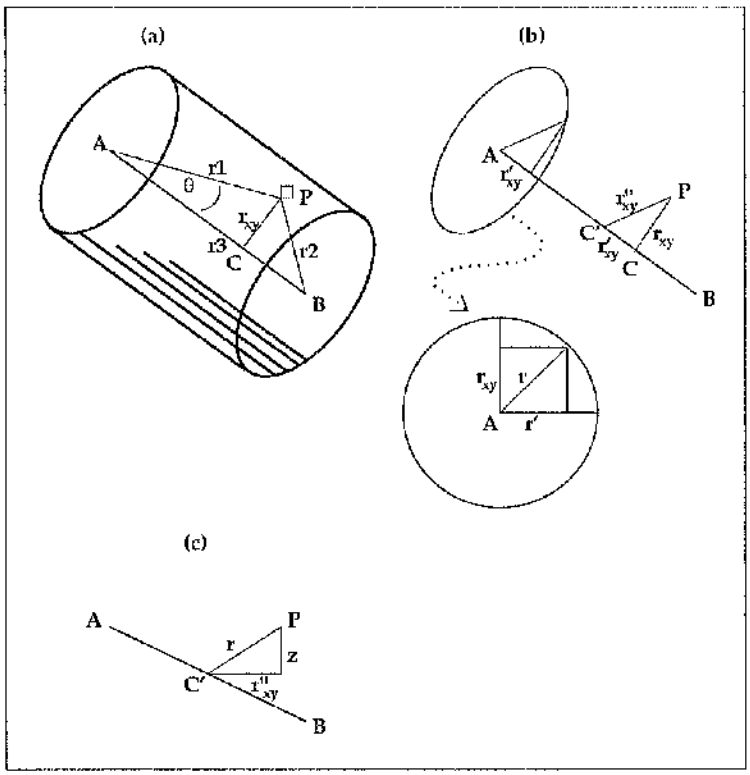

Figure C.2: Raymarking cylinders

Cylinders from more of a challenge than spheres: the problem lies in working out the position of the nearest point on the cylinder axis to the line of sight intersection P, before the normal and z value can be calculated.

First the nearest point on the cylinder axis in the xy plane is found (a). Then an offset along the axis to the real intersection is found, by recognising that the distance **r'** *is simply the projection of the coordinate perpendicular to PC on the cylinder base, which can be found by pythagoras.*

Once the nearset axial position has been found, the problem reduces to an identical calculation to that on the sphere. An extra square root calculation has been introduced.

Appendix D

Fast Spheres

D.1 A very fast space-filling model rendering algorithm

Given the importance of spheres to molecular graphics, a good deal of attention has been given to speeding up their display. Appendix C shows that there is an unavoidable and computationally costly square root calculation in the evaluation of the z-value, the axis perpendicular to the screen, for each pixel the sphere covers.

Typical methods for speeding up sphere display have relied on some form of pre-calculation: approaches based on Blitting pre-calculated spheres, fixing the lighsource behind the viewer and drawing circles of constant colour, or using a lookup table for the z values have been applied with some degree of success. However, none is truly flexible under a range of perspective and lighting conditions. An algorithm which reduces the calculation to integer arithmetic akin to the scan line polygon display techniques used for most interactive displays would have obvious appeal. John Patterson's Fast Spheres [52] is just such an algorithm.

This method generates sphere-like surfaces, using a parabolic approximation delimited by Bresenham's circle-generation algorithm. The original version used first and second order differencing, and calculated values for a z-buffer at the same time. Implementation allowed interactive manipulation of space filling models at around 4 frames/second for a 200 atom model in a 400^2pixel window, even on machines without any hardware graphics support other than 2D pixel blitting and using an X windows interface which involved an extra memory copy for the whole scene. 10 frames per second were actually being

generated, a fairly respectable animation rate; the bottleneck was all in the display system.

Use of the algorithm interactively highlighted one problem: the parabolic approximations did not intersect evenly, leading to a visual popping effect when the object was moved. Corrections to the geometry were introduced by adding third and fourth order difference terms to the z-calculation. This resulted in sphere approximations within 1 pixel of the values for a real sphere.

Since the perspective transformation of a sphere is roughly spherical, it is possible to extend the technique into real 3D by scaling the sphere according to the perspective transformation. Fast spheres can then form the basis of an interactive visualiser with perspective and possibly stereo views generated by the researcher as he/she works.

For a sphere, the lighting is given by a diffuse term based on the angle between light source and surface normal plus an ambient component.

$$colour = k_D N.L + k_A$$

$$N.L = 1/r(x.l_x + y.l_y + z.l_z)$$

The z-value should be given by

$$z = -(r^2 - x^2 - y^2)^{1/2}$$

but for this fast approximation, this term is replaced by:

$$z = (r^2 - x^2 - y^2)/r$$

the parabolic approximation, which removes the need for square root evaluation. Even a divide operation normally required is removed and replaced with a pre-multiplication and a divide by a power of 2, both of which are considerably faster.

Figure D.1 shows some of the order of evaluation of pixels. There are a large number of related initialisation parameters, which are tabulated in the original reference.

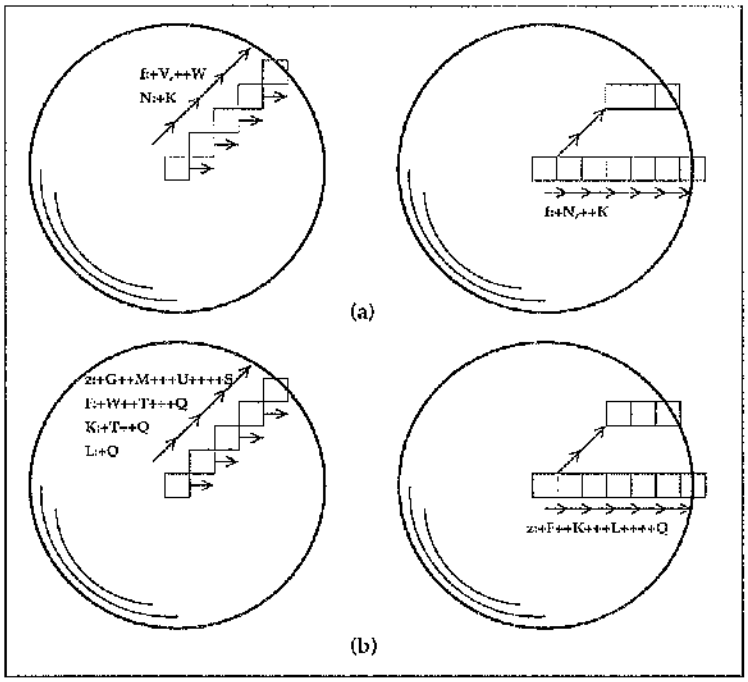

Figure D.1: Fast spheres

Calculating the lighting and z value functions for spheres by Patterson's difference method requires two types of operation, one a scan-line difference for each line of pixels between a diagonal and the circle edge and the other a propagation of the function and the difference terms up the diagonal. Only one octant needs to be considered for the case where the light-source is behind the viewer.

The two significant terms are f(x,y), the colour intensity propagation term and z(x,y), the z buffer propagation term, and the difference terms associated with each. In the figure "+V++W" means "+V(x,y),V(x+1,y+1)=V(x,y)+W": so all terms except K, W, Q, and S are functions of x and y.

Bibliography

[1] L.Pauling *The Nature of the Chemical Bond*

[2] L.Pauling, R.B Corey (1953) Compund Helical Configurations of polypeptide chains: structure of proteins of the α-keratin type. *Nature* 171, 59-61.

[3] J.S.Richardson (1981) Protein Anatomy *Advan.Protein Chem.* 34, 167-339

[4] J.Hermans, J.C.Berendsen, W.F.van Gunsteren, J.P.M.Postman (1984) GROMOS *Biopolymers* 23, 1513-1575

[5] P.Dauber-Osguthorpe, V.A.Roberts, D.J.Osguthorpe, J.Wolff, M.Genest, A.T.Hagler (1988) Structure and energetics of ligand binding to proteins. *Proteins:Struct.Fun.Gen* 4, 31-47.

[6] S.J.Weiner, P.A.Kollman, D.Case, U.L.Singh, C.Chio, G.Alagona, P.S.Profeta, P.Weiner (1984) AMBER: A new force field for molecular mechanical simulation of nucleic acids and proteins. *J.Am.Chem.Soc* 106, 765-784

[7] B.R.Brooks, R.E.Bruccoleri, B.D.Olafson, D.J.States, S.Swaminathan, M.Karplus (1983) CHARMm: a program for macromolecular energy minimisation and dynamics calculations. *J.Comput.Chem.* 4, 187-217

[8] E.N.Baker, R.E.Hubbard (1984) Hydrogen bonding in globular proteins. *Prog.Biophys.molec.Biol.* 44, 97-179

[9] D.H.Williams (1991) The molecular basis of biological order. *Aldrichimica Acta* 24, 71-80

[10] K.Belhadj-Mostefa, R.Poet, E.J.Milner-White (1991) Displaying inter-main-chain hydrogen bond patterns in proteins. *J.Mol.Graph.* 9, 194-197

[11] E.J.Milner-White (1988) Recurring loop motif in proteins that occurs in right-handed and left-handed forms. *J.Mol.Biol.* 199, 503-511

[12] F.C.Bernstein, T.F.Koetzle, G.J.B.Williams, E.F.Meyer, M.D.Brice, J.R.Rogers, O.Kennard, T.Shimanouchi, M.Tasumi (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112, 535-542.

[13] P.H.Maccallum, R.Poet, E.J.Milner-White (1995) Coulombic Interactions between partially charged main-chain atoms stabilise the right-handed twist found in most β-strands. *J.Mol.Biol.* 248, 374-384

[14] P.H.Maccallum, R.Poet, E.J.Milner-White (1995) Coulombic Interactions between partially charged main-chain atoms not hydrogen-bonded to each other influence the conformations of α helices and antiparallel β-sheet. *J.Mol.Biol.* 248, 361-373

[15] J.Singh, J.M.Thornton (1992) Atlas of protein side chain interactions, Vols I and II. *IRL Press, Oxford*

[16] J.A.Ippolito, R.S.Alexander, D.W.Christianson (1990) Hydrogen bond stereochemistry in protein structure and function. *J.Mol.Biol.* 215, 457-471

[17] J-Y.Le Questel, D.G.Morris, P.H.Maccallum, R.Poet, E.J.Milner-White (1993) Common ring motif in proteins involving asparagine or glutamine amide groups hydrogen-bonded to main chain atoms. *J.Mol.Biol.* 231, 888-896

[18] L.D.Bergman, J.S.Richardson, D.C.Richardson (1995) An algorithm for smoothly tesselating β-sheet structures in proteins. *J.Molec.Graph.* 13, 36-45

[19] L.D.Bergman, J.S.Richardson, D.C.Richardson, F.P.Brooks (1993) VIEW - an exploratory molecular visualisation system with user-definable interaction sequences. *Comput.Graphics* 27(4) 117-126

[20] W.Kabsch, C.Sander (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers* 22, 2577-2637

[21] A.T.Hagler, S.Lifson (1974) Energy functions for polypeptides and proteins. I. The amide hydrogen bond and calculations of amide crystal properties. *J.Am.Chem.Soc* 96, 5327-5335

[22] A.T.Hagler, S.Lifson, P.Dauber (1979) CFF studies of intermolecular forces in hydrogen bonded crystals II. A benchmark for the objective comparison of alternative force fields. *J.Am.Chem.Soc* 101, 5122-5130

[23] S.Lifson, A.T.Hagler, P.Dauber (1979) Consistent force field studies of intermolecular forces in hydrogen bonded crystals I. *J.Am.Chem.Soc.* 101, 5111-5121

[24] G.N.Ramachandran, V.Sasiekharan, C.Ramakrishnan (1963) Stereochemistry of polypeptide chain configurations. *J.Mol.Biol.* 7, 95-99

[25] G.N.Ramachandran, V.Sasiekharan, C.Ramakrishnan (1966) Molecular structure of polyglycine II. *Biochim.Biophys.Acta* 112, 168-170

[26] E.N.Baker, R.E.Hubbard (1984) Hydrogen bonding in globular proteins. *Prog.Biophys.molec.Biol.* 44, 97-179

[27] C.Toniolo, E.Benedetti (1991) The fully extended polypeptide conformation, in *Molecular conformation and biological interactions*, Indian Academy of Sciences, Bangalore.

[28] E.J.Milner-White, R.Poet (1986) Four classes of beta-hairpins in proteins. *Biochem.J.* 240, 289-292

[29] E.J.Milner-White, Ron Poet (1987) Loops, bulges, turns and hairpins in proteins. *TIBS* 189-192

[30] E.J.Milner-White, B.M.Ross, R.Ismail, K.Belhadj-Mostefeda, R.Poet (1988) One type of gamma turn, rather than the other gives rise to chain-reversal in proteins. *J.Mol.Biol.* 204, 777-782

[31] E.James Milner-White (1990) Situations of gamma-turns in proteins. *J.Mol.Biol.* 216, 385-397

[32] C.Chothia (1973) Conformations of β-pleated sheets in proteins. *J.Mol.Biol.* 75, 295-302

[33] K-C.Chou, M.Pottle, G.Nemethy, Y.Ueda, H.A.Scheraga (1982) Structure of beta sheets: Origin of the right-handed twist and of the increased stability of antiparallel over parallel sheets. *J.Mol.Biol.* 162, 89-112

[34] K.C.Chou, G.Nemethy, H.A.Scheraga (1990) Energetics of interactions of regular structural elements in proteins. *Acc.Chem.Res* 23, 134-141

[35] A.A.Adzhubei, M.J.E.Sternberg (1993) Left-handed polyproline II helices commonly occur in globular proteins. *J.Mol.Biol.* 229, 472-493

[36] O.Schueler, H.Margalit (1995) Conservation of salt bridges in protein families. *J.Mol.Biol.* 248, 125-135

[37] J.Singh, J.M.Thornton (1992) Atlas of protein side chain interactions, Vols I and II. *IRL Press, Oxford*

[38] P.Artimiuk, C.C.F.Blake (1981) Refinement of human lysozyme at 1.5 A. Analysis of non-bonded and hydrogen bond interactions. *J.Mol.Biol.* 152, 737-762

[39] J.S.Richardson, D.C.Richardson (1988) Amino acid preferences for specific locations at the ends of alpha helices. *Science* 240, 1648-1652

[40] L.G.Presta, G.D.Rose (1988) Helix signals in proteins. *Science* 240, 1632-1641

[41] D.W.Heinz, W.A.Baase, F.W.Dahlquist, B.W.Matthews (1993) How amino acid insertions are allowed in an alpha helix of T4-lysozyme. *Nature* (361, 561-564

[42] M.Sundaralingam, Y.C.Sekharudu (1989) Water-inserted alpha-helical segments implicate reverse turns as folding intermediates. *Science* 244, 1333-1337

[43] M.Bycroft, A.Matouschek, J.T.Kellis Jr, L.Serrano, A.R.Fersht (1990) Detection and Characterisation of a folding intermediate in barnase by NMR. *Nature* 346, 488-490

[44] A.Matouschek, J.T.Kellis Jr, L.Serrano, M.Bycroft, A.R.Fersht (1990) Transient folding intermediates characterised by protein engineering. *Nature* 346, 440-445

[45] C.Sander, R.Schneider (1991) Database of homology-derived protein structures. *Proteins:Struct.Fun.Gen.* 9, 56-68

[46] L.J.Stern, J.H.Brown, T.S.Jardetzky, J.C.Gorga, R.G.Urban, J.L.Strominger, D.C.Wiley (1994) Crystal structure of the human class II MHC protein HLA-DR1 complexed with an influenza virus peptide. *Nature* 368, 215-221