# Detecting configurational motifs in protein structures

Submitted in accordance with the requirements for the degree of

Master of Science in Statistics

The University of Leeds, School of Mathematics

September 2009

The candidate confirms that the work submitted is his/her own and that appropriate credit has been given where reference has been made to the work of others.

# Abstract

The theory that the three-dimensional structure of a protein is completely determined by it's amino acid sequence has been known for the past 50 years. In this time several approaches to mapping structure from sequence have had some success, most notably when using data-based techniques. However, the accuracy of results from protein structure prediction experiments remain limited. Currently, homologous proteins play a leading role in protein structure prediction but this approach is likely to be overturned by the newly formed conjecture that protein folds are constrained in local sequence-structure space. Using the three-dimensional coordinates of all heavy atoms from 30 protein structures within the PDB, this project aims to detect configurational motifs composed of parts of amino acid residues, which are spatially compact but distant in sequence.

A database of compact structural configurations is compiled using the statistical package R. These atomic configurations are investigated by performing hierarchical agglomerative clustering using both complete and Ward's linkage. The results are masked by the overwhelming influence of a peptide bond, resulting in a constraint, stating that configurations must contain atoms from residues distant in sequence, being applied. The results from cluster analysis on the new dataset are much more promising, with two configurational motifs being identified. The first of these motifs is explained by two distant cysteine residues forming a disulphide bond, which brings the residues into spatial proximity. The second motif is accounted for by the presence of salt bridges within the protein structures. Multiple clusters are shown to contain this motif, which is shown to have an equal shape for all clusters the motif appears in.

Although these results are already known in the field of bioinformatics, the conclusions drawn here confirm the validity and effectiveness of the methodology. Further work is suggested to build on the approach used here, identifying scenarios in which the methodology requires adaptation. It is hoped that by implementing these changes, systematic patterns representing new discoveries will become visible. These motifs could then be used as 'building-blocks' to predict the structure of a given protein.

# Acknowledgements

I would like to express my gratitude towards my supervisor [redacted], for both his guidance and enthusiasm over the past three months. I am very grateful for the discussions with [redacted], who I am sincerely thankful to, not just for this project, but throughout the whole of my time at the University of Leeds. I would also like to thank [redacted] for his advice and expertise in shape analysis. Finally, I would like to thank [redacted], [redacted] and [redacted] all of whom have supported and encouraged me during the course of this dissertation.

# Contents

# List of Figures

# List of Tables

# Declaration

No portion of the work referred to in this dissertation has been submitted in support of an application for another degree or qualification of this or any other university or institution of learning.

In the attached submission I have not presented anyone else's work as my own. Where I have taken advantage of the work of others, I have given full acknowledgment. I have read and understood the University's published rules on plagiarism and also any more detailed rules specified at School or module level. I know that if I commit plagiarism I can be expelled from the University and that it is my responsibility to be aware of the University's regulations on plagiarism and their importance.

I re-confirm my consent to the University copying and distributing any or all of my work in any form and using third parties (who may be based outside the EU/EEA) to verify whether my work contains plagiarised material, and for quality assurance purposes.

I confirm that I have declared all mitigating circumstances that may be relevant to the assessment of this piece of work and that I wish to have taken into account. I am aware of the School's policy on mitigation and procedures for the submission of statements and evidence of mitigation.

Signed: _____

# Chapter 1

# Introduction

## 1.1 Overview

The so-called 'second central dogma' of molecular biology states that the three-dimensional structure of a protein is completely determined by it's amino acid sequence (Fasnacht, 2002). This hypothesis originated from experiments carried out by Christian Anfinsen back in the 1960s, who showed that protein structure is dictated by both its sequence and its environment, without the obligatory role of extrinsic factors (Tramontano, 2005).

Moult (1999) defines protein structure prediction to be 'all methods of modeling protein structure from sequence information'. All proteins contain a flexible polypeptide backbone which can adopt an infinite number of spatial configurations (Tang *et al.*, 2005). This large degree of structural freedom, combined with complicated interactions between residues both proximal and distant in sequence, heightens the complexity of modeling the three-dimensional structure of proteins. Determining protein structure from the amino acid sequence is acknowledged as one of the biggest challenges in the field of bioinformatics. It is hoped that the exploratory analysis undertaken in this project will give insight into this problem.

A multitude of factors, including the increasing size of sequence and structural databases, technological advances in computational power and a deeper understanding of the principles of protein structure, have resulted in significant progress being made. However, the difference between the number of known protein sequences and the number of experimentally determined structures is increasing rapidly, meaning that a solution to this problem still eludes bioinformaticians. This reflects a combination of factors, including the relative ease of identifying protein sequences, contrasted by the monumental cost and time requirements related to determining three-dimensional structures.

At present, research demonstrates limited ability to predict protein structure from local sequence information. To improve the accuracy of predictions, a better understanding of the interactions between elements of secondary structure, which are inherently non-local is needed. These interactions create small substructures, known as *motifs* (Gu & Bourne, 2009). The main aim of this project is not to simply rediscover recurring structural patterns or motifs which have

already been discovered, such as $\alpha$-helices and $\beta$-sheets, but to instead identify more subtle motifs, which may not necessarily be a function of a subsequence. The work involves a large proportion of exploratory research, beginning with the compilation of a database of compact structural configurations of heavy atoms, such as carbon, oxygen, nitrogen and sulphur. The size of the database is restricted by computational power, a clear limitation of using a database approach. Despite this restriction, it is hoped that any patterns discovered might contribute to the understanding of the principles of protein folding and hence lead to an improved accuracy in protein structure prediction.

The next part of this chapter recognises the biological background required by the reader to gain a better understanding of the task at hand. The remainder of the chapter adopts a more statistical approach to the problem. Chapter 2 discusses the methodology used in constructing the database, as well as the programming requirements for this problem, along with a brief overview of how the original data was obtained. The results from the cluster analysis are shown in Chapter 3 together with results from principal coordinate analysis and statistical shape analysis. These results are discussed in more detail and interpreted fully in the following chapter. Chapter 4 also discusses the limitations of the procedures used and identifies scenarios for which the methodology becomes unsatisfactory. Chapter 5 summarises the main conclusions drawn from the project and also presents potential future work for this problem.

## 1.2 Biological background

### 1.2.1 Biochemistry review

Proteins are composed of polypeptide chains of amino acid residues (Branden & Tooze, 1999). In total there are 20 amino acids, all of which have a central carbon atom, sometimes referred to as the *backbone*, or *alpha* carbon. A hydrogen atom, amino group and carboxyl group are all attached to the backbone carbon. Also joined to the backbone carbon is a side-chain which determines the type of amino acid present (Lesk, 2000). It is important to note that the carboxyl group contains a carbon frequently referred to as the *carbonyl* carbon. Whilst all amino acids contain the alpha and carbonyl carbon, most contain further carbon atoms, for example, a *lysine* residue contains six carbon atoms (Figure 1.1)[1]. The carbon atom which is attached to the alpha carbon but is not the carbonyl carbon is known as the *beta carbon*. The carbon attached to this is the *gamma carbon*, with subsequent carbon atoms named with chronological Greek alphabet descriptions.

### 1.2.2 Protein structure

The sequence of amino acids is known as the primary structure and is unique for that particular protein (Figure 1.2)[2]. The secondary structure refers to local folding of the polypeptide chain,

[1] http://en.wikipedia.org/wiki/File:Lysine_fisher_struct_num.png
[2] http://barleyworld.org/css430_09/lecture%209-09/figure-09-03.jpg

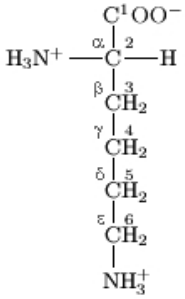

*Figure 1.1: Lysine residue with labeled carbon atoms.*

such as in $\alpha$-helices and $\beta$-sheets. The tertiary structure is the three-dimensional structure of the protein, showing the arrangement of the secondary structures with respect to one another (Branden & Tooze, 1999) (Figure 1.3). The three-dimensional coordinates of every atom in each amino acid residue in a large number of protein structures are required to detect for configurational motifs composed of parts of amino acid residues, which are spatially compact but sequentially dispersed.

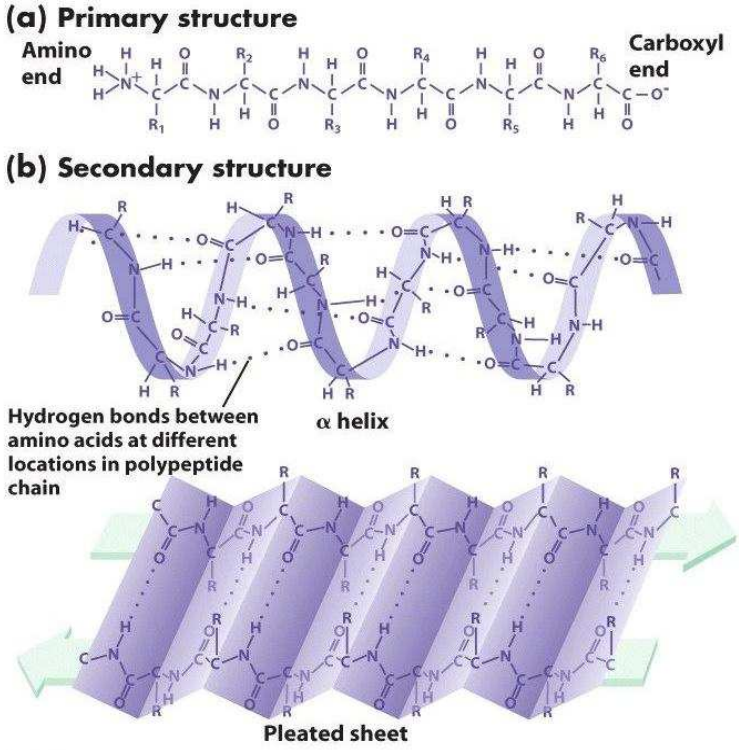


*Figure 1.2: Primary and secondary structure of a protein.*

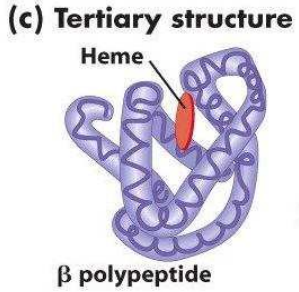


*Figure 1.3: Tertiary structure of a protein.*

### 1.2.3 Bonds and interactions

A polypeptide chain can place any possible set of residues in proximity (Niggemanna & Steipe, 2000). The interactions of the side and main chains determine the energy of the conformation and the true folding pattern of the chain is the conformation which produces a set of interactions that is significantly favourable (Branden & Tooze, 1999). For proteins to function, a stable configuration is required. Interactions between different residues within the protein contribute to its stability. These interactions can be formed between residues close in sequence or residues which are distant from one another in linear sequence but brought into spatial proximity due to folding of the protein Branden & Tooze (1999). Intuitively, interactions between residues distant in the linear sequence add to the complexity of predicting the three-dimensional structure.

Interactions between residues include peptide bonds, electrostatic interactions, disulphide bonds, hydrogen bonds, salt bridges and van-der-Waals interactions, although this is not an exhaustive list (Gu & Bourne, 2009). The four main types of bonding required for the interpretation of results are peptide bonds, disulphide bonds, hydrogen bonds and salt bridges (Figures 1.4-1.7)[3][4]. Peptide bonds essentially hold the polypeptide chain together, joining the carboxyl group of one amino acid with the amino group of the adjacent amino acid (Lesk, 2000). However, these bonds do not contribute to the stabilisation of the conformation as heavily as other interactions (Gu & Bourne, 2009). Ramachandran & Sasisekharan (1968) state that the majority of peptide bonds are found to be in *trans*-conformation, where the dihedral angle is defined to be $180^o$, giving the bond a typically rigid structure. MacArthur & Thornton (1991) and Tramontano (2005) point out that the exception to this rule occurs most visibly in *proline* residues, in which approximately 5% of residues are found to be in *cis*-transformation, where the dihedral angle is very close to $0^o$.

Branden & Tooze (1999) remark that disulphide bonds are only present in proteins which contain sulphur atoms and are restricted to extracellular proteins. Despite these restrictions they are found quite frequently in the Protein Data Bank [5](PDB). Disulphide bridges allow different

[3]http://xray.bmc.uu.se/∼kurs/BiostrukfunkX2/practicals/practical_1/figs
[4]http://www.biog1105-1106.org/demos/105/unit1/
[5]http://www.rcsb.org/pdb/

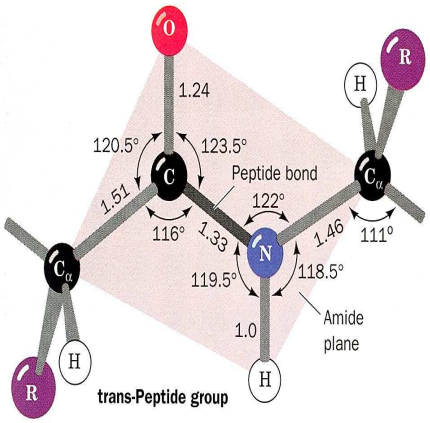


*Figure 1.4: Peptide bond.*

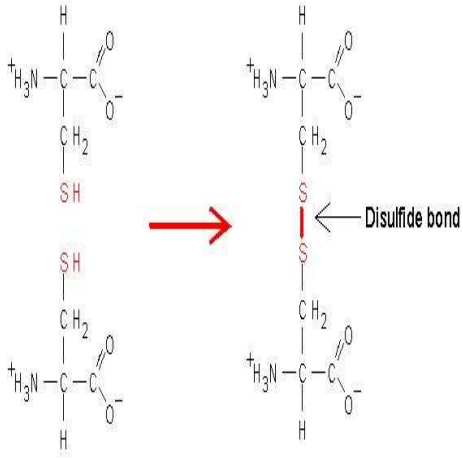


*Figure 1.5: Disulphide bond.*

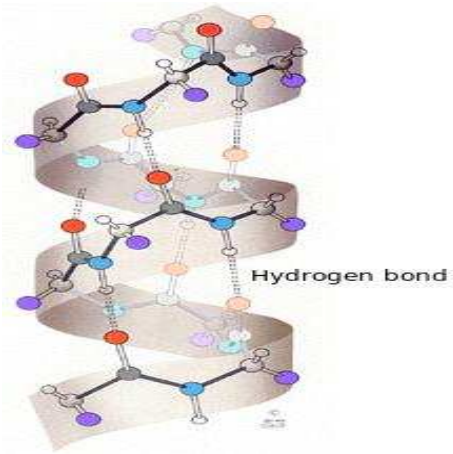


*Figure 1.6: Hydrogen bond.*

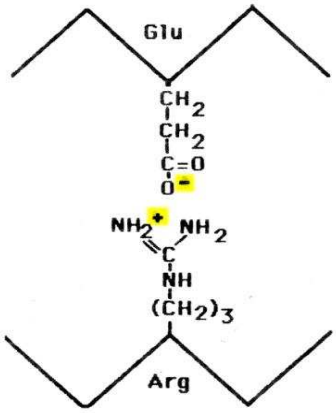


*Figure 1.7: Salt bridge.*

parts of the polypeptide chain to be covalently bound and as a consequence are understood to help stabilise the three-dimensional structure of a protein (Schulze-Kremer, 1996). This stabilisation is believed to be accomplished by either linking two polypeptide chains, or by stabilising the folding of a single chain. Disulphide bridges allow two cysteine residues to be sequentially dispersed but adjacent in three-dimensional space, creating a configurational motif. This is extremely useful as it allows the methodology used here to be informally tested.

Hydrogen bonds play a fundamental role in the secondary structure of proteins, forming between backbone oxygen atoms and amide hydrogen atoms. The bond occurs when two electronegative atoms interact with the same hydrogen. This hydrogen atom is covalently bonded to one of the atoms, known as the *donor*, and interacts electrostatically with the other atom, known as the *acceptor* (Lodish *et al.*, 2007). The well known secondary structures, $\alpha$-helices and $\beta$-sheets can be formed depending on the spacing of the amino acid residues involved in the hydrogen bond (Figure 1.2).

Salt bridges describe hydrogen bonds formed by the interaction between two charged residues

(Lesk, 2000). There are two positively charged (*arginine* and *lysine*) and two negatively charged amino acids (*asparatic acid* and *glutamic acid*), thus salt bridges can form between four different combinations of amino acid residues.

To summarise, interactions between atoms influence the folding of a protein. Interactions can occur between residues close in sequence or between sequentially distant residues. Recent work has been shaped by the belief that a newly created polypeptide may form local folds in parts before completion of the folding process Tang *et al.* (2005). As a result of this conjecture, attempting to observe small structural motifs appears to be the next step in protein structure prediction. These small configurational motifs could then be viewed as "building-blocks" which can be pieced together in order to predict the whole three-dimensional structure of a given protein.

## 1.3 Statistical background

It is important to note that specific amino acid residues occur in differing frequencies in different proteins. Moreover, different types of atoms are present in different amino acids; for example, sulphur atoms only occur in two out of the 20 amino acids, namely *cysteine* and *methionine*, emphasising the need for the protein structures used in the database to be representative of protein structures as a whole. Ulmschneider & Sansom (2001) show that the frequency of amino acid varies depending on what type of secondary structure the atom is in, making statistical analysis more challenging.

Protein sequences are strongly autocorrelated, for example, if one residue is in a helix, then the next one is also likely to be in the same helix, adding to the complexity of statistical analysis. Interestingly, despite helices and sheets being regular secondary structure elements, a method of assigning these structures from atom coordinates does not exist. All of these factors emphasise the huge challenge in protein structure prediction. There are many ways to tackle this particular problem but, intuitively, methods involving exploratory statistical analysis appear to be a sensible starting point. The next part of this chapter addresses the techniques used here.

### 1.3.1 Cluster analysis

Cluster analysis is the main statistical tool used to detect configurational motifs in protein structures. Clustering techniques are of course exploratory and are thus used in conjunction with other statistical methods in order to further analyse any results obtained. Unlike many other statistical procedures, cluster analysis methods are typically used for exploratory research when no a priori hypotheses are known; a technique ideal for this type of project. Unsupervised learning methods, such as cluster analysis, are clearly appropriate here because of the exploratory nature of the work, but this does not guarantee definite clustering solutions. Clustering methods depend on many sources of human judgement, such as the type of distance measure used to calculate the dissimilarity between objects, as well as the type of clustering algorithm used. Results can be

subjective and, as expected, there are advantages and disadvantages of each method. Therefore, serious justification and validation is required for each step of the clustering procedure.

To briefly summarise, cluster analysis is essentially concerned with identifying any 'natural groupings' in a set of objects. Some clustering techniques work well on certain types of data but not on others. Everitt & Dunn (2001) remark that the method used can make implicit assumptions about the structure present in the sample. If these assumptions do not hold, then the clustering method itself may impose a structure on the data set, rather than find a natural clustering of the observations, if of course there is one, posing the question of whether or not natural groupings of the data actually exist.

### 1.3.2 Hierarchical agglomerative clustering

As expected, exploratory analysis of finding structurally equivalent sets of motifs begins by implementing hierarchical agglomerative methods. These techniques are ideal because they require no initial knowledge about the clustering solution, such as the number of clusters present in the dataset. Since no prior knowledge concerning either the number or the size of possible structural motifs is known, this technique is clearly useful for the initial phase of research.

There are of course several drawbacks to using such methods; Kaufman & Rousseeuw (1990) state that the key disadvantage of using these techniques is that "*a hierarchical method suffers from the defect that it can never repair what was done in previous steps*". However, if configurational motifs are distinct from one another, this potential problem becomes negligible. The major difficulty of this particular technique lies in the timing of the termination of the clustering algorithm. Usually the iterations are run up to a threshold point, decided by the investigator in order to identify the 'optimal' number of clusters. This is achieved by cutting the dendrogram at a specific height, sometimes known as the *best cut* and Everitt, Landau & Leese (2001) note that this is indicated by large changes in fusion levels. If increases in mean internal distance are small then this suggests that merging clusters are similar to one another; however a noticeable jump indicates that dissimilar clusters are merging and the clustering algorithm should be stopped prior to this agglomeration. More formal methods are reviewed by Milligan & Cooper (1985).

Hierarchical clustering algorithms can differ in relation to the type of inter-group proximity measure used. This measure defines the distance between two clusters and different measures will subsequently give rise to different clustering solutions. The clustering methods considered here are *complete linkage* and *Ward's method*. The former defines the distance between clusters $A$ and $B$ as the *maximum* distance between one element of $A$ and one element of $B$. Ward (1963), on the other hand, introduced a method which defines the distance between clusters $A$ and $B$ by the increase in sum of squares within clusters, after fusion of the two clusters. The main goal at each stage of the clustering process is to minimise the increase in the total within-cluster error sum of squares. Everitt, Landau & Leese (2001) remark that that this can be a highly efficient method which tends to find small sized, spherical clusters.

### 1.3.3 Principal coordinate analysis

Principal coordinate analysis (PCoA) is a type of scaling procedure, commonly referred to as *classical scaling*. The main objective of PCoA is to produce a *map* of the objects in a small number of dimensions. This attribute of dimension reduction makes comparisons between objects clearer, especially when the original dataset is of high dimensionality Chatfield & Collins (1980). It is essentially a method of graphical display, plotting objects as points against principal coordinate axes and can prove extremely useful for examining results of a cluster analysis; a feature which complements the exploratory statistical analysis carried out in this project.

Recall that the atom coordinates have been experimentally determined in three-dimensional space. Since it is reasonably difficult to compare these configurations visually, it would be ideal to reduce the number of dimensions from three to two, allowing for easier comparisons between configurations.

Not to be confused with principal coordinate analysis, principal component analysis (PCA) is a variable-directed technique, where the main objective is to replace the original variables in the dataset by a smaller number of underlying variables, thereby reducing the dimensionality of the dataset. The problem in using PCA directly relates to the arbitrariness of the $x$, $y$, and $z$ axis. The atoms could undergo similarity transformations, such as rotation, reflection and translation, and yet the distance between the atoms would remain the same, implying that the axes have no real intrinsic meaning. PCoA, on the other hand, makes use of the distance between pairs of points rather than the actual coordinates. Therefore, it is applied here as a useful tool for visualising results from a cluster analysis.

Chatfield & Collins (1980) note that given a set of Euclidean distances, no unique representation of the points exist since distance preserving transformations do not affect the solution. Hence the points are only defined up to location, reflection and rotation, meaning that the location and orientation of the configuration cannot be directly determined.

Suppose that the Euclidean distances, $d_{rs}$, between a set of objects, are contained in an $n \times n$ matrix, $D$ say. The coordinates of the configuration can be estimated, by first introducing a matrix $B$, where the $(r, s)$th element of $B$ is given by

$$b_{rs} = \sum_{i=1}^{p} x_{ri} x_{si}, \tag{1.1}$$

and then by factorising it to be of the form $B = XX^T$, where $X$ is the original data matrix Chatfield & Collins (1980). Let $d_{rs}^2$= squared Euclidean distance between points $r$ and $s$. Then

$$\begin{aligned} d_{rs}^2 &= \sum_{i=1}^{p}(x_{ri}-x_{si})^2 \\ &= \sum_{i=1}^{p}(x_{ri}-x_{ri}) + \sum_{i=1}^{p}(x_{si}-x_{si}) - 2\sum_{i=1}^{p}(x_{ri}-x_{si}) \\ &= b_{rr}+b_{ss}-2b_{rs}. \end{aligned} \tag{1.2}$$

Equation (1.2) can then be inverted to find the elements of $B$ in terms of the $\{d_{rs}^2\}$. A location constraint must be applied first though to obtain a unique solution. This constraint usually involves centering the data, i.e setting $\sum_r^n x_{ri} = 0$ for all $i = 1, \ldots, p$. Combining these constraints with equation (1.1) implies that the rows and columns of $B$ sum to zero. Now, summing equation (1.2) over $r$ yields

$$\begin{aligned} \sum_{r=1}^{n} d_{rs}^2 &= \sum_{r=1}^{n} b_{rr} + \sum_{r=1}^{n} b_{ss} - 2\sum_{r=1}^{n} b_{rs} \\ &= T + nb_{ss}, \end{aligned} \tag{1.3}$$

since the third term vanishes due to summing over a centered variable, and where $T = \sum_{i=1}^{n} b_{rr}$. Similarly, summing over $s$ and $r$ and $s$ together gives

$$\sum_{s=1}^{n} d_{rs}^2 = nb_{rr} + T, \tag{1.4}$$

$$\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2 = 2nT. \tag{1.5}$$

Now substituting equations (1.3)-(1.5) into equation (1.1) gives

$$b_{rs} = -\frac{1}{2}\left[d_{rs}^2 - \frac{1}{n}\sum_{s=1}^{n} d_{rs}^2 - \frac{1}{n}\sum_{r=1}^{n} d_{rs}^2 + \frac{1}{n^2}\sum_{r=1}^{n}\sum_{s=1}^{n} d_{rs}^2\right].$$

Setting $A = -\frac{1}{2}d_{rs}^2$ means that $B$ can be written in the form $B = HAH$, where $H$ is the standard $n \times n$ centering matrix, i.e. $H = I_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$. $B$ is symmetric and positive semi-definite and thus can be written in terms of its spectral decomposition

$$B = \Gamma\Lambda\Gamma^T, \tag{1.6}$$

where $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ denotes the ordered eigenvalues of $B$ with corresponding unit eigenvectors $\Gamma = [\gamma_{(1)}, \ldots, \gamma_{(n)}]$.

If $D$ is formed from an $n \times q$ matrix of full rank, then $B$ has rank $q$ ($q \leq n$) and

$\lambda_{q+1} = \ldots = \lambda_n = 0$ (Everitt, 2005). Putting this into equation (1.6) means that $B$ can be rewritten as

$$B = \Gamma_1 \Lambda_1 \Gamma_1^T,$$

where $\Lambda_1 = \text{diag}(\lambda_1, \ldots, \lambda_q)$ with corresponding unit eigenvectors $\Gamma_1 = [\gamma_{(1)}, \ldots, \gamma_{(q)}]$. The coordinate values can be recovered by setting $X = \Gamma_1 \Lambda_1^{\frac{1}{2}}$. A $k$-dimensional principal coordinates representation is then achieved by plotting the coordinates with respect to the first $k$ principal coordinates, giving the best fitting $k$-dimensional representation of the objects.

Jeffers (1992) points out that there is no simple interpretation for the eigenvalues obtained. The first principal coordinate simply maximises the total squared distance between the objects, whilst the second maximises the squared distance along an axis orthogonal to the first. Interpreting axes is even more complex, and is thus more complicated than in PCA. However, a graphical representation of the configurations in two dimensions should give a helpful insight into any recurring structural motifs observed from the results of cluster analysis.

Chatfield & Collins (1980) note that despite making little use of an underlying stochastic model, principle coordinate analysis and cluster analysis are both techniques used to give an initial summary of the data, from which hypotheses can be generated and then tested in a statistical setting.

### 1.3.4 Statistical shape analysis

Recall that proteins can be expressed as a sequence of amino acids which are made up of different atoms. Representing each atom as a single point allows statistical shape analysis to be applied to each configuration of atoms. Shape consists of those properties of a geometric object that are invariant under similarity transformations, i.e. under translation, rotation and scaling. Configurational motifs which have been clustered together should have a similar shape to one another since the distance between corresponding atoms in different configurations grouped in the same cluster should be roughly equal. A recurring structural motif will, therefore, have the same shape regardless of the three-dimensional coordinates of the index atom or the rotation of the configuration about this index atom. To summarise, the shape of the configuration is regarded as its fundamental property, whereas the registration parameters, such as location, orientation and to some extent size, are merely artefacts of the observation process. It is of course important to consider the possibility that there could be recurring structural motifs which have the same shape but are of varying size. Each configuration can be represented by a $k \times p$ matrix,

$$X = \begin{bmatrix} \mathbf{x}_1^T \\ \ldots \\ \mathbf{x}_k^T \end{bmatrix}, \tag{1.7}$$

of $k$ landmarks in $\mathbb{R}^p$, where the $j$th row of $X$, denoted by $\mathbf{x}_j^T$, $j = 1, \ldots, k$, represents the

coordinates in $\mathbb{R}^p$ of the $j$th landmark. Note that in (1.7), $\mathbf{x}_j$ is a column vector. Since the atoms are determined in three-dimensional space, $p = 3$ here.

### Bookstein coordinates

One approach could be to construct Bookstein coordinates (Bookstein, 1986) for each configuration and perform statistical analysis on these coordinates. Dryden & Mardia (1998) point out that Bookstein coordinates can give a basic overview of the shape of a set of configurations since interpretation of the shape coordinates is simple in the majority of cases. Another advantage of Bookstein coordinates is that graphical displays are relatively easy to produce and can be carried out using standard computer packages. However, there are some undesirable features of using these coordinates. Not only are Bookstein coordinates inappropriate for non-concentrated shapes but they are also inadvisable when two landmarks are nearly coincident. Moreover, the choice of baseline is an arbitrary one and since different baseline landmarks give different results, this arbitrary choice makes Bookstein coordinates somewhat unsatisfactory. Dryden & Mardia (1998) also note that registering landmarks with respect to a common edge can induce correlations in the data, hindering interpretation of shape coordinates.

### Procrustes tangent coordinates

A more fitting approach is to use Procrustes tangent coordinates (Dryden & Mardia, 1998). In this coordinate system, configurations are registered with respect to one another, rather than with respect to a common baseline. Procrustes analysis is subdivided into 2 settings - ordinary (OPA) and generalised (GPA), both of which exist in 2 forms - partial and full. The aim is to create an "average" shape using one of the settings and then represent configurations as "residuals" about this mean shape. Three different Procrustes distances are available, namely full, partial and angular; however for concentrated data there is minimal difference between each distance and in fact it can be shown that the three distance measures are asymptotically equivalent for concentrated configurations (Dryden & Mardia, 1998).

To summarise, the basic object of interest is the shape of the configuration, represented in Procrustes tangent coordinates about a mean shape, $\hat{\mu}$ say, as a $3k$-dimensional vector. The mean shape, $\hat{\mu}$, is usually taken to be the GPA estimate of mean shape for the combined dataset, although it can be chosen to be the GPA estimate of mean shape of either sets of configurations, $X$ or $Y$, say, where $X$ and $Y$ are $k \times p \times n_x$ and $k \times p \times n_y$ arrays, respectively. The Procrustes tangent vectors, $\mathbf{w}_i^{(x)}, i = 1, \ldots, n_x$ and $\mathbf{w}_i^{(y)}, i = 1, \ldots, n_y$, are then calculated for both sets of configurations, where each $\mathbf{w}_i$ is a $3k$-dimensional vector. Standard multivariate analysis can then be performed in this linear space in order to perform tests comparing the shapes of sets of configurations.

### Hotelling's $T^2$ test

Two sample comparisons can be carried out to compare two sets of configurations using the two-sample Hotelling's $T^2$ test. This test is used primarily to test the null hypothesis, $H_0$: the two sets of configurations have equal mean shapes, measuring the distance in $M$ space between the mean shapes in dimensionless units. The test can also be used to gain further insight into differences between the two groups should the null hypothesis be rejected. The test is essentially an extension of the standard *t-test* into multidimensional space.

More formally, let $[\mu_x]$ and $[\mu_y]$ denote the mean shapes of two groups of configurations. The two hypotheses

$$H_0 : [\mu_x] = [\mu_y] \text{ vs. } H_1 : [\mu_x] \neq [\mu_y],$$

can be tested using Hotelling's $T^2$ test in Procrustes tangent space. Let

$$\mathbf{w}_i^{(x)} \sim N(\xi_x, \Sigma), \quad i = 1, \ldots, n_x, \text{ and } \mathbf{w}_i^{(y)} \sim N(\xi_y, \Sigma), \quad i = 1, \ldots, n_y,$$

where the $\mathbf{w}_i^{(x)}$ and $\mathbf{w}_i^{(y)}$ are mutually independent and the covariance matrices are assumed to be the same. Before Hotelling's $T^2$ test statistic can be calculated, the mean vector of each group is required. This will be denoted by $\mathbf{w}_{ave}^{(x)}$ and $\mathbf{w}_{ave}^{(y)}$ for the two sets of configurations. The difference between the two means is simply

$$\mathbf{d} = \mathbf{w}_{ave}^{(x)} - \mathbf{w}_{ave}^{(y)},$$

where $\mathbf{d}$ is a $3k$-dimensional vector. Finally, a pooled sample covariance matrix, $S$, is required.

### Moore-Penrose generalised inverse

The tangent coordinates cannot incorporate changes in either location, orientation or scaling of $\hat{\mu}$, and therefore $S$ is singular since it has 7 zero eigenvalues (3 for location, 3 for orientation and 1 for scaling). As a result of this $S$ is not invertible; however, a Moore-Penrose generalised inverse $S^-$ exists, where

$$S^- = \sum_{j=1}^{3k-7} l_j^{-1} \mathbf{g}_j \mathbf{g}_j^T, \tag{1.8}$$

and $\mathbf{g}_j$ is the $j$th column of $G$, where the spectral decomposition of $S$ is $GLG^T$ (Penrose, 1955). Note that $L$ is a diagonal matrix with ordered eigenvalues along the diagonal and therefore the last 7 elements will all be equal to zero. The squared Mahalanobis distance between $\mathbf{w}_{ave}^{(x)}$ and $\mathbf{w}_{ave}^{(y)}$ is

$$D^2 = \left(\mathbf{w}_{ave}^{(x)} - \mathbf{w}_{ave}^{(y)}\right)^T S^- \left(\mathbf{w}_{ave}^{(x)} - \mathbf{w}_{ave}^{(y)}\right), \tag{1.9}$$

where the Moore-Penrose generalised inverse, $S^-$ is defined as in equation (1.8). Hotelling's

$T^2$ test statistic can then be calculated using the squared Mahalanobs distance, $D^2$, as defined in equation (1.9), and the following formula

$$T^2 = \left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{-1} D^2 \quad \sim T^2(3k-7, n_x+n_y-2).$$

Using the fact that $\frac{q-p+1}{pq}T^2(p,q) = F(p, q-p+1)$, a $p$-value can be calculated to either accept or reject the null hypothesis. Note that under $H_0 : \xi_x = \xi_y$, the test statistic is

$$F = \frac{n_x n_y (n_x+n_y-M-1)}{(n_x+n_y)(n_x+n_y-2)M} D^2 \sim F_{M,n_x+n_y-M-1},$$

where $M$ is the dimension of the shape space. Since each atom is determined in three-dimensional space the shape space will have dimension $3k-7$, where $k$ is the number of atoms in each configuration. The null hypothesis is rejected for large values of $F$.

## 1.4 Summary

The three-dimensional structure of a protein is dictated by its amino acid sequence. Determining the structure of a given protein can provide important information about the function of the protein. Hence, there is an obvious link between a protein's amino acid sequence and it's function, emphasising the importance of both awareness and understanding of the three-dimensional structures of proteins. The study of protein structure prediction, i.e. predicting the three-dimensional structure of a protein given its amino acid sequence, is a colossal challenge but is receiving increasing levels of attention, resulting in a larger number structures known today.

### 1.4.1 Recent work

Recent work in this field includes a number of ad hoc data-based approaches, which have had some limited success in protein structure prediction, as well as ab initio structure prediction methods; an area described by Moult (1999) as the "Holy Grail of the protein modeling field". Ab initio techniques are designed to model protein structures without direct knowledge of experimentally determined structures. Despite this idealistic approach, progress has been limited with prediction methods so far proving elusive. If successful, these methods will provide a test of current understanding of the principles of protein structure. Moreover, ab initio techniques will be needed to model the differences between structures even when databases are present, highlighting the significance of such methods.

Fasnacht (2002) introduces a method for automatically finding structural motifs in proteins. The work mainly focuses on a special case of secondary structure interaction called coupled helical motifs, which consist of two interacting helices. However, this method can be applied more

generally to any type of structure. Based on hierarchical agglomerative clustering, the method is shown to rediscover important features of the known structural motif using the *root-mean-square* (rms) distance between atoms as a distant metric to compare pairs of helix couples. The rms distance between a pair of structures, $i$ and $j$ say, involves the sum of the differences in corresponding atoms and is defined as

$$d_{ij} = \min \sqrt{\frac{\sum_{k=1}^{N} ||\mathbf{r}_{i,j} - \mathbf{r}_{j,k}||^2}{N}}, \tag{1.10}$$

where $\mathbf{r}_{i,k}$ is the position vector of the $k$th atom in the $i$th structure. It is viewed as a standard measure of structural distance between coordinate sets and can be used to compare both the complete three-dimensional structure of a protein and also any subset of this structure, such as a configurational motif. This distance measure could be used to identify configurational motifs when new discoveries are made.

Qian & Sejnowski (1988) suggested that the best method of predicting the structure of an unknown protein was to observe the structure of a homologous protein. Not only are very few structures currently known, the set of solved structures could potentially be biased towards particular types of protein, meaning that some proteins may not have homologous proteins with known structures, rendering this method insufficient.

Currently it seems that the best approach is to try and identify small configurational motifs, since interactions between residues in proteins can occur between sequential residues as well as residues distant in sequence. As a consequence, these structural motifs may not necessarily contain residues adjacent in linear sequence. If prominent, these recurring patterns could support current understanding of the principles involved in protein folding, and therefore improve the accuracy of protein structure prediction. Interestingly, Bystroff *et al* (1999) point out that if there is a finite number of unique local structural motifs contained within all proteins, then multiple sequence alignments of protein sequences should also exhibit a finite number of patterns of sequence variation.

Sternberg (1996) raises the interesting point that protein structures are more highly conserved than their sequences. During evolution, insertions and deletions occur predominantly in loop regions, the areas between secondary structures, therefore altering the sequence of the protein but not the folding process, resulting in the protein structure remaining unchanged. This partially explains why a large proportion of the experimentally determined structures in the PDB are very similar to each other.

### Limited number of protein folds

As the number of determined protein structures increases, the number of observed folds increases. Thornton (2001) notes that in 2001 however, only 5% of all newly determined structures adopted a novel fold, suggesting that there could be a relatively small number of protein folds from which all proteins have evolved. Chothia (1993) suggests that there could possibly

be fewer than 1000 folds in total, although this figure remains highly questionable. Xu *et al.* (2007) remark that both theoretical and statistical studies (Murzin *et al.*, 1995), (Brenner *et al.*, 1996) and (Wang, 1996) appear to suggest that the number of unique structural folds in nature is most likely to be somewhere in the region of a few hundred to a few thousand. For proteins to function a stable conformation is required, restricting the number of allowed amino acid sequences. However, it may be case that some stable folds are not present in nature, since the folds that exist today are the result of an evolutionary process in which not all possible sequence combinations have been attempted.

Currently only around a few hundred unique protein folds are known. It is hoped that if there is indeed a limited number of protein folds, then an experimental determination of each type of fold can be produced. Methods which can identify fold types from sequences can consequently be implemented, enabling the prediction of the three-dimensional structure of a protein from sequence alone.

## Review

Overall, it is hoped that identifying recurring structural motifs will lead to an improved understanding of the principles of protein folding in three-dimensional space. The methods implemented here are inherently exploratory and are used primarily to help reveal any systematic patterns in the atomic configurations studied. These patterns can then be further investigated with the long term goal of statistically analysing all identifiable patterns in as many protein structures as possible. Obviously some observed 'patterns' may occur purely by chance and as a result do not actually represent a link between sequence and structure. Hence, the major objective of this work is to eventually test the significance of any systematic patterns in the hope of developing an understanding of this link between primary and tertiary structure.

# Chapter 2

# Methodology

This chapter briefly summarises the methods used in this project with reference to the construction of the database and the identification of recurring structural motifs from this database. Also included is an overview of how protein structures are determined and the accessibility and credibility of structures found online. Several limitations relating to the construction of the database are also considered here. The last part of the chapter describes the compilation of the database, stating the distance measures and metrics used when comparing individual configurations.

## 2.1 PDB

The Protein Data Bank (PDB) is a repository for the three-dimensional structural data of tens of thousands of protein structures. It is perceived as a primary data bank, from which secondary databases can be constructed in the interest of research. Governed by the Worldwide Protein Data Bank (wwPDB) to help maintain the credibility and quality of the information stored in the data bank, the PDB is vital for the integrity of this project.

### 2.1.1 X-ray crystallography

For each protein structure within the database the three-dimensional coordinates are given for each atom and this information is used as the main driving-force in detecting recurring structural patterns. All of the information in the PDB archive has been determined experimentally, using either x-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. NMR is restricted in its use on small protein molecules and therefore protein structures determined by this particular method are not considered here. Hence, for this project only protein structures determined by x-ray crystallography shall be considered. This method involves directing a beam of x-rays onto a regular, repeating array of identical molecules arranged in a crystal. The interaction of the x-rays with the electrons in the molecules produces an electron-density map, from which the structure of the molecule can be obtained (Branden & Tooze, 1999).

Ideally all possible protein structures within the PDB would be used to compile the database; however, using even just a reasonable number of protein structures would be deemed computationally expensive. Instead, several protein structures shall be chosen at random from the data bank. Before the protein structures are selected to use in the database, a constraint on the resolution of the protein structure must be applied. X-ray structures are determined at different levels of resolution, measured in Ångströms (Å). The smaller the number, the higher the resolution and the greater the amount of detail that can be seen. To give a brief summary, a protein structure is considered to have a low resolution if the resolution is greater than or equal to 5Å, medium if around 3Å, and high if the resolution is less than this value. Since a high resolution implies that most atomic positions within the structure can be determined to a high degree of accuracy, it will be of greater benefit to select protein structures which have a high resolution to use in the database. Therefore, to begin with only protein structures with of a resolution $\leq 2$Å shall be considered.

### 2.1.2 Drawbacks

There are several limitations regarding the process used in compiling the database. Firstly the database relies on the quality of the information contained in the PDB. Moreover, x-ray crystallography requires well-ordered crystals to produce accurate results; however, these crystals are difficult to grow due to the irregular nature of the protein molecules in question (Branden & Tooze, 1999). Since some forms of protein molecules are easier to crystallise than others, the set of solved structures is biased towards those proteins which are better suited to the conditions required in x-ray crystallography. Branden & Tooze (1999) give one example relating to this idea of suitability, stating that globular proteins are, in general, easier to crystallise than membrane proteins. As a consequence of this, fewer membrane proteins are present in the PDB. The limitations discussed above suggest that great care must be taken when compiling the database of compact structural configurations and this should be taken into consideration when drawing conclusions.

Protein structures in the PDB often contain hydrogen atoms, which are frequently unreported or inaccurately recorded. Those structures in the PDB which report the coordinates of hydrogen atoms often employ imputation methods for positioning hydrogen atoms, rather than experimental determination, therefore increasing the error in the accuracy of the atomic coordinates. As a result of these inconsistencies, hydrogen atoms shall be disregarded from the database. It should be noted that approximately half of the atoms in a protein are hydrogen molecules and thus by ignoring these, a large proportion of individual atoms in each observed structure will be disregarded. However, due to such vast discrepancies in the reporting of hydrogen atoms, this appears to be the most appropriate way to progress with this project.

## 2.2 Procedure

The protein structures, selected randomly from the PDB are stored as *pdb* files and read into **R** with the assistance of the specially created *bio3d* package. An idealistic approach would be to run in all possible protein structures from the PDB; nonetheless the computational requirements increase exponentially as the number of configurations in the database increases. The limitation of computational power can be overcome to some extent by addressing the number of calculations required to produce results of interest, allowing more protein structures to be incorporated into the database.

### 2.2.1 Database compilation

The analysis begins with the compilation of a database of compact structural configurations of heavy atoms, including carbon, nitrogen, oxygen and sulphur. Each compact configuration $C_k$, say, will be centred on one atom from a PDB protein structure. This atom, $k$, will be referred to as the *index* atom of the configuration. Interest lies in the $m-1$ atoms closest to the index atom in three-dimensional space, using the most appropriate value of $m$. The idea is that some exploratory work is required to find a value of $m$ large enough to ensure that the database is not dominated by local sequence substructures, such as $\alpha$-helix fragments, but small enough so that recurring configurations of that size actually exist.

#### Calculating the Euclidean distance

For each configuration, $C_k$, the $m-1$ atoms closest to atom $k$ in three-dimensional space must first be discovered. For a protein $p$ containing $n$ heavy atoms, the Euclidean distance between the index atom $k$, and any atom $i$, $i-1,\ldots,n$, can be calculated using

$$d_{ki}^{(p)} = \sqrt{\left(x_k^{(p)} - x_i^{(p)}\right)^2 + \left(y_k^{(p)} - y_i^{(p)}\right)^2 + \left(z_k^{(p)} - z_i^{(p)}\right)^2}, \tag{2.1}$$

for all index atoms $k = 1\ldots,n$, where $(x_i^{(p)}, y_i^{(p)}, z_i^{(p)})$ denotes the spatial coordinates of the $i$th atom of protein $p$.

This gives the Euclidean distance between each index atom and every other atom in the protein. However, only the closest $m-1$ atoms to the index atom $k$ are of interest, and hence the spatial coordinates of the closest non-trivial atom can be denoted by $(x_2^{(k)}, y_2^{(k)}, z_2^{(k)})$ and the second closest by $(x_3^{(k)}, y_3^{(k)}, z_3^{(k)})$, and so on, until the $m$th closest atom is reached. More generally, the spatial coordinates of the $j$th closest non-trivial atom are labelled $(x_{j+1}^{(k)}, y_{j+1}^{(k)}, z_{j+1}^{(k)})$, $j = 1\ldots,m-1$. The spatial coordinates of the index atom are denoted by $(x_1^{(k)}, y_1^{(k)}, z_1^{(k)})$.

The database then contains the Euclidean distances, $d_{ij}^{(k)}$, between each pair $(i,j)$, $i = 1,\ldots m$, $j = 1,\ldots m$, of these $m$ atoms for each configuration $k$. As in (2.1), the distances are calculated as follows

$$d_{ij}^{(k)} = \sqrt{\left(x_i^{(k)} - x_j^{(k)}\right)^2 + \left(y_i^{(k)} - y_j^{(k)}\right)^2 + \left(z_i^{(k)} - z_j^{(k)}\right)^2}, \tag{2.2}$$

where $(x_i^{(k)}, y_i^{(k)}, z_i^{(k)})$ represents the spatial coordinates of the $i$th atom of configuration $C_k$. Thus each configuration is represented by a symmetric $m \times m$ distance matrix with each cell denoting the Euclidean distance between each of the $m$ atoms. Trivially, the leading diagonal elements will be zero and since the matrix is symmetric, only $\frac{m(m-1)}{2}$ distances are required. The storage of distance matrices rather than the original three-dimensional atomic coordinates is advantageous since it eliminates unimportant orientation information. For example, a set of $m$ atoms may produce the same distance matrix as a different set of $m$ atoms but in another area of the three-dimensional space or possibly at a different orientation about the index atom.

Each heavy atom for each protein structure will contribute one configuration towards the database, resulting in an enormous database. As stated earlier, each configuration is an $m \times m$ distance matrix. To make analysis universal, a standard ordering of the rows and columns of each distance matrix has been implemented. The matrix will contain the index atom in the first position, followed by the remaining $m - 1$ atoms positioned in order of their relative proximity to the index atom, i.e. the atom nearest to the index atom is placed in the second position of the matrix and so forth. This is achieved by using the specially created function named *protein* (see A.2.1)

According to Brocchieri & Karlin (2005), a protein of average length contains approximately 300 amino acid residues. Ignoring hydrogen atoms, each amino acid has between 5 and 15 heavy atoms and each of these atoms relates to a configuration $C_k$. This implies that, on average, for each protein added to the database, around 3000 configurations are produced. This is a very basic estimate since amino acids occur in differing frequencies, but it does give the reader, who may be unfamiliar with the biological background, a rough idea of the size of the database.

### 2.2.2 Dissimilarity measure

To compare two configurations, $C_k$ and $C_l$, a metric must first be formulated. An initial choice is to define the metric as follows

$$D_{kl} = \frac{2}{m(m-1)} \sum_{j=2}^{m} \sum_{i=1}^{j-1} |d_{ij}^{(k)} - d_{ij}^{(l)}|, \tag{2.3}$$

where $d_{ij}^{(k)}$ is defined as in (2.2) and $D_{kl}$ is formally the dissimilarity between the two configurations, $C_k$ and $C_l$. The implementation of the standard ordering of the columns and rows justifies the choice of metric with the factor $\frac{2}{m(m-1)}$ representing the reciprocal of the total number of distances summated.

Using (2.3), the dissimilarity between each configuration in the database can then be calculated and a dissimilarity matrix compiled with cluster analysis being performed on all config-

urations. If recurring structural patterns exist within the database, then these structural motifs will have very similar configurations, or distance matrices, and hence the dissimilarity between two such configurations will be minimal. These configurations will then be grouped within the same cluster with other configurations of a similar nature.

The *cutree* function in **R** can be used to observe the groupings of the configurations at different points along the clustering algorithm. If the dendrogram is cut too low and agglomerative clustering techniques are applied say, then all configurations could be individual clusters, giving the observer no information regarding the similarity of any of the configurations. On the other hand, if the dendrogram is cut too high, then the majority of the configurations will all belong to one large cluster. A brief visualisation of the dendrogram should give an idea of a sensible range of cut-off points.

Once a potential clustering has been identified, the cluster membership of each individual configuration is needed. The atoms within each configuration must then be identified (see A.2.2) in order to observe any potentially recurring patterns within each cluster.

The initial metric in (2.3), used to compare two configurations, $C_k$ and $C_l$, can detect configurational motifs of the same size and of approximately spherical shape. However, there is a distinct possibility that commonly recurring structural motifs vary in shape and size, rendering the search methodology unsatisfactory. One way of adapting the methodology is to create a metric which allows for a proportion of the $m$ atoms to be disregarded. The number of atoms excluded from the distance measure can be dependent on the size of the differences between elements $d_{ij}^{(k)}$ and $d_{ij}^{(l)}$, relative to a critical value. A threshold value is thus required to determine how many atoms should be ignored.

In some sense, similar to the principle of the bootstrap philosophy, this novel metric will allow the data to 'speak for itself', choosing the most appropriate value of $m$. As a consequence of this, a major advantage of this new metric is that the methodology is much less sensitive to the choice of $m$. This adaptation to the methodology is further discussed in Chapters 4 & 5.

# Chapter 3

# Data Analysis

In this chapter the results of the statistical exploratory data analysis are reported, concentrating primarily on systematic patterns of interest. The methods used here include cluster analysis, principal coordinate analysis and statistical shape analysis, as well as the implementation of a chi-squared test of independence.

## 3.1 Hierarchical agglomerative clustering: original database

Hierarchical agglomerative clustering is performed on a total of 8,336 configurations from 5 protein structures, using both complete linkage (Figure 3.1) and Ward's linkage (Figure 3.2). Analysis begins by setting the number of atoms in each configuration equal to 10, i.e. $m = 10$. This means that each configuration contains the index atom and the 9 closest atoms to this point in three-dimensional space. The metric used to calculate the dissimilarity between two configurations is as in equation (2.3), which essentially calculates the average difference between corresponding cells of the two $m \times m$ matrices.

The two cluster dendrograms (Figures 3.1-3.2) appear to be very different from one another; the agglomerations of configurations when using complete linkage appear to occur at roughly the same height, whereas when Ward's linkage is considered the clusters appear to be far more distinct. Figure (3.2) shows a large number of tight clusters, possibly indicating the presence of configurational motifs. The dendrogram is cut at a height of 2 giving rise to 736 clusters. Table 3.1 shows the 7 configurations found in cluster number 1. Each configuration contains the atom number, element type, amino acid residue and residue number of each of the atoms in the configuration, together with the PDB accession number of the protein containing the atoms. It can be seen that the amino acids belonging to each configuration are almost exclusively different from one another, indicating that there is no immediately obvious pattern. Moreover, the residue numbers are sequential demonstrating that these configurations are proximal in both three-dimensional space and in sequence.

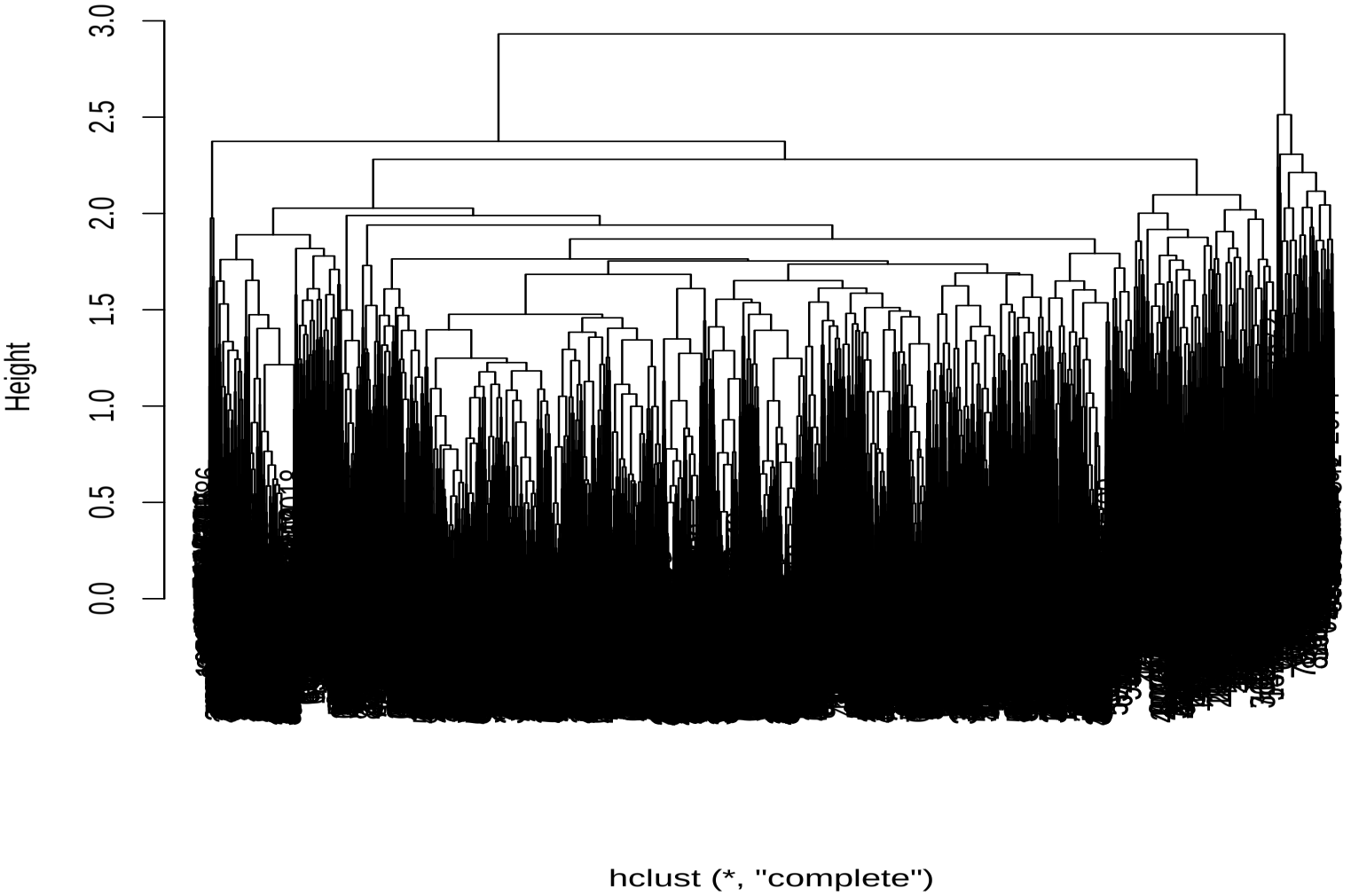


*Figure 3.1: Cluster dendrogram of configurations using complete linkage and $m = 10$.*

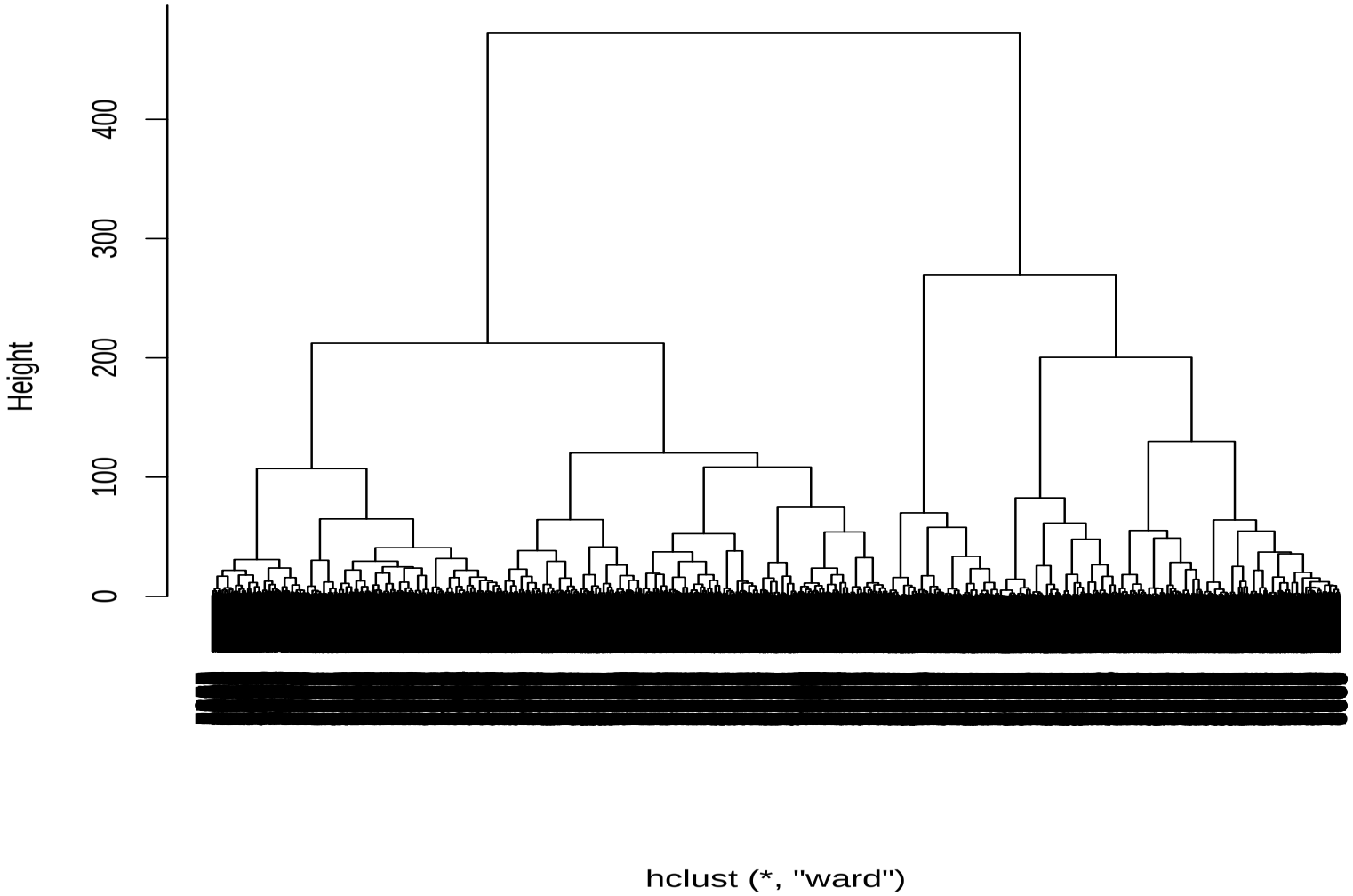


*Figure 3.2: Cluster dendrogram of configurations using Ward's linkage and* $m = 10$.

Interestingly, 6 out of the 7 configurations in this particular cluster have an index atom which is an oxygen atom (O). This is always followed by a carbonyl carbon (C), a nitrogen atom (N), and then by two alpha-carbon atoms (CA), indicating an (O-C-N-CA-CA) pattern. This pattern clearly represents a peptide bond between two adjacent amino acids. Similar patterns are observed in the remaining clusters (see Table A.1), suggesting that a constraint on the residue numbers of each configuration could provide the key to a more 'interesting' cluster solution. It is important to note that configurations from all 5 different protein structures appear in this cluster, suggesting that this particular type of configuration is visible in all chosen structures.

Applying a constraint to the database ensures that each configuration contains a set of atoms which are spatially compact but sequentially distant, i.e. the atoms in the configuration cannot simply all come from the same amino acid residue or residues which are adjacent in linear sequence. As a consequence of applying this constraint, a large proportion of configurations are discarded. This eases the pressure on computational requirements and therefore allows for an increase in the number of different protein structures used in the database.

*Table 3.1: Configurations in cluster 1.*

| Atom number | Element | Amino acid | Residue number | PDB accession number |
|---|---|---|---|---|
| 1 | N | ILE | 16 | 3tpi |
| 2 | CA | ILE | 16 | 3tpi |
| 5 | CB | ILE | 16 | 3tpi |
| 3 | C | ILE | 16 | 3tpi |
| 6 | CG1 | ILE | 16 | 3tpi |
| 4 | O | ILE | 16 | 3tpi |
| 9 | N | VAL | 17 | 3tpi |
| 8 | CD1 | ILE | 16 | 3tpi |
| 7 | CG2 | ILE | 16 | 3tpi |
| 10 | CA | VAL | 17 | 3tpi |
| 1169 | O | MET | 180 | 3tpi |
| 1168 | C | MET | 180 | 3tpi |
| 1174 | N | PHE | 181 | 3tpi |
| 1167 | CA | MET | 180 | 3tpi |
| 1175 | CA | PHE | 181 | 3tpi |
| 1166 | N | MET | 180 | 3tpi |
| 1170 | CB | MET | 180 | 3tpi |
| 1178 | CB | PHE | 181 | 3tpi |
| 1148 | O | THR | 177 | 3tpi |
| 1507 | CE | LYS | 230 | 3tpi |

Continued on next page

*Table 3.1 – continued from previous page*

| Atom number | Element | Amino acid | Residue number | PDB accession number |
|---|---|---|---|---|
| 102 | O | GLU | 13 | 3paz |
| 101 | C | GLU | 13 | 3paz |
| 108 | N | GLY | 14 | 3paz |
| 100 | CA | GLU | 13 | 3paz |
| 109 | CA | GLY | 14 | 3paz |
| 103 | CB | GLU | 13 | 3paz |
| 99 | N | GLU | 13 | 3paz |
| 890 | NH2 | ARG | 114 | 3paz |
| 110 | C | GLY | 14 | 3paz |
| 106 | OE1 | GLU | 13 | 3paz |
| 621 | O | LEU | 81 | 1aew |
| 620 | C | LEU | 81 | 1aew |
| 626 | N | PHE | 82 | 1aew |
| 619 | CA | LEU | 81 | 1aew |
| 627 | CA | PHE | 82 | 1aew |
| 618 | N | LEU | 81 | 1aew |
| 622 | CB | LEU | 81 | 1aew |
| 628 | C | PHE | 82 | 1aew |
| 124 | CG | ASN | 21 | 1aew |
| 125 | OD1 | ASN | 21 | 1aew |
| 425 | O | ARG | 61 | 1a7s |
| 424 | C | ARG | 61 | 1a7s |
| 433 | N | ARG | 62 | 1a7s |
| 423 | CA | ARG | 61 | 1a7s |
| 434 | CA | ARG | 62 | 1a7s |
| 426 | CB | ARG | 61 | 1a7s |
| 422 | N | ARG | 61 | 1a7s |
| 435 | C | ARG | 62 | 1a7s |
| 437 | CB | ARG | 62 | 1a7s |
| 427 | CG | ARG | 61 | 1a7s |
| 478 | O | GLN | 66 | 1a7s |
| 477 | C | GLN | 66 | 1a7s |
| 484 | N | SER | 67 | 1a7s |
| 476 | CA | GLN | 66 | 1a7s |
| 485 | CA | SER | 67 | 1a7s |
| 479 | CB | GLN | 66 | 1a7s |

Continued on next page

*Table 3.1 – continued from previous page*

| Atom number | Element | Amino acid | Residue number | PDB accession number |
|---|---|---|---|---|
| 475 | N | GLN | 66 | 1a7s |
| 486 | C | SER | 67 | 1a7s |
| 488 | CB | SER | 67 | 1a7s |
| 480 | CG | GLN | 66 | 1a7s |
| 1157 | O | GLN | 153 | 1a7s |
| 1156 | C | GLN | 153 | 1a7s |
| 1163 | N | CYS | 154 | 1a7s |
| 1155 | CA | GLN | 153 | 1a7s |
| 1164 | CA | CYS | 154 | 1a7s |
| 1158 | CB | GLN | 153 | 1a7s |
| 1154 | N | GLN | 153 | 1a7s |
| 1167 | CB | CYS | 154 | 1a7s |
| 1165 | C | CYS | 154 | 1a7s |
| 1481 | CG | PRO | 200 | 1a7s |

## 3.2 Hierarchical agglomerative clustering: new database

Reading in 30 different protein structures, each selected at random from the PDB, gives rise to 54,972 configurations. A constraint is then introduced to these configurations, stating that one of the first 5 atoms in each standard ordered matrix must have a difference in residue number greater than two from any of the remaining 4 atoms in the top half of the configuration. After applying the constraint the number of configurations is reduced to 4,579. Hierarchical agglomerative clustering of these configurations, again using complete linkage and Ward's linkage, can be viewed in the following dendrograms (Figures 3.3-3.4).

It should be noted that configurations, where the underlying feature of the set of atoms is a peptide bond but a single atom from an amino acid residue distant in sequence has appeared proximal to the index atom in three-dimensional space purely by chance, will still be included in the clustering. However, these anomalies are easy to identify when comparing configurations within a given cluster and hence are a negligible problem.

Cutting the dendrograms at a height of 2 generates 24 and 474 clusters for complete linkage and Ward's linkage respectively. All configurations have now had the constraint applied, meaning that any structural motifs identified will contain amino acid residues which are distant in sequence.

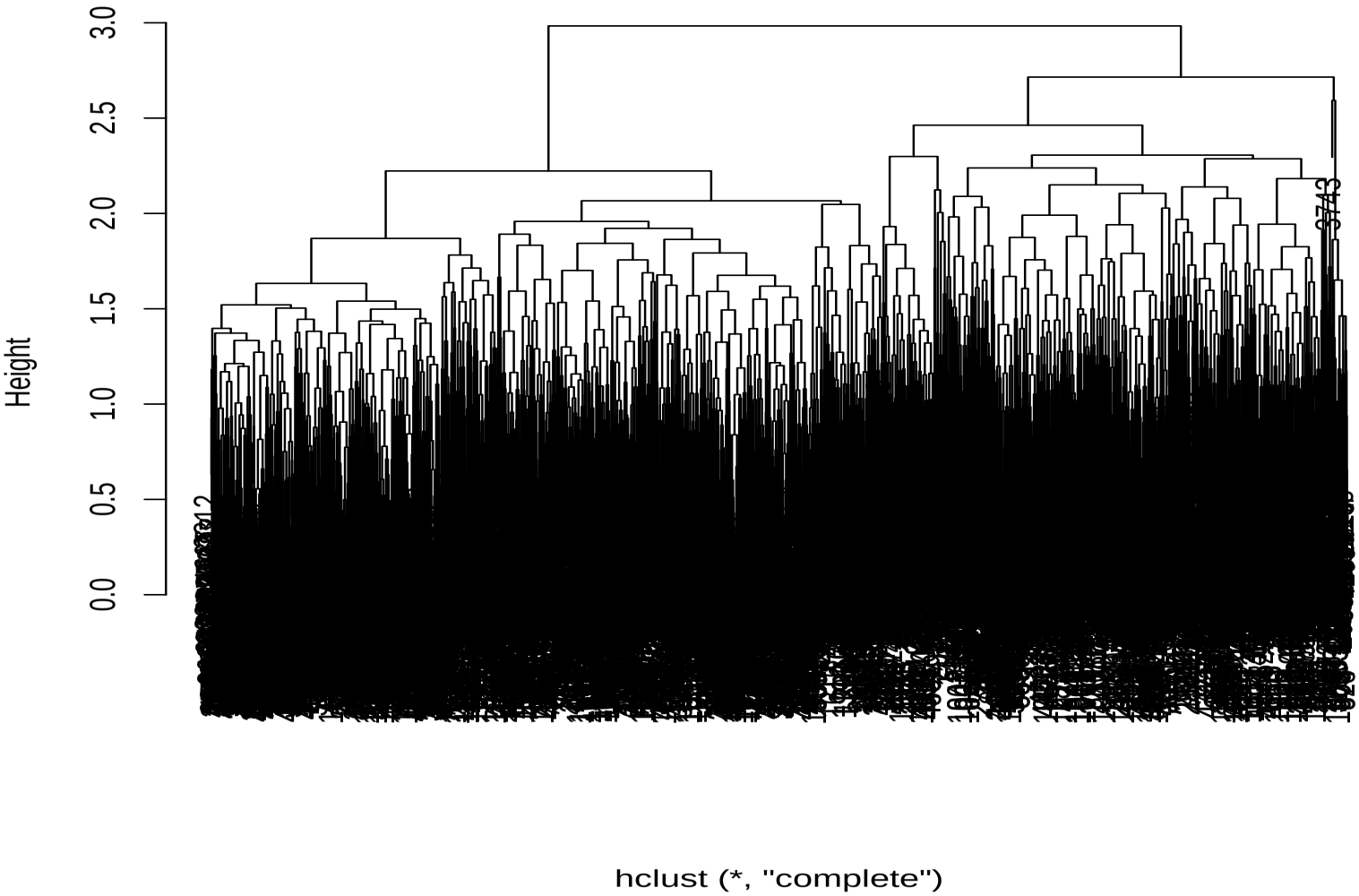


*Figure 3.3: Cluster dendrogram of new configurations using complete linkage and $m = 10$.*

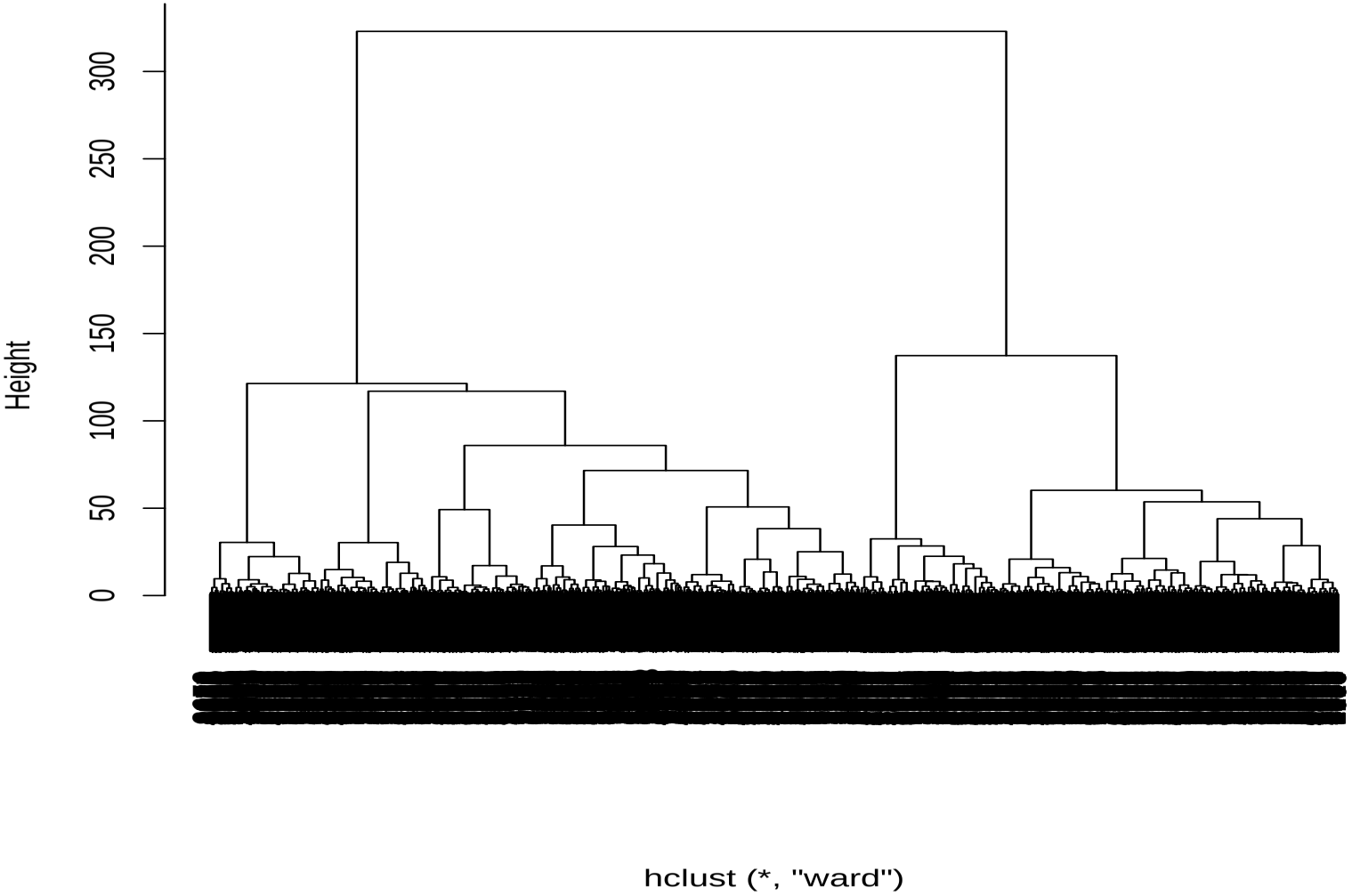


*Figure 3.4: Cluster dendrogram of new configurations using Ward's linkage and* $m = 10$.

The first thing to note is that there is a large difference in the number of clusters for each clustering algorithm; far fewer clusters are observed when complete linkage (Figure 3.3) is used when the dendrograms are cut at the same height. Fusions again appear to occur around the same height making it difficult to differentiate between clusters and hence there appears to be no explicit solution. On the other hand, Ward's linkage (Figure 3.4) appears to be more beneficial than complete linkage, since the algorithm groups the configurations clearly into distinct clusters. To further explore these groupings in order to help identify configurational motifs, the cluster membership of each configuration is observed. For the hierarchical agglomerative clustering algorithm using Ward's linkage, the number of configurations in each cluster ranges from between 1 and 30, averaging just less than 10 configurations per cluster. The configurations associated with the largest cluster appear to reveal no systematic patterns, indicating that this particular cluster does not represent a specific structural motif.

*Table 3.2: Configurations in cluster 17.*

| Atom number | Element | Amino acid | Residue number | PDB accession number |
|---|---|---|---|---|
| 1521 | SG | CYS | 232 | 3tpi |
| 1520 | CB | CYS | 232 | 3tpi |
| 811 | SG | CYS | 128 | 3tpi |
| 1517 | CA | CYS | 232 | 3tpi |
| 810 | CB | CYS | 128 | 3tpi |
| 1518 | C | CYS | 232 | 3tpi |
| 1516 | N | CYS | 232 | 3tpi |
| 1522 | N | ASN | 233 | 3tpi |
| 1506 | CD | LYS | 230 | 3tpi |
| 807 | CA | CYS | 128 | 3tpi |
| 1932 | SG | CYS | 38 | 3tpi |
| 1931 | CB | CYS | 38 | 3tpi |
| 1740 | SG | CYS | 14 | 3tpi |
| 928 | CA | CYS | 38 | 3tpi |
| 1739 | CB | CYS | 14 | 3tpi |
| 1929 | C | CYS | 38 | 3tpi |
| 1930 | O | CYS | 38 | 3tpi |
| 1927 | N | CYS | 38 | 3tpi |
| 1735 | N | CYS | 14 | 3tpi |
| 608 | CD2 | LEU | 99 | 3tpi |

### Motif one

However, some clusters do show signs of systematic patterns, such as the configurations grouped in cluster 17. Table (3.2) displays the atoms involved in the first two configurations in this cluster. The first seven atoms in both configurations belong to cysteine residues. Moreover, both configurations have a sulphur atom as the index atom and a beta carbon as the closest atom to the index atom in three-dimensional space. The 3rd, 4th, 5th and 6th positions in the configurations contain another sulphur atom, an alpha carbon, a beta carbon and a carbonyl carbon, respectively. Four of the atoms, those in the 1st, 2nd, 4th and 6th positions, all belong to the same cysteine residue. Atoms in the 3rd and 5th positions are part of another cysteine residue which is distant in sequence to the first residue. For example, in the first configuration the 1st, 2nd, 4th and 6th atoms are found in residue number 232, whereas atoms in the 3rd and 5th position are from residue number 128. Since residues are numbered in linear sequence according to the backbone of the polypeptide chain, it is immediately apparent that these two cysteine residues are non-local with respect to sequence.

This pattern is clearly visible for all remaining configurations within this cluster (see Table A.2). Hence this particular cluster contains a group of configurations which represent a set of atoms which are spatially compact but not adjacent in amino acid sequence. A similar pattern is observed in cluster 99, although these two clusters do not merge until the dendrogram reaches a height of over 300, i.e. the clusters do not merge until the final agglomeration of all configurations in the clustering. The possible reason for this is discussed in Chapter 4.

### 3.2.1 Principal coordinate analysis: motif one

Recall that principle coordinate analysis provides a method for reducing the dimensionality of a set of objects, allowing for easier comparison between objects which depend on multiple variables. The first 6 atoms of each configuration from cluster 99 are plotted in two dimensions (Figure 3.5). The points are defined up to rotation, reflection and location and hence distance preserving similarity transformations can be applied to the configurations to allow observation of any structural similarity. Each colour represents a different configuration within the cluster. The lines between the points indicate the presence of covalent bonds joining the atoms together. Recall that atoms 3 and 5 are from a different cysteine residue from the remaining atoms in the plot. There appears to be a degree of similarity between each of the configurations; however, it is difficult to determine the statistical significance of this similarity.

In order to apply shape analysis techniques it becomes a much simpler problem if only the first 5 atoms of each configuration are considered. Figure (3.6) shows the PCoA coordinates of the first 5 atoms from each configuration in cluster 17. The lines have now been uniformly ordered to allow for a representation of the shape of each configuration and thus no longer represent covalent bonds. The same method is applied to the configurations found in cluster 99 (Figure 3.7). It is interesting to note that one of the configurations in cluster 99 has a very similar shape to the mean shape of the configurations found in cluster 17 when PCoA is applied. The

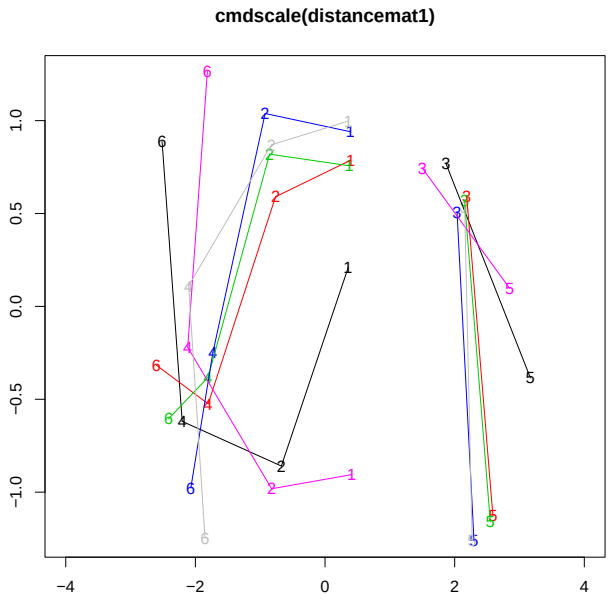


*Figure 3.5: PCoA of first 6 atoms of configurations in cluster 99.*

mean shape of each set of configurations is shown in Figure (3.8). The main difference between the two configurations appears to involve the interaction between the two cysteine residues. There appears to be relatively little difference in the cysteine residue containing the 1st, 2nd and 4th atoms from each configuration. However, the difference lies in the cysteine residue which is distant in linear sequence to this first residue but contributes the 3rd and 5th atoms of each configuration. The distance between the two sulphur atoms (1st and 3rd atoms) is much greater for the configurations in cluster 17. Also there is a large difference in the positioning of the 5th atom relative to the 3rd atom between both mean shapes.

### 3.2.2 Hotelling's $T^2$ test: motif one

We can test whether the configurations in cluster 17 have a significantly different shape to the configurations in cluster 99 using Hotelling's $T^2$ test. First, Procrustes tangent coordinates are calculated for both sets of configurations about the GPA estimate of mean shape for the combined dataset. Both the difference between the mean shapes and the covariance matrix of all configurations from both clusters are calculated and then used to calculate Hotelling's $T^2$ test statistic.

The dimension of the shape space, $M$, is equal to 8. Each configuration consists of 5 atoms in 3 dimensions giving rise to 15 dimensions. However, 3 are removed for location, 3 for orientation and 1 for scaling, resulting in a shape space of 8 dimensions. There are 5 configurations in each cluster, thus $n_x = 5$ and $n_y = 5$.

A test statistic of $T^2 = 128.03 \sim T^2(8, 8)$ is obtained, with a $p$-value of 0.5003. Since the $p$-value is much greater than 0.05, the null hypothesis, $H_0$: the two sets of configurations have equal mean shapes, cannot be rejected. This is interesting because not only is a large test statistic produced but the PCoA plots (Figures 3.6-3.8) appear to show a reasonable difference in shape between the two sets of configurations. The sample sizes are both extremely small as there are only 5 objects in each group. As a result of these small sample sizes, Hotelling's $T^2$ test is not very powerful.

Goodall's $F$ test (Goodall, 1991) can also be used to test for a significant difference between the two mean shapes. This test uses a diagonal covariance matrix, where the diagonal elements are all equal and proportional to the identity matrix. Implementing Goodall's $F$ test results in the same conclusion as reached above, i.e. the null hypothesis is accepted.

Increasing the number of structures in the database directly increases the number of configurations available. By reading in four extra protein structures, $n_x$, from above, increases from 5 to 9 configurations. With the number of configurations remaining the same in the other cluster, i.e. $n_y = 5$, Hotelling's $T^2$ test statistic becomes $T^2 = 127.48 \sim T^2(8, 12)$. The $p$-value is now 0.0259, which is less than the critical threshold value of 0.05. The null hypothesis is rejected and hence there is evidence to suggest that there is a statistically significant difference in mean shapes between the two clusters of configurations.

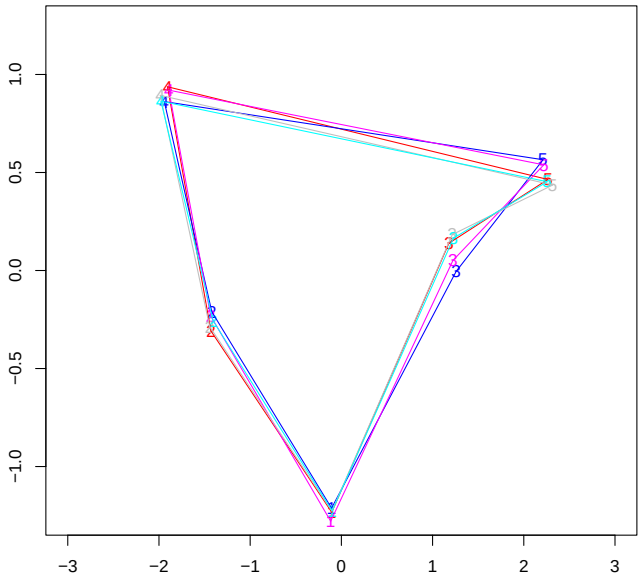


*Figure 3.6: PCoA of configurations in cluster 17.*

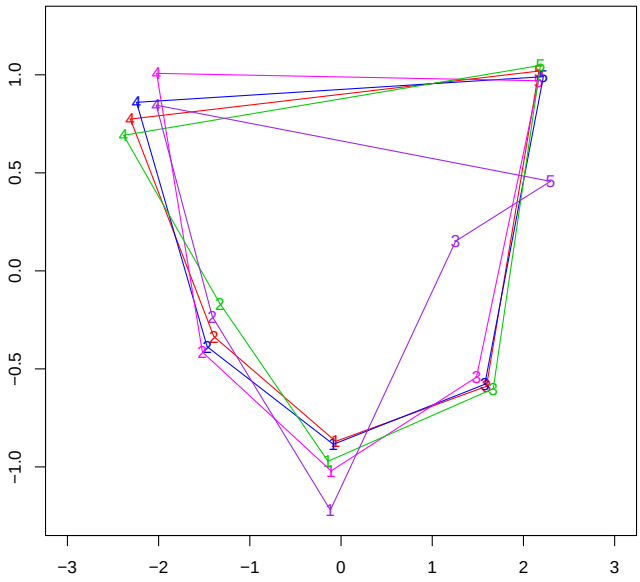


*Figure 3.7: PCoA of configurations in cluster 99.*

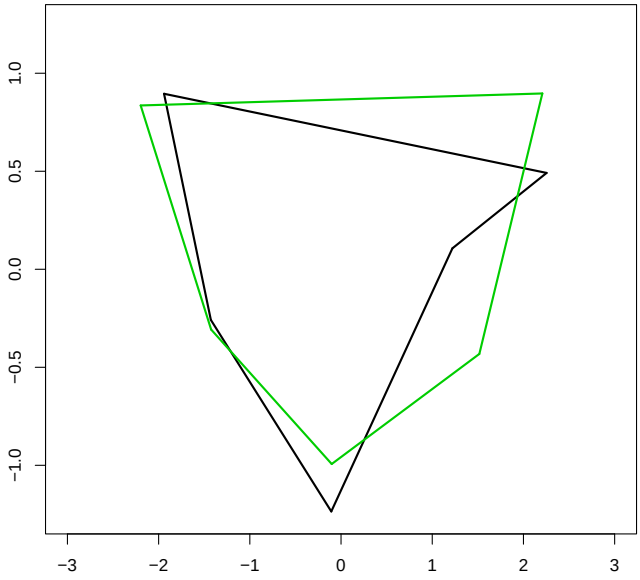


*Figure 3.8: PCoA of mean configurations of cluster 17 (black) and cluster 99 (green).*

*Table 3.3: Configurations in cluster 396.*

| Atom number | Element | Amino acid | Residue number | PDB accession number |
|---|---|---|---|---|
| 1236 | OD2 | ASP | 157 | 16gs |
| 1234 | CG | ASP | 157 | 16gs |
| 1235 | OD1 | ASP | 157 | 16gs |
| 1233 | CB | ASP | 157 | 16gs |
| 784 | NH2 | ARG | 100 | 16gs |
| 783 | NH1 | ARG | 100 | 16gs |
| 782 | CZ | ARG | 100 | 16gs |
| 1231 | C | ASP | 157 | 16gs |
| 1230 | CA | ASP | 157 | 16gs |
| 1232 | O | ASP | 157 | 16gs |
| 2163 | OD2 | ASP | 42 | 1k3y |
| 2161 | CG | ASP | 42 | 1k3y |
| 2162 | OD1 | ASP | 42 | 1k3y |
| 2160 | CB | ASP | 42 | 1k3y |
| 3630 | NH2 | ARG | 221 | 1k3y |
| 3629 | NH1 | ARG | 221 | 1k3y |
| 3628 | CZ | ARG | 221 | 1k3y |
| 2129 | O | ALA | 38 | 1k3y |
| 2157 | CA | ASP | 42 | 1k3y |
| 2156 | N | ASP | 42 | 1k3y |

### Motif two

Table (3.3) shows two of the configurations from cluster 396. The first four atoms in both configurations are all from aspartic acid residues. Moreover, the type of element of each of the heavy atoms is the same for corresponding positions in each cluster. For example, the first atom is an oxygen, the second atom is a gamma carbon, and so on. The atom in the 5th position of both configurations is a nitrogen from an arginine residue. This residue is clearly distant in linear sequence from the aspartic acid residue seen in the first four positions of both configurations since the residue numbers are very different from one another. The remaining configurations in this cluster resemble a similar pattern, with one of configurations containing a lysine residue in the 5th position instead of arginine. Hence, this cluster appears to show configurations which are structurally compact but sequentially distant.

The whole database can then be searched to locate all configurations with this pattern. The results obtained from this search can then be tested to see whether or not, after the constraint dis-

cussed earlier is applied, the number of configurations with this pattern is significantly different from the expected number of results, under the null model that the amino acid residue appearing after 4 aspartic acid atoms is random. A chi-squared test of independence can be implemented with the purpose of testing the significance of this observed pattern.

### 3.2.3 Chi-Squared test of independence: motif two

Setting one variable to be the amino acid residue in the 5th position of each configuration and the other variable to be the first four atoms in each corresponding configuration, a chi-square test of independence can be carried out on these results to formally test whether the two categorical variables are independent of one another, i.e. test whether or not the atom in the 5th position of each configuration depends significantly on the first four atoms. The idea is to show that the amino acid residues *arginine* and *lysine* occur more frequently than expected given that four *aspartic acid* atoms are observed in the first four positions of the configuration. The null and alternative hypotheses are set out as follows

$H_0$ = the amino acid residue appearing in the 5th position of the configuration is independent of the atoms which appear before it,

$H_A$ = the two categorical variables are related.

| | Arginine | Lysine | Threonine | Other | Row total |
|---|---|---|---|---|---|
| 4 Aspartic acid atoms | 69 | 65 | 33 | 186 | 353 |
| Any other combination | 342 | 425 | 306 | 3153 | 4226 |
| Column total | 411 | 490 | 339 | 3339 | 4579 |

*Table 3.4: $\chi^2$-test of independence: observed frequencies.*

Table (3.4) shows the observed frequencies of 8 different events for 4,579 configurations. The column totals indicate the frequency of each amino acid occurring in the 5th position of a configuration, where 'other' denotes the remaining 17 amino acids. The first row total corresponds to the number of configurations where four atoms from aspartic acid residues are present in the first four positions. The second row total accounts for any other sequence of amino acids in the first four positions.

The expected frequencies, $E_{ij}$, of each event can be calculated using the formula

$$E_{ij} = \frac{\sum_{t=1}^{c} O_{it} \sum_{t=1}^{r} O_{tj}}{N},$$

where $N$ denotes the total number of observed events, $c$ and $r$ are the number of columns and rows, respectively, in the contingency table. $O_{it}$ denotes the sum of the observed frequencies of all cells in row $i$ and $O_{tj}$ denotes the sum of the observed frequencies of all cells in column $j$. Table (3.5) shows the expected frequencies for each event. The cells which represent configurations which have four atoms from aspartic acid residues in the first four positions followed

by either an atom from a arginine or lysine residue have much larger values for the observed frequencies compared to the expected frequencies. In fact, 69 configurations with atoms from aspartic acid residues occupying the first four positions, followed by an atom from an arginine residue, are observed, which is over double the expected frequency of this event (31.7). Threonine is another amino acid which has a larger observed frequency than expected frequency; however the difference in observed and expected frequency is not as great as seen with arginine and lysine. On the other hand, the converse is true for the remaining 17 amino acids, with the table indicating that there is a much larger expected frequency than observed frequency for this particular event.

The opposite is true for any other combination of amino acids for all cells, although the absolute differences between observed and expected frequencies appear to be much smaller relative to the large expected frequencies.

| | Arginine | Lysine | Threonine | Other | Row total |
|---|---|---|---|---|---|
| 4 Aspartic acid atoms | 31.7 | 37.8 | 26.1 | 257.4 | 353 |
| Any other combination | 379.3 | 452.2 | 312.9 | 3081.6 | 4226 |
| Column total | 411 | 490 | 339 | 3339 | 4579 |

*Table 3.5: $\chi^2$-test of independence: expected frequencies.*

Table (3.6) shows the calculations required for the test statistic. The fourth column of the table shows the sign of each contribution so that it can easily be seen whether the observed frequencies are greater than, or less than, the expected frequencies of each cell.

| Observed frequency | Expected frequency | $(O-E)^2/E$ | $(O-E)/\sqrt{E}$ |
|---|---|---|---|
| 69 | 31.7 | 44.0 | 6.6 |
| 65 | 37.8 | 19.6 | 4.4 |
| 33 | 26.1 | 1.8 | 1.3 |
| 186 | 257.4 | 19.8 | -4.5 |
| 342 | 379.3 | 3.7 | -1.9 |
| 425 | 452.2 | 1.6 | -1.3 |
| 306 | 312.9 | 0.2 | -0.4 |
| 3153 | 3081.6 | 1.7 | 1.3 |

*Table 3.6: $\chi^2$-test of independence: test statistic.*

The test statistic is calculated as follows

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}},$$

where $O_{ij}$ denotes the observed frequency of cell $(i,j)$ and $E_{ij}$ denotes the expected frequency of cell $(i,j)$.

Summing over the rows and columns gives

$$\begin{aligned}\chi^2 = &= 44.0 + 19.6 + 1.8 + 19.8 + 3.7 + 1.6 + 0.2 + 1.7 \\ &= 92.4.\end{aligned}$$

The $\chi^2$-test has $(c-1)(r-1) = (4-1)(3-1) = 3$ degrees of freedom and thus the critical value of interest is $\chi^2_3(0.05) = 7.815$. Since $\chi^2_{obs} >> \chi^2_3(0.05)$, the null hypothesis, $H_0$: the amino acid residue appearing in the 5th position of the configuration is independent of the atoms which appear before it, is rejected. The test shows that there is a statistically significant association between the atom which appears in the 5th position of a configuration and the atoms which appear in the first four positions. Clearly there are a greater number of configurations with atoms from arginine and lysine residues in the 5th position than expected when the first four atoms are all from aspartic acid residues. The possible reasoning behind this large discrepancy is discussed in detail in Chapter 4.

*Table 3.7: Configuration pattern in cluster 359.*

| Atom number | Element | Amino acid | Residue number | PDB accession number |
|---|---|---|---|---|
| 777 | CE | LYS | 107 | 2abh |
| 778 | NZ | LYS | 107 | 2abh |
| 776 | CD | LYS | 107 | 2abh |
| 775 | CG | LYS | 107 | 2abh |
| 1149 | OE1 | GLU | 155 | 2abh |
| 774 | CB | LYS | 107 | 2abh |
| 1150 | OE2 | GLU | 155 | 2abh |
| 1148 | CD | GLU | 155 | 2abh |
| 736 | OD1 | ASP | 102 | 2abh |
| 771 | CA | LYS | 107 | 2abh |

### Another systematic pattern

Cluster 359 also represents a recurring pattern. Table (3.7) shows an example of one of the configurations found in this cluster. The first four atoms, an epsilon carbon, nitrogen, delta carbon and a gamma carbon, all come from the same lysine residue. An oxygen atom from a glutamic acid residue, which is distant in linear sequence to the lysine residue, is situated in the 5th position. In total there are 9 configurations in cluster 359. Two of these configurations contain an atom which is assigned the first position, i.e. the index atom, in one of the other configurations

within this cluster, but in one of the remaining 9 positions. Hence only 7 structural motifs can be considered as independent, since 2 of the configurations are essentially repetitions of the same motif, differing only by the choice of index atom within the motif. Of these 7 configurations, 6 have the same pattern as seen in the table. The only difference in the remaining configuration is that an oxygen atom from an aspartic acid residue is present in the 5th position. This particular cluster therefore represents a set of atoms which are proximal in three-dimensional space but are not local in amino acid sequence.

Table (3.8) shows that a similar pattern is observed in cluster 426; the only difference being that an oxygen atom from an aspartic acid residue is found in the 5th position after four atoms from a lysine residue. In this cluster there is one configuration with a glutamic acid residue in the 5th position, suggesting that clusters 359 and 426 contain structural motifs which are similar to one another. Again, this topic is discussed further in Chapter 4.

*Table 3.8: Configuration pattern in cluster 426.*

| Atom number | Element | Amino acid | Residue number | PDB accession number |
|---|---|---|---|---|
| 133 | CE | LYS | 16 | 104m |
| 134 | NZ | LYS | 16 | 104m |
| 132 | CD | LYS | 16 | 104m |
| 131 | CG | LYS | 16 | 104m |
| 977 | OD2 | ASP | 122 | 104m |
| 130 | CB | LYS | 16 | 104m |
| 975 | CG | ASP | 122 | 104m |
| 99 | NE2 | HIS | 12 | 104m |
| 976 | OD1 | ASP | 122 | 104m |
| 98 | CE1 | HIS | 12 | 104m |

### 3.2.4 Principal coordinate analysis: motif two

Principle coordinate analysis can be applied to both of these sets of configurations to produce a two-dimensional plot of the first five atoms. Figures (3.9-3.10) shows these PCoA plots for configurations in cluster 359 and 426, respectively. In both plots, the points in positions 1-4 represent atoms from the same amino acid residue, whereas point 5 represents an oxygen atom from a different residue which is distant in sequence. The first four atoms in all configurations form part of a lysine residue. The difference between the two figures is that in figure (3.9) the atom in the 5th position is from a glutamic acid residue, whereas the corresponding atom in figure (3.10) is part of an aspartic acid residue. Despite this difference, both of these atoms appear to be positioned in such a way that when all five atoms of each configuration are joined,

the shapes of the configurations in both plots are very similar to one another. Figure (3.11) shows the means shape for both sets of configurations.

### 3.2.5 Hotelling's $T^2$ test: motif two

Again a Hotelling's $T^2$ test is implemented to test the null hypothesis, $H_0$: the two mean shapes are equal. A test statistic of $T^2 = 72.02$ is obtained, with a $p$-value of 0.3756, suggesting that there is no evidence to reject the null hypothesis. Therefore, as expected, it is concluded that the two mean shapes are equal.

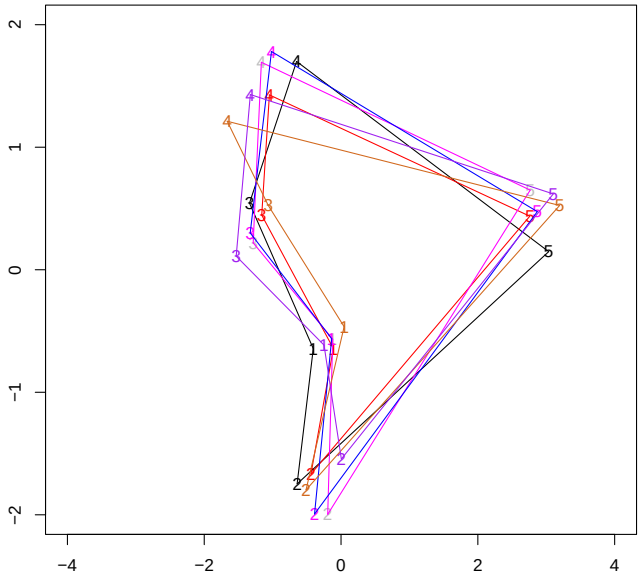


*Figure 3.9: PCoA of configurations in cluster 359.*

### 3.2.6 Other results

Apart from the results shown so far, none of the remaining clusters appear to show any systematic patterns, i.e. there appears to be no obvious relationship between the configurations grouped in any other cluster not mentioned above, suggesting that the methodology used here is somewhat unsatisfactory. The strengths and limitations of the methodology are discussed in Chapter 4 with the implications of possible future work reviewed in the final chapter.

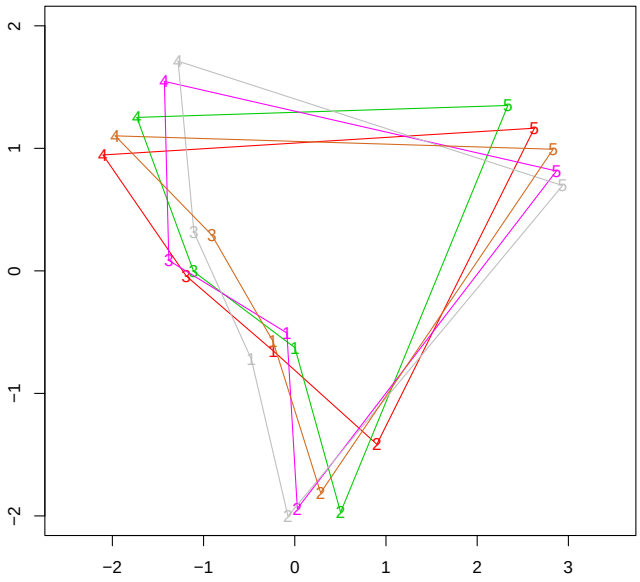

*Figure 3.10: PCoA of configurations in cluster 426.*

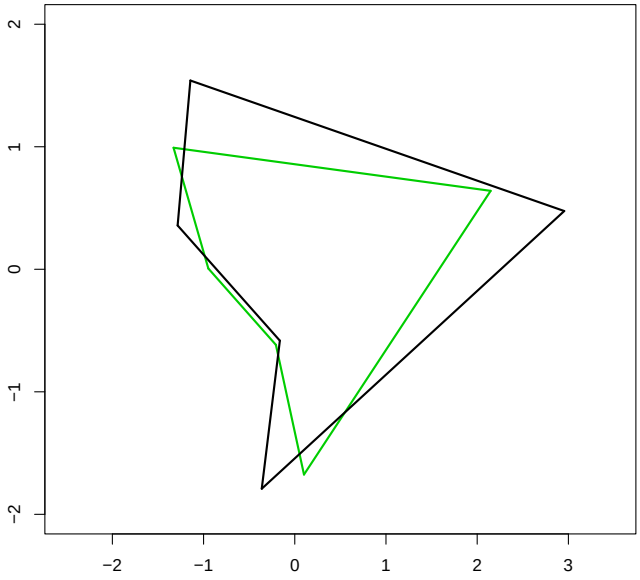

*Figure 3.11: PCoA of mean configurations of cluster 359 (black) and cluster 426 (green).*

# Chapter 4

# Discussion

This project presents a method for identifying structural motifs in protein sequences using exploratory data analysis techniques. The methodology is unique and is designed to reveal systematic patterns in atomic configurations using the three-dimensional coordinates of atoms from protein structures in the PDB. The PDB is a highly regarded repository for experimentally determined protein structures, adding to the credibility of the study.

## 4.1 Results

### 4.1.1 Original database

The results from the hierarchical agglomerative clustering of the original database showed that Ward's linkage is a more appropriate choice of inter-group proximity measure than complete linkage, although both measures tend to form reasonably similar clusters here. The choice of measure is not a fundamental issue and as a result any future work adding to this project should not completely disregard complete linkage when performing cluster analysis on configurations. The results of the first cluster analysis show sets of configurations which have been grouped together due to the positioning of a peptide bond. Recall that a peptide bond constitutes six atoms; one of which is a hydrogen atom. Since hydrogen atoms are disregarded from the database, each peptide bond contains five heavy atoms from the database. The initial results show that the configurations are clustered depending on which of these five atoms is assigned to be the index atom. This is because each peptide bond is a rigid 6-atom structure, meaning that the distance between atoms within the bond remain the same in all protein structures (Figure 1.4). Since the metric used for the cluster analysis depends on these differences in Euclidean distances between corresponding atoms in different configurations, it is no surprise that the intransigent feature of the peptide bond is so influential with regards to the clustering solution.

The reason that not all configurations with the same type of index atom element are clustered together relates to the fact that despite the rigid structure of the peptide bond, these bonds can rotate by an angle $\psi$, about the alpha carbon-carbonyl carbon bond, and an angle $\phi$, about the

alpha carbon-nitrogen bond (Figure 4.1)[6]. Depending on the values of $\psi$ and $\phi$, different atoms will be brought into spatial proximity of the index atom, and as a result, the configurations can differ from one another. Inevitably, setting the number of atoms in each configuration equal to five increases the influence of the peptide bond on the clustering solution.

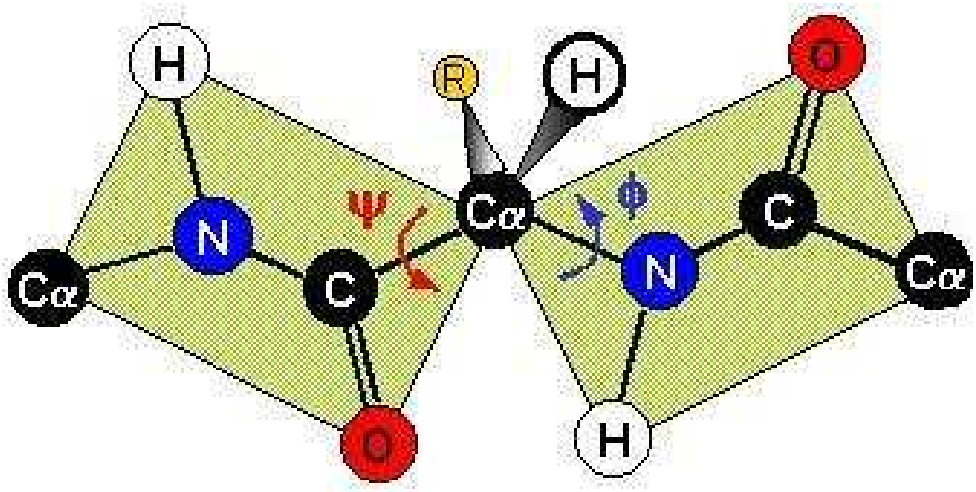


*Figure 4.1: Peptide bond with torsion angles - $\psi$ and $\phi$.*

### 4.1.2 New database

These results are, of course, of little interest, and hence a constraint is required to guarantee that atomic configurations contain atoms which are proximal in three-dimensional space but from residues which are distant in linear sequence. The results from hierarchical agglomerative clustering of the new dataset containing configurations restricted to the constraint, on the other hand, are much more promising. Moreover, the number of protein structures contained in the database is increased. This is because the process of discarding 'uninteresting' configurations eases the pressure on computational requirements, therefore allowing a greater volume of data to be analysed.

### 4.1.3 Motif one

Again, Ward's linkage is preferred over complete linkage since the former produces a much more distinct and regulated clustering solution. One of the most prominent patterns observed from the clustering solution is the presence of two cysteine residues, which are consistently far apart in linear sequence, but brought into spatial proximity in three-dimensional space by the folding of the protein. Two separate clusters both contain this configurational motif, which essentially represents the presence of disulphide bonds. Branden & Tooze (1999) state that disulphide bonds can only occur between the sulphur atoms of two cysteine residues. These

[6]http://employees.csbsju.edu/HJAKUBOWSKI/classes/ch331/protstructure/

cysteine residues are distant in sequence but are brought into spatial proximity by the disulphide bonds, which help stabilise the structure of the protein (Sevier & Kaiser, 2002). As a consequence of this, disulphide bonds play an important role in the folding and stability of proteins containing cysteine residues. Although this result is not new to the field of bioinformatics, it does show that the methodology used here works to a high standard. This suggests that to detect other more subtle motifs, future work needs to be directed at creating a more astute metric to describe the dissimilarity between two configurations.

Since not all protein structures contain cysteine residues, only a proportion of structures contain disulphide bonds. Hence this configurational motif is not present in a large number of structures. Despite this, the motif is visible in two separate clusters. The reason for this separation is initially unclear, however the results from principal coordinate analysis and shape analysis shed some light into the differences between the two sets of configurations.

### Hotelling's $T^2$ test

The mean shapes of both sets of configurations appear to be significantly different from each other and hence the null hypothesis, $H_0$: the two mean shapes are equal, is likely to be rejected. The value of the test statistic is large but is not large enough to reject the null hypothesis. The fact that $H_0$ is accepted is most likely to have arisen because of the small sample sizes used in the test. Recall that there are only five configurations in each group, the minimum number required to be able to construct the test statistic. Clearly, this is a case where although statistical inference can be carried out in the appropriate tangent space, the sample size is simply too small, resulting in a test which is not very powerful. Dryden & Mardia (1993) note that this is often a problem generated by a tangent space which is over-dimensioned and therefore a Hotelling's $T^2$ test will only be powerful given that there are a large number of observations available.

Dryden & Mardia (1993) also remark that the Hotelling's $T^2$ test is less powerful than Goodall's $F$ test (Goodall, 1991), when the isotropic normal model holds. Recall that Goodall's $F$ test does not estimate the off-diagonal elements of the covariance matrix, whereas Hotelling's $T^2$ test does. Hence the reason for this loss of power in the Hotelling's $T^2$ test is explained by the increase in the number of degrees of freedom used in estimating the covariance matrix. Despite this, both Hotelling's $T^2$ test and Goodall's $F$ test accept the null hypothesis that the two mean shapes are the same.

An obvious way to overcome the sample size problem is to increase the number of protein structures in the database. Increasing the number of structures will, of course, increase the number of configurations used in the cluster analysis. Disulphide bonds from between the thiol groups of cysteine residues, however not all protein structures contain cysteine residues (Sevier & Kaiser, 2002). Recall that methionine residues contain sulphur atoms, but these cannot form disulphide bonds. This suggests that only by increasing the number of protein structures containing cysteine residues will the sample size of the configurations in cluster 17 and 99 be increased. Ideally all structures in the PDB could be used; however, this is not possible here.

### Increasing size of database

By running in four extra structures containing cysteine residues, the number of configurations, in what was previously known as cluster 17, increases to 9 observations, whereas the number remains the same in the cluster previously known as cluster 99, i.e. $n_x = 9$ and $n_y = 5$. The Hotelling's $T^2$ test statistic was then shown to be large enough to reject the null hypothesis. The two mean shapes are therefore significantly different, helping to explain why the configurations are in different clusters, which don't actually join until the final agglomerations, despite representing the same type of bond.

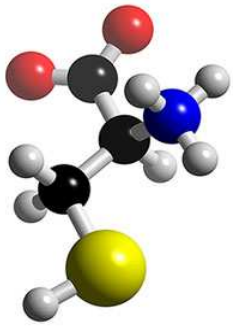

*Figure 4.2: Three-dimensional image of a cysteine residue.*

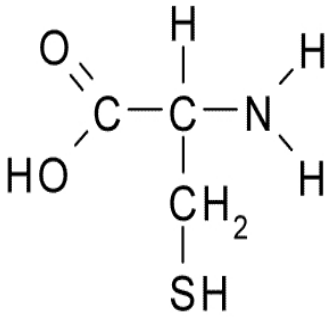


*Figure 4.3: Chemical structure of a cysteine molecule.*

Figures (4.2)[7]-(4.3)[8] show both the three-dimensional and chemical structure of a cysteine residue. The large yellow atom is a sulphur atom and the disulphide bond forms between this and a corresponding sulphur atom in another cysteine residue.

### 4.1.4 Motif two

The configurational motif discussed above is not the only recurring pattern within the database. Configurations involving two oppositely charged amino acid residues distant in sequence also appear in close proximity in three-dimensional space. Recall that aspartic acid and glutamic acid are negatively charged, whereas lysine and arginine are positively charged. Any combination of two oppositely charged residues provides a systematic pattern prominent in the clustering solution. The type of configurational motif observed in these situations is shaped by the presence of a salt bridge. Recall that a salt bridge is essentially a special type of hydrogen bond, and hence it is interesting to note that despite hydrogen atoms being disregarded from the database, these configurational motifs remain identifiable.

[7] http://www.3dchem.com/imagesofmolecules/Cysteine.jpg
[8] http://upload.wikimedia.org/wikipedia/commons/5/5e/L-Cysteine.png

The results from the chi-squared test of independence show that the configurational motifs, comprising of an aspartic acid residue and either an arginine or lysine residue, appear in two clusters. There is no evidence, however, to suggest that the main shapes of the two sets of configurations are significantly different to each other, despite the configurations forming two reasonably distinct clusters. Perhaps this is due to the small sample size of both groups, or possibly because the differences are subtle enough to be registered by cluster analysis if the dendrogram is cut at a low height, as it was here. Again, these motifs are already recognised by the field of bioinformatics, however, these results reinforce the astuteness of the methodology.

The chi-squared test of independence suggests that a possible configurational motif may exist which compromises of an aspartic acid and a threonine residue distant in amino acid sequence. However, there does not appear to be sufficient evidence to suggest that this observed pattern is non-random and hence further work must be applied to test for the presence of this particular motif. It is important to note that although no new discoveries have been made here, the potential for exploration is huge. Further work relating to how this can achieved is discussed in Chapter 5.

## 4.2 Limitations

There are, of course, several limitations to the approaches used in this project. Firstly, the methodology states that hydrogen atoms should be disregarded from the database because these atoms are frequently unreported or inaccurately recorded in the PDB. However, hydrogen atoms account for approximately half of all atoms in protein structures. Therefore, by ignoring such a large proportion of atoms from the database, a large volume of data is lost. Moreover, there could be a distinct possibility that hydrogen atoms play a fundamental role in being able to detect configurational motifs in protein structures. This would mean that the inclusion of these atoms could potentially be crucial in future work. The PDB is notoriously one of the most credible protein databases available internationally, yet the number of protein structures containing the atom coordinates of hydrogen atoms is very small. Furthermore, those recorded are often calculated using imputation and hence carry a large degree of uncertainty. Including hydrogen atoms as part of the configurations has great potential for future work but obviously this will depend highly on the availability of such data.

One of the main limitations of this project is the effect of computational requirements on the size of the database. Due to the large number of protein structures available in the PDB, there is an enormous volume of data available when considering that each atom from each structure constitutes one configuration. This project has been constrained to observing only a small proportion of the total number of known protein structures, clearly restricting the success of the methodology. The size of the database could easily be increased by a combination of using additional computational power as well as reducing the number of calculations required by the programs used here. This could make systematic patterns more accessible.

It is important to consider the point made earlier that, although increasing, the number of experimentally determined protein structures is actually very small in comparison to the number of known protein sequences. This could suggest that the experimentally determined structures, in general, might not be representative of all protein structures. This point is highlighted by the fact that the set of solved structures, i.e. those contained in the PDB, might be biased towards certain types of protein. One possibility is that smaller proteins are easier to solve and hence there may exist a larger proportion of smaller proteins in the PDB. It was already stated in Chapter 1 that there are fewer membrane proteins than globular proteins in the PDB, since the latter are, in general, easier to crystallise than the former, resulting in more globular proteins being determined experimentally by x-ray crystallography techniques. Since configurational motifs are likely to depend heavily on heavily on the type of protein observed, it is hoped that generalised motifs are available to apply to all protein structures. These structural motifs could then perhaps be used to decrease the time requirements for experimental determination of structures, resulting in a dramatic increase in the number of known protein structures. Any recurring patterns then observed could be used to help predict the structure of previously unknown protein structures.

### 4.2.1 Potential problems with the methodology

It should be noted that the methodology used in this project forms an elementary approach to detecting configurational motifs in protein structures. There are many scenarios where the methodology used becomes unsatisfactory. The next part of this discussion summaries a few of these scenarios and outlines possible solutions to these problems.

Consider the case where two configurational motifs are considered to be the same, i.e. the distances between corresponding atoms in each configurations are almost identical, and hence there is very little difference between the two distance matrices. Now, suppose there is an atom which happens to be situated somewhere in or around one of the recurring motifs but is not actually a part of it, i.e. the atom is *interloping*. Then the distance matrix will be affected since this interloping atom is likely to be represented in this distance matrix. If some prior knowledge is available, stating that this particular atom is close to the configurational motif purely by chance and is not actually part of the motif, then ideally this atom should be ignored and as a consequence disregarded from the configuration.

Suppose the configurational motif consists of 8 atoms say, but the interloping atom is located next to one of the central atoms within this structural motif. One approach to avoid the aforementioned problem could be to set $m = 9$ and then search for a subset of this $9 \times 9$ matrix. Obviously the known 8 atom configurational motif will become visible when the interloping atom is disregarded, suggesting that this particular motif could still be clustered amongst recurring motifs of a similar nature, despite having the interloping atom in the centre of the motif. It is important to note that all atoms will become the index atom at some point due to the format of the methodology. Thus the configurational motif will be present providing that only one

atom, which does not form part of the motif, is in the configuration. Obviously this idea can be generalised to be in a form that allows any proportion of the $m$ atoms in each configuration to be disregarded.

Suppose that a configurational motif consists of 8 atoms, say. One might believe that each time any one of these 8 atoms is chosen to be the index atom, the configuration produced will represent the motif fully. However, this will not always be true since other atoms, not in the structural motif, may be closer in three-dimensional space to a proportion of the 8 atoms than all atoms in the motif. This would result in the distance matrix of the configuration changing and it would no longer represent the motif.

Now consider the case where two atoms, $A_1$ and $A_2$, say, are approximately equidistant from an atom, $A_3$, within a structural motif, but are far apart from one another in three-dimensional space. Now suppose that one configuration has $A_1$ closer than $A_2$ to $A_3$ but another configuration has $A_2$ closer to $A_3$. Since the matrix has a standard ordering of its columns and rows, the two resulting distance matrices could be very different from each other despite the atoms forming the same structural motif. This particular scenario suggests that a metric which does not depend on the ordering of the matrices columns and rows could be a more robust method.

# Chapter 5

# Conclusion

## 5.1 Conclusions

Motivated by the theory that the three-dimensional structure of a protein is completely determined by it's amino acid sequence, this project introduces a novel method for detecting configurational motifs in protein structures. The work illustrates the great potential for mapping sequence to structure using a database approach, constructed by extracting data from the PDB. The PDB contains the three-dimensional coordinates of heavy atoms from thousands of protein structures which have been determined experimentally using a technique known as x-ray crystallography.

The modeling of three-dimensional conformations is a complex issue despite this knowledge of the link between sequence and structure. This complexity is attributed to the large degree of structural freedom of residues, combined with complicated interactions, both local and distant in sequence, between residues. Methods of predicting structure include ad hoc data-based approaches, which have shown limited success over the past decades, as well as ab initio prediction methods, which have so far proved elusive.

Currently homologous proteins play a leading role in protein structure prediction. This approach, however, relies on the availability of a homologous protein, whose structure has been determined experimentally; a feature which is not always possible. Even if a structure has been experimentally determined, it does not guarantee an accurate prediction and thus a more general approach to the problem is required. This approach is shaped by the conjecture that protein folds are constrained in local sequence-structure space and hence this project identifies configurational motifs which are proximal in three-dimensional space but distant in linear sequence.

The database was constructed by extracting 30 different protein structures from the PDB, chosen at random, with the only requirement being that each structure must have a resolution $\leq$ 2Å. Statistical exploratory data analysis methods, including cluster analysis and principal coordinate analysis are applied to the dataset. Statistical shape analysis techniques are also implemented to test for significant differences in the shapes of any configurational motifs identified.

The initial results are masked by the overwhelming influence of the rigid 6-atom structure of the peptide bond on the clustering solution. The recurring pattern appears to be that each cluster contains configurations with the same type of index atom from the peptide bond. However, not all configurations with the same type of index atom are grouped within the same cluster, suggesting that there is some other underlying feature influencing the clustering solution. Peptide bonds can be rotated by an angle $\psi$, about the alpha carbon-carbonyl carbon bond, and an angle $\phi$, about the alpha carbon-nitrogen bond. The values of $\psi$ and $\phi$ have an effect on the configurations. Hence the solution appears to be determined solely by the index atom element and the degree of rotation of the torsion angles.

The methodology was adapted accordingly and the results from hierarchical agglomerative clustering of the new dataset are much more promising. Two different systematic patterns are identified after applying a constraint stating that the residues contained within the configuration must be distant in sequence. Disulphide bonds and salt bridges are accountable for the configurational motifs identified. Two different clusters contain the disulphide bond motif. Results from principal coordinate analysis and statistical shape analysis show that the mean shapes of these two sets of configurations are significantly different from one another, explaining the reason behind the motif forming two distinct clusters.

Salt bridge motifs, on the other hand, are formed within multiple clusters depending on which combination of positively charged and negatively charged amino acid residues are contained in the configuration. Interestingly, statistical shape analysis shows that there is no evidence to suggest the mean shapes of these sets of configurations are not equal.

Although these discoveries are familiar to the field of bioinformatics, the results here show that the methodology functions as intended. Future work is discussed below, reviewing ideas about how adapting the methodology could lead to more interesting discoveries.

## 5.2 Future work

There is great potential for future work in the field of protein structure prediction using a data-based approach such as this one. The methodology used here is elementary and can be adapted significantly with the aid of additional computing resources. If more computational power was available, then a greater number of protein structures could be used in the compilation of the database. This would allow the potential for greater identification of configurational motifs in a larger number of structures.

If three-dimensional coordinates of hydrogen atoms became easily accessible in the future, then these atoms could be included in the database. These additional atoms could play a key role in the detection of motifs.

The metric considered in this project is very basic and is too simplistic for the scenarios considered in Chapter 4. For example, the case of the interloping atom present in the centre of a conserved motif, is an extremely plausible scenario in which the methodology becomes

unsatisfactory. To account for this, a metric which searches for subsets of the original $m \times m$ matrix appears to be an appropriate solution to this particular problem. This would make the configurations less sensitive to the initial choice of $m$ and would ensure that configurational motifs would be identifiable in a greater range of configurations. A metric which disregarded outliers would also have a similar effect on the results.

It is also noted that the standard ordering of the matrices columns and rows has a large impact on the clustering solution. A metric which does not depend on this ordering will certainly be a more appropriate tool in some situations, allowing configurational motifs, where two atoms are roughly equidistant from the index atom, to be grouped together.

Another idea could be to look at the use of hidden Markov models (HMMs) in local sequence-structure data. Schütz & Delorenzi (2008) comment that these models could be used for predicting the occurrence of configurational motifs and could be extremely effective when patterns consist of motifs which vary in length.

# Bibliography


Bookstein, F.L., (1986), Size and shape spaces for landmark data in two dimensions, *Statistical Science* **1**, 181-222.

Branden, C., & Tooze, J., (1999), *Introduction to Protein Structure, 2nd edition*, New York: Garland Pub.

Brenner, S.E., Chothia, C., Hubbard, T.J., & Murzin, A.G., (1996), Understanding protein structure: Using scop for fold interpretation, *Methods in Enzymology* **266**, 635-643.

Brocchieri, L., & Karlin, S., (2005), Protein length in eukaryotic and prokaryotic proteomes, *Nucleic Acids Research* **33**, 3390-3400.

Bystroff, C., Simons, K.T., Han, K.F., & Baker, D., (1996), Local sequence-structure correlations in proteins, *Current Opinion in Biotechnology* **7**, 417-421.

Chatfield, C., & Collins, A.J, (1980), *Introduction to Multivariate Analysis*, London : Chapman and Hall.

Chothia, C., (1993), One thousand families for the molecular biologist, *Nature* **357**, 543-544.

Dryden, I.L., & Mardia, K.V., (1993), Multivariate shape analysis, *Sankhya: The Indian Journal of Statistics* **55**, 460-480.

Dryden, I.L., & Mardia, K.V., (1998), *Statistical Shape Analysis*, Chichester: Wiley.

Everitt, B.S., (2005), *An R and S-plus Companion to Multivariate Analysis*, London: Springer.

Everitt, B.S., & Dunn, G., (2001), *Applied Multivariate Data Analysis, 2nd edition*, London: Hodder Arnold.

Everitt, B.S., Landau, S., & Leese, M., (2001), *Cluster Analysis, 4th edition*, New York: Oxford University Press Inc.

Fasnacht, M., (2001), A method for automatically finding structural motifs in proteins, School of Computer Science: Carnegie Mellon University.

Goodall, C. R., (1991), Procrustes methods in the statistical analysis of shape, *Journal of the Royal Statistical Society Series B* **53**, 285-339.

Gu, J., & Bourne, P.E., (2009), *Structural Bioinformatics, 2nd edition*, Chichester: Wiley-Blackwell.

Jeffers, J.N.R., (1992), *Microcomputers in Environmental Biology*, Carnforth: Parthenon.

Kaufman, L., & Rousseeuw, P.J., (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, New York: Wiley-Interscience.

Lesk, A.M., (2000), *Introduction to Protein Architecture : The Structural Biology of Proteins*, Oxford: Oxford University Press.

Lodish, H.F., Berk, A., & Kaiser, C.A., (2007), *Molecular Cell Biology*, New York: W.H. Freeman.

MacArthur, M., & Thornton, J., (1991), Influence of proline residues on protein conformation, *Journal of Molecular Biology* **218**, 397-412.

Milligan, G.W., & Cooper, M.C., (1985), An examination of procedures for determining the number of clusters in a data set, *Psychometrika* **50**, 159-179.

Moult, J., (1999), Predicting protein three-dimensional structure, *Current Opinion in Biotechnology* **10**, 583-588.

Murzin, A.G., Brenner, S.E., Hubbard, T., & Chothia, C., (1995), SCOP: A structural classification of proteins database for the investigation of sequences and structures, *Journal of Molecular Biology* **247**, 536-540.

Niggemanna, M., & Steipe, B., (1995), Exploring local and non-local interactions for protein stability by structural motif engineering, *Journal of Molecular Biology* **296**, 181-195.

Penrose, R., (1955), A generalized inverse for matrices, *Proceedings of the Cambridge Philosophical Society* **51**, 406-413.

Qian, N., & Sejnowski, T.J., (1988), Predicting the secondary structure of globular proteins using neural network models, *Journal of Molecular Biology* **202**, 865-884.

Ramachandran, G., & Sasisekharan, V., (1968), Conformation of polypeptides and proteins, *Advances in Protein Chemistry* **23**, 283-438.

Schulze-Kremer, S., (1996), *Molecular Bioinformatics : Algorithms and Applications*, Berlin; New York: Walter de Gruyter.

Schütz, F., & Delorenzi, M., (2008), MAMOT: hidden Markov modeling tool, *Bioinformatics* **24**, 1399-1400.

Sevier, C.S., & Kaiser, C.A., (2002), Formation and transfer of disulphide bonds in living cells, *Nature Reviews Molecular and Cellular Biology* **3**, 836-847.

Sternberg, J.E., (1996), *Protein Structure Prediction: A Practical Approach*, New York: Oxford University Press Inc.

Tang, T., Xu, J., & Li, M., (2005), Discovering sequence-structure motifs from protein segments and two applications, *Pacific Symposium on Biocomputing* **10**, 370-381.

Thornton, J.M., (2001), Proteins: A testament to physics, chemistry, and evolution, *Protein Science* **10**, 3-11.

Tramontano, A., (2005), *Protein Structure Prediction : Concepts and Applications*, Weinheim: Wiley-VCH.

Ulmschneider, M.B., & Sansom, M.S.P., (2001), Amino acid distributions in integral membrane protein structures, *Biochimica et Biophysica Acta* **1512**, 1-14.

Wang, Z.X., (1996), How many fold types of protein are there in nature?, *Proteins: Structure, Function and Genetics* **26**, 186-191.

Ward, J.H., (1963), Hierarchical groupings to optimize an objective function, *Journal of the American Statistical Association* **58**, 236-244.

Xu, Y., Xu, D., & Liang, J., (2005), *Computational Methods for Protein Structure Prediction and Modeling*, New York: Springer.

# Appendix A

# Supplementary information

## A.1 Additional configurations

*Table A.1: Configurations in cluster 2.*

| Atom number | Element | Amino acid | Residue number | PDB accession number |
|---|---|---|---|---|
| 147 | N | TYR | 39 | 3tpi |
| 145 | C | GLY | 38 | 3tpi |
| 148 | CA | TYR | 39 | 3tpi |
| 146 | O | GLY | 38 | 3tpi |
| 149 | C | TYR | 39 | 3tpi |
| 151 | CB | TYR | 39 | 3tpi |
| 144 | CA | GLY | 38 | 3tpi |
| 150 | O | TYR | 39 | 3tpi |
| 143 | N | GLY | 38 | 3tpi |
| 152 | CG | TYR | 39 | 3tpi |
| 697 | N | ALA | 111 | 3tpi |
| 693 | C | SER | 110 | 3tpi |
| 698 | CA | ALA | 111 | 3tpi |
| 694 | O | SER | 110 | 3tpi |
| 701 | CB | ALA | 111 | 3tpi |
| 699 | C | ALA | 111 | 3tpi |
| 692 | CA | SER | 110 | 3tpi |
| 700 | O | ALA | 111 | 3tpi |
| 695 | CB | SER | 110 | 3tpi |
| 702 | N | ALA | 112 | 3tpi |
| 959 | N | ASP | 126 | 1aew |

Continued on next page

*Table A.1 – continued from previous page*

| Atom number | Element | Amino acid | Residue number | PDB accession number |
|---|---|---|---|---|
| 956 | C | ALA | 125 | 1aew |
| 960 | CA | ASP | 126 | 1aew |
| 957 | O | ALA | 125 | 1aew |
| 961 | C | ASP | 126 | 1aew |
| 963 | CB | ASP | 126 | 1aew |
| 955 | CA | ALA | 125 | 1aew |
| 933 | O | GLY | 121 | 1aew |
| 954 | N | ALA | 125 | 1aew |
| 962 | O | ASP | 126 | 1aew |
| 385 | N | SER | 50 | 135l |
| 383 | C | GLY | 49 | 135l |
| 386 | CA | SER | 50 | 135l |
| 384 | O | GLY | 49 | 135l |
| 387 | C | SER | 50 | 135l |
| 389 | CB | SER | 50 | 135l |
| 382 | CA | GLY | 49 | 135l |
| 388 | O | SER | 50 | 135l |
| 381 | N | GLY | 49 | 135l |
| 390 | OG | SER | 50 | 135l |
| 684 | N | THR | 89 | 135l |
| 678 | C | ILE | 88 | 135l |
| 685 | CA | THR | 89 | 135l |
| 679 | O | ILE | 88 | 135l |
| 686 | C | THR | 89 | 135l |
| 688 | CB | THR | 89 | 135l |
| 677 | CA | ILE | 88 | 135l |
| 689 | OG1 | THR | 89 | 135l |
| 676 | N | ILE | 88 | 135l |
| 691 | N | ALA | 90 | 135l |
| 969 | N | CYS | 127 | 135l |
| 967 | C | GLY | 126 | 135l |
| 970 | CA | CYS | 127 | 135l |
| 968 | O | GLY | 126 | 135l |
| 971 | C | CYS | 127 | 135l |
| 973 | CB | CYS | 127 | 135l |
| 966 | CA | GLY | 126 | 135l |

Continued on next page

*Table A.1 – continued from previous page*

| Atom number | Element | Amino acid | Residue number | PDB accession number |
|---|---|---|---|---|
| 972 | O | CYS | 127 | 135l |
| 965 | N | GLY | 126 | 135l |
| 974 | SG | CYS | 127 | 135l |
| 551 | N | SER | 75 | 1a7s |
| 546 | C | SER | 74 | 1a7s |
| 552 | CA | SER | 75 | 1a7s |
| 547 | O | SER | 74 | 1a7s |
| 553 | C | SER | 75 | 1a7s |
| 555 | CB | SER | 75 | 1a7s |
| 545 | CA | SER | 74 | 1a7s |
| 554 | O | SER | 75 | 1a7s |
| 544 | N | SER | 74 | 1a7s |
| 556 | OG | SER | 75 | 1a7s |
| 997 | N | SER | 133 | 1a7s |
| 988 | C | ARG | 132 | 1a7s |
| 998 | CA | SER | 133 | 1a7s |
| 989 | O | ARG | 132 | 1a7s |
| 1001 | CB | SER | 133 | 1a7s |
| 999 | C | SER | 133 | 1a7s |
| 987 | CA | ARG | 132 | 1a7s |
| 1000 | O | SER | 133 | 1a7s |
| 990 | CB | ARG | 132 | 1a7s |
| 1003 | N | GLY | 134 | 1a7s |
| 707 | N | LEU | 91 | 131l |
| 703 | C | SER | 90 | 131l |
| 708 | CA | LEU | 91 | 131l |
| 704 | O | SER | 90 | 131l |
| 711 | CB | LEU | 91 | 131l |
| 709 | C | LEU | 91 | 131l |
| 702 | CA | SER | 90 | 131l |
| 710 | O | LEU | 91 | 131l |
| 701 | N | SER | 90 | 131l |
| 712 | CG | LEU | 91 | 131l |

*Table A.2: Configurations in cluster 17.*

| Atom number | Element | Amino acid | Residue number | PDB accession number |
|---|---|---|---|---|
| 1521 | SG | CYS | 232 | 3tpi |
| 1520 | CB | CYS | 232 | 3tpi |
| 811 | SG | CYS | 128 | 3tpi |
| 1517 | CA | CYS | 232 | 3tpi |
| 810 | CB | CYS | 128 | 3tpi |
| 1518 | C | CYS | 232 | 3tpi |
| 1516 | N | CYS | 232 | 3tpi |
| 1522 | N | ASN | 233 | 3tpi |
| 1506 | CD | LYS | 230 | 3tpi |
| 807 | CA | CYS | 128 | 3tpi |
| 1932 | SG | CYS | 38 | 3tpi |
| 1931 | CB | CYS | 38 | 3tpi |
| 1740 | SG | CYS | 14 | 3tpi |
| 928 | CA | CYS | 38 | 3tpi |
| 1739 | CB | CYS | 14 | 3tpi |
| 1929 | C | CYS | 38 | 3tpi |
| 1930 | O | CYS | 38 | 3tpi |
| 1927 | N | CYS | 38 | 3tpi |
| 1735 | N | CYS | 14 | 3tpi |
| 608 | CD2 | LEU | 99 | 3tpi |
| 237 | SG | CYS | 30 | 135l |
| 236 | CB | CYS | 30 | 135l |
| 881 | SG | CYS | 115 | 135l |
| 233 | CA | CYS | 30 | 135l |
| 880 | CB | CYS | 115 | 135l |
| 234 | C | CYS | 30 | 135l |
| 940 | NE1 | TRP | 123 | 135l |
| 235 | O | CYS | 30 | 135l |
| 877 | CA | CYS | 115 | 135l |
| 941 | CE2 | TRP | 123 | 135l |
| 927 | SG | CYS | 123 | 1a7s |
| 926 | CB | CYS | 123 | 1a7s |
| 1355 | SG | CYS | 181 | 1a7s |
| 923 | CA | CYS | 123 | 1a7s |
| 1354 | CB | CYS | 181 | 1a7s |

Continued on next page

*Table A.2 – continued from previous page*

| Atom number | Element | Amino acid | Residue number | PDB accession number |
|---|---|---|---|---|
| 922 | N | CYS | 123 | 1a7s |
| 913 | C | ARG | 122 | 1a7s |
| 914 | O | ARG | 122 | 1a7s |
| 1122 | CG2 | VAL | 148 | 1a7s |
| 1351 | CA | CYS | 181 | 1a7s |
| 1355 | SG | CYS | 181 | 1a7s |
| 1354 | CB | CYS | 181 | 1a7s |
| 927 | SG | CYS | 123 | 1a7s |
| 1351 | CA | CYS | 181 | 1a7s |
| 926 | CB | CYS | 123 | 1a7s |
| 1350 | N | CYS | 181 | 1a7s |
| 1345 | C | VAL | 180 | 1a7s |
| 1346 | O | VAL | 180 | 1a7s |
| 1340 | CG | LEU | 179 | 1a7s |
| 1342 | CD2 | LEU | 179 | 1a7s |
| 240 | SG | CYS | 30 | 1b2k |
| 239 | CB | CYS | 30 | 1b2k |
| 894 | SG | CYS | 115 | 1b2k |
| 236 | CA | CYS | 30 | 1b2k |
| 893 | CB | CYS | 115 | 1b2k |
| 952 | NE1 | TRP | 123 | 1b2k |
| 237 | C | CYS | 30 | 1b2k |
| 238 | O | CYS | 30 | 1b2k |
| 890 | CA | CYS | 115 | 1b2k |
| 269 | CE2 | PHE | 34 | 1b2k |
| 1245 | SG | CYS | 30 | 1b2k |
| 1244 | CB | CYS | 30 | 1b2k |
| 1903 | SG | CYS | 115 | 1b2k |
| 241 | CA | CYS | 30 | 1b2k |
| 1902 | CB | CYS | 115 | 1b2k |
| 1242 | C | CYS | 30 | 1b2k |
| 1961 | NE1 | TRP | 123 | 1b2k |
| 1243 | O | CYS | 30 | 1b2k |
| 1278 | CE2 | PHE | 34 | 1b2k |
| 1899 | CA | CYS | 115 | 1b2k |
| 240 | SG | CYS | 30 | 194l |

Continued on next page

*Table A.2 – continued from previous page*

| Atom number | Element | Amino acid | Residue number | PDB accession number |
|---|---|---|---|---|
| 239 | CB | CYS | 30 | 194l |
| 896 | SG | CYS | 115 | 194l |
| 236 | CA | CYS | 30 | 194l |
| 895 | CB | CYS | 115 | 194l |
| 237 | C | CYS | 30 | 194l |
| 954 | NE1 | TRP | 123 | 194l |
| 238 | O | CYS | 30 | 194l |
| 892 | CA | CYS | 115 | 194l |
| 269 | CE2 | PHE | 34 | 194l |
| 349 | SG | CYS | 45 | 4pep |
| 348 | CB | CYS | 45 | 4pep |
| 380 | SG | CYS | 50 | 4pep |
| 345 | CA | CYS | 45 | 4pep |
| 379 | CB | CYS | 50 | 4pep |
| 346 | C | CYS | 45 | 4pep |
| 788 | O | GLU | 105 | 4pep |
| 350 | N | SER | 46 | 4pep |
| 356 | N | SER | 47 | 4pep |
| 361 | OG | SER | 47 | 4pep |
| 965 | SG | CYS | 388 | 3b8z |
| 64 | CB | CYS | 388 | 3b8z |
| 1591 | SG | CYS | 471 | 3b8z |
| 961 | CA | CYS | 388 | 3b8z |
| 1590 | CB | CYS | 471 | 3b8z |
| 1586 | N | CYS | 471 | 3b8z |
| 960 | N | CYS | 388 | 3b8z |
| 1567 | O | HIS | 468 | 3b8z |
| 1587 | CA | CYS | 471 | 3b8z |
| 1575 | CA | GLY | 469 | 3b8z |

## A.2 R functions

### A.2.1 Function 1

```
protein=function(m,w,t){
mat=matrix(w$xyz,t,byrow = T)
mx=seq(1, 3*t, by = 3)
my=seq(2, 3*t, by = 3)
mz=seq(3, 3*t, by = 3)
co1=w$xyz[mx]
co2=w$xyz[my]
co3=w$xyz[mz]

g=array(0, c(m,m,t))
for(a in 1:t){
d=t(matrix(rep(0,t*3),3,byrow=T))
for(i in 1:t){
for(j in 1:3){
d[i,j]=mat[i,j]-mat[a,j]
}}

e=rep(0,t)
for(k in 1:t){
e[k]=sqrt(d[k,1]^2+d[k,2]^2+d[k,3]^2)
}
y=sort(e)

zz=t(matrix(rep(0,m*3),3,byrow=T))
for(l in 1:t){
for(n in 1:m){
if(y[n]==e[l]){zz[n,1]=(co1[l])}&{zz[n,2]=(co2[l])}&{zz[n,3]=(co3[l])}
}}

g[,,a]=matrix(dist.xyz( zz[1:m,], zz[1:m,]),nrow=m)
}
return(g)
}
```

### A.2.2 Function 2

```
protein2=function(m,w,t){
mat=matrix(w$xyz,t,byrow = T)
mx=seq(1, 3*t, by = 3)
my=seq(2, 3*t, by = 3)
mz=seq(3, 3*t, by = 3)
co1=w$xyz[mx]
co2=w$xyz[my]
co3=w$xyz[mz]

F=t(matrix(rep(0,t*m),m,byrow=T))
for(a in 1:t){
d=t(matrix(rep(0,t*3),3,byrow=T))
for(i in 1:t){
for(j in 1:3){
d[i,j]=mat[i,j]-mat[a,j]
}}

e=rep(0,t)
for(k in 1:t){
e[k]=sqrt(d[k,1]^2+d[k,2]^2+d[k,3]^2)
}
y=sort(e)

for(l in 1:t){
for(n in 1:m){
if(y[n]==e[l]){F[a,n]=l}
}}
}
return(F)
}
```