A Method for Automatically Finding Structural Motifs in Proteins

Marc Fasnacht

August 2002

CMU-CALD-02-105

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Abstract

A lot of work has gone into predicting the secondary (small scale) structure of proteins from their amino acid sequence. Current research indicates that there are limits on how well secondary structure can be predicted from local sequence information. To further advance prediction, the interactions between elements of secondary structure which are inherently non-local, have to be better understood.

This project studies a special case of secondary structure interaction, coupled helical motifs, consisting of two interacting helices. The underlying hypothesis of this work is that there are different types of coupled helical motifs, which can be characterized by different sets of rules governing the underlying amino acid sequence of the protein. In order to learn such rules, a classification of the coupled helical motifs needs to be introduced. This can be achieved by unsupervised learning methods such as clustering.

We present a method to automatically extract structural motifs in proteins. The method uses hierarchical agglomerative clustering to find structurally equivalent sets of motifs in proteins. These motifs can be used for study of the underlying amino acid sequence. We test the method on a set of coupled helical motifs from the globin family of proteins. It rediscovers important aspects of the well known structural hierarchy of this protein family.

Keywords: data mining, agglomerative clustering, cluster validation, cluster visualization, protein structure, protein structure classification

Contents

1 Introduction

Understanding and predicting the structure of proteins is one of the most important problems in modern biology. Chemically, proteins are polypeptides - linear chains of amino acids - which fold into complex three-dimensional structures. (see fig. 1 for examples). Each protein has a unique sequence of amino acids. The so-called "Second Central Dogma" of molecular biology states that the amino acid sequence of a protein completely determines its three-dimensional structure and that it is this structure that determines the function of the protein [3]. It is relatively easy to measure the amino acid sequence of a protein experimentally, so that most protein sequences are known. On the other hand it is extremely costly and time intensive to determine the three dimensional structure [7]. Only about 18000 structures, a small fraction of the total number of proteins, has been solved.

Proteins molecules play important roles in most biological processes. Given this, it is essential for any kind of biological or medical applications to know and understand the three-dimensional structure. Most drugs work by interfering with the function of specific proteins in the body. The discovery of drugs is much easier if the structure of the protein is known. Molecules can then be specifically designed to bind to certain parts of the protein to block or enhance its functions (e.g. blocking the docking of AIDS viruses to cells). The current method for drug discovery is to systematically screen hundred thousands of chemical compounds experimentally with a trial and error method, which is very costly. Computational methods for structure prediction from sequence are therefore extremely valuable in pharmaceutical research.

Given that there is a sufficiently large number of known protein structures, one of the main approaches to protein structure prediction is based on machine learning and data mining techniques [1]. Since directly predicting the structure is much too difficult (some studies suggest NP-hard [8]), it is important to solve intermediate steps such as finding repeating structural motifs that can be predicted from the sequence. If a sufficient number of these motifs can be identified for a given sequence, they could be used as building blocks of sorts to determine the structure of the protein.

An equally important task is the inverse prediction problem. Here the task is to find sequences that fold into a given three dimensional structure. This approach might be used to find sequences in a database that fold into a particular three-dimensional structure. It is also important in designing proteins to perform a specific function, which requires them to fold in a certain way.

The first step in solving the inverse folding problem is to understand how elements of secondary structure interact and form specific motifs. In order to study this problem, we need

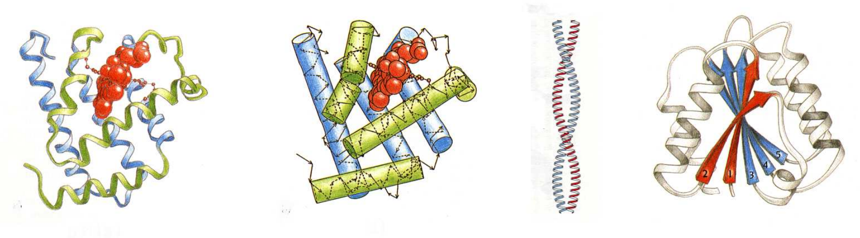

Figure 1: Different schematic representations of protein structures. The protein is represented by the helical thin bands on the first picture from the left. We can see several coupled helical motifs (helices crossing at different angles). The middle picture shows a coiled coil (much smaller scale). The protein on the right has helices as well as sheets (colored)

to have specific examples of secondary structure. Given enough examples of a certain type of secondary structure motifs, we can use statistical and machine learning tools to analyze the characteristics of these amino acid sequences and derive rules about what amino acid patterns define these motifs.

There are several known examples of such motifs, however currently there is no systematic classification of the motifs. The set of known protein structures is too large to do this by hand and automated techniques will need to be applied. The method we present in this paper addresses this problem: it automatically extracts instances of a specific type of secondary structure motifs from a protein structure database, and groups them by similarity using clustering methods. The members of the resulting groups belong to the same secondary structure motif, and can be used for further study. In this project we examine a particular subset of structural motifs: *coupled helical motifs*. We define coupled helical motifs as motifs consisting of two helices in contact with each other (e.g. parallel or crossing at a certain angle). Examples can be seen in figure 1. In order to avoid confusion, we will use the term *helix couple* to refer to two helices in contact from now on.

The remainder of this paper is organized as follows: the section2 gives background information on protein structure. It is followed by a detailed description of the individual steps of our method. The fourth section presents and discusses our results from applying the method to a dataset with proteins from the globin family. We conclude with a summary and a discussion of future work.

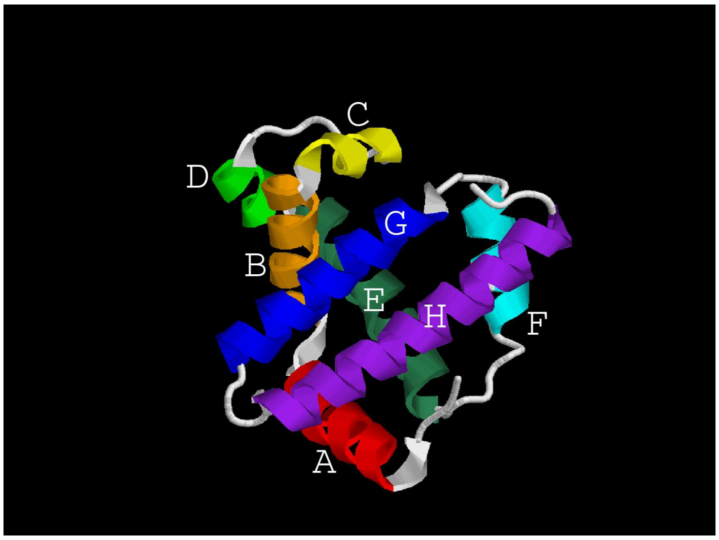

Figure 2: Systematic representation of a myoglobin, a protein responsible for oxygen storage in muscle tissue. The colored segments, labeled A-H, represent helical sections of the protein backbone

2 Background

In biology the spacial structure of proteins is classified at different levels: Primary structure refers to the amino acid sequence (i.e. directly adjacent to each other). Secondary structure refers to the relative spacial arrangement of amino acids that are near each other in the sequence. Typical examples of secondary structure are helices or sheets (see figures 1 and 2). Tertiary structure describes the spatial arrangements of amino acids that are far apart in the sequence.

Proteins can be classified into different families. One such classification, which we have used extensively in this project, is the CATH protein structure classification at

http://www.biochem.ucl.ac.uk/bsm/cath_new/.

It groups proteins hierarchically at four major levels. The top level, the Class-level (C), describes secondary structure composition and packing. There are three major C-classes: mainly alpha helix, mainly beta-sheet and mixed, alpha-beta. Figure 2 shows a protein made up mainly of alpha helices. The next level, Architecture (A), depends on the overall shape of the structure (i.e. the relative orientations of the secondary structure, ignoring their connectivity). The Topology (T), or fold family level, describes both, the overall shape and connectivity of the secondary structure. The last major level, the Homologous Superfamily or H-level, groups structures that are thought to have a common evolutionary ancestor into families. Typically, structures with the same H-level classification have certain degree of sequence similarity. There are further sub-levels in the CATH classification. We have used the S-level, which demands sequence similarity of at least 35%, and, the N-level requires yet an even higher sequence identity. Domains in the same H-level class virtually have the same sequence.

3 Approach

3.1 Overview

The goal of the method introduced here is to automatically find structural motifs in proteins. The approach we take can be summarized as follows:

- Define a general type of motif we are interested in. For this project we focussed on the family of coupled helical motifs, which we define as motifs consisting of two helices in contact with each other, or helix couples
- Scan the protein deposited in the Protein Database for motifs that fall into the selected category (helix couples in our case).
- Extract the coordinates of the atoms making up the helix couple and calculate the pairwise rms-distances, as defined in equation 1 below, between all helix couples found.
- Use the rms-distances to cluster the data using hierarchical agglomerative clustering. This results in clusters of helix couples that are similar to each other.
- Use multidimensional scaling to visualize and interpret the data.
- Identify interesting clusters by looking for jumps in the mean internal distance (MID).[1]

The details of the different steps in the method are described in more detail in the remainder of this section.

3.2 Selection of data

As mentioned above, we have focussed our work on coupled helical motifs. The first step in identifying sub-structures that fit this description is to find all the helices in a protein structure file. We do this with a structure analyis program such as the *define_structure* program by Richards and Kundrot (1988). This program will examine the atomic coordinates of the protein and on the basis of bonding angles of the backbone return a list with all the starting and end points of helices in the file. This step is fairly fast, since it is linear in the number

[1] The mean internal distance of a cluster is defined as the weighted mean of the average pairwise distance between points in a cluster, over all the clusters.

of protein structures we examine. It took less than an hour on an Intel Pentium based Linux machine to do this for our data-set.

The next step was to find which helices are in contact and make up a helix couple. By being in contact, we mean that the helices have some amount of buried surface area.[2] We calculated the buried surface area for each possible pair of helices for a given protein using the *naccess* program by S. Hubbard and J. Thornton and kept the pairs with a buried surface greater than a certain threshold. We also imposed a fixed length for the helices in order to have a constanct number of atoms for the distance calculation. Finding helix couples is quadratic in the number of helices per protein structure and linear in the number of protein structures examined. However, since there are usually only few helices per protein[3], this step took just a bit over an hour on an Intel Pentium based Linux machine for our data-set.

3.3 Similarity Measure: rms-distance

In order analyze the data-space, we need to have a distance metric to compare pairs of helix couples. Given two helix couples, different similarity measures between the pairs of couples are possible. The most general and most widely used measure in structural biology is the rms-distance. It is determined as follows: the two pairs of helix couples are ridigly rotated and translated on top of each other such that the root-mean-square (rms) distance between corresponding atoms on the helices is minimized. (This corresponds to finding the orientation of maximal overlap.) Mathematically, for a pair of structures, i and j, the rms-distance is given by

$$d_{rms,ij} = \min_{rot,trans} \sqrt{\frac{\sum_{k=1}^{N} \|\mathbf{r_{i,k}} - \mathbf{r_{j,k}}\|^2}{N}} \tag{1}$$

where $\mathbf{r_{i,k}}$ and $\mathbf{r_{i,k}}$ are the position vectors of the k-th atom in the respective structure. The sum runs over pairs of equivalent atoms. The minimal rms-distance is unique and it is used as the distance function. We have calculated the rms-distance between all helix couples using the *ProFit* program by Dr. Andrew C. R. Martin which can be obtained at *http://www.biochem.ucl.ac.uk/ martin/*. This step is by far the computationally most intensive of our method. It took on the order of two days, using all nodes on an eight node, 1.2 GHz AMD Athlon based Beowulf cluster, to do this calculation for our data-set.

[2]The buried surface area of two objects is defined as the sum of the surace areas of the individual objects minus the surface area of the two objects put in contact.

[3]globin, which is a alpha-helical protein, has eight helices.

3.4 Multi-Dimensional Scaling (MDS)

The calculation of all the pairwise distances in the previous step results in a very large[4] distance matrix. It is difficult to analyze and visualize such a table. One of the main purposes of multidimensional scaling (MDS) is to provide a coordinate representation of the similarity or distance relations among a set of objects. Often it also results in a dimensionality reduction of the problem. This allows easier visualization of the data and also allows the application of methods that rely on a coordinate representation (e.g., k-means clustering). The description of MDS given below closely follows the one given by Cox [5].

Suppose we are given a set of N objects for which we do not have a coordinate representation but are instead given a matrix of pairwise distances δ_{ij}. MDS allows us to find a set of N points $\{\mathbf{x_i}\}$, $i = 1, 2, \ldots, N$ in a p-dimensional Euclidian space such that the distances between the points in that space d_{ij} obey

$$d_{ij} \approx f(\delta_{ij}). \tag{2}$$

Here f is a monotonic function of the distances. There are several methods of solving this problem. Given a set of coordinates $\{\mathbf{x_i}\}$ for the points, we can define a stress function which measures how well the spatial configuration of the points satisfies equation 2. A commonly used function is

$$S = \frac{\sum_{i \neq j} (d_{ij} - f(\delta_{ij}))^2}{\sum_{i \neq j} d_{ij}^2} \tag{3}$$

Finding a coordinate representation $\{\mathbf{x_i}\}$ of the data with the desired distances then corresponds to minimizing this stress function. Minimization can be done in standard fashion by using gradient descent or annealing techniques.

Another approach is the following: Let us define the matrix $\mathbf{A}$ as $A_{ij} = -1/2\delta_{ij}^2$, with δ_{ij} as defined above. We define a second matrix $\mathbf{B}$ as $\mathbf{B} = \mathbf{XX^T}$ where $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}]^\mathbf{T}$. These are the coordinates of the points we want to determine. It can be shown (c.f.[5]) that

$$\boldsymbol{B} = \boldsymbol{HAH} \tag{4}$$

where $\mathbf{H}$ is given by $\mathbf{H} = \mathbf{I} - \mathbf{N^{-1} 1 1^T}$ with the length-N vector $\mathbf{1} = (\mathbf{1}, \mathbf{1}, \ldots, \mathbf{1})^\mathbf{T}$. $\mathbf{B}$ is symmetric and positive semi-definite of rank p. It can be decomposed into

$$\boldsymbol{B} = \boldsymbol{V \Lambda V^T} \tag{5}$$

where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues λ_i and $\mathbf{V} = [\mathbf{v_1}, \ldots, \mathbf{v_N}]$ is the matrix of eigenvectors of $\mathbf{B}$. The problem of finding the coordinates x_i therefore reduces to solving

[4] 3459 by 3459 in the case of our data set.

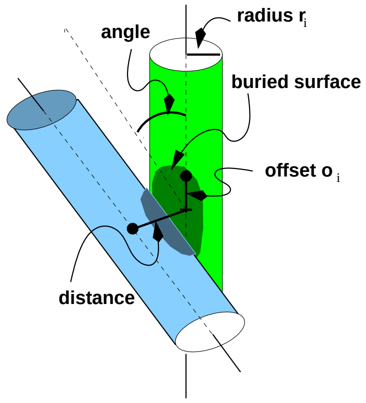

Figure 3: Systematic representation of a helix couple as two cylinders.

the decomposition problem in equation 5, since $\mathbf{X} = \mathbf{V}\Lambda^{\frac{1}{2}}$. This is a standard problem in linear algebra. Generally, some of the eigenvalues are small so that they can be neglected. This simplifies the numerical solution of the problem. The advantage of this approach over direct minimization of the stress function 3 is that the eigenvectors found are equivalent to those found by a principal component analysis in the projection space. This often simplifies interpretation of the data.

3.5 Interpretation of dimensions

Once we have a coordinate representation of the data, we can visualize it more easily. However, we do not usually know what the different MDS coordinates mean. In the case of the helix couples, we have auxiliary features, properties such as crossing angle, distance, etc (see figure 3) which describe the data. We would like to relate those features to the MDS coordinates.

Sometimes it is possible to find an interpretation in terms of auxiliary features by visual inspection [2], but this is not very practical, even for a modest number of dimensions. In this project we have used an approach based the nearest neighbor method to interpret the MDS coordinates. The method allows us to find mappings between reduced dimension representations (MDS coordinates) and auxiliary features that describe the data. It finds groups of dimensions that taken together preserve local structure in the auxiliary feature space. We have described this method in detail in an earlier report [6]. Here we only give a brief qualitative summary.

The problem the method solves is to relate a set of MDS coordinates $\{c_1, \ldots, c_N\}$ to some auxiliary feature f. The intuition behind the approach is the following: Suppose two points are near each other in a subspace defined by MDS coordinates $\{c_i, \ldots, c_j\}$. If this subspace is well characterized by feature some f, the points should have similar values of f.

This property can be measured by looking for the nearest neighbor in the MDS subspace and calculating an average distance with respect to feature f over all the data. We then compare this value to the average distance between all the points in the feature space f.

If the space $\{c_i, \ldots, c_j\}$ is unrelated to f, then picking the closest point in that space would be equivalent to picking a random point in f. The average f-distance of a nearest neighbor point $\{c_i, \ldots, c_j\}$ should be similar to the average distance in the f. So the ratio of the average f-distance over nearest neighbors to the average f-distance over the entire space should be close to 1. If there is a strong relationship, picking a close point in $\{c_i, \ldots, c_j\}$ should correspond to picking a point that is close in terms of f too. Therefore the average nearest neighbor distance in terms of f should be much smaller than the average f-distance over all the points, and the ratio of the two numbers should be small.

By looking at these ratios for all possible combinations of features and MDS coordinates, we can identify the features that have a correlation with a given set of coordinates.

3.6 Hierarchical Agglomerative Clustering

The next step is to find sets of similar helix couples. We use Hierarchical Agglomerative Clustering (HAC) for this purpose. Given a distance function defined between two points, HAC starts putting all the points in the dataset in singleton clusters (i.e. number of clusters equals number of data points). It then proceeds to merge the clusters that have the smallest distance between each point in one and each point in the other. This corresponds to merging clusters in such a way that the mean internal distance of the clusters is increased by the smallest amount. This is repeated until we have one cluster containing all the points. HAC can be described by a dendrogram or clustering tree which can serve to visualize the clustering. (The description given here follows [4]). Figure 4 shows an example.

The main advantage of agglomerative clustering is that it gives a hierarchical clustering tree, which will be useful for the rule learning stage of the problem. Other clustering methods are worth exploring, but this will be left to a later stage of the project.

3.6.1 Importance of mean internal distance

During HAC, the mean internal distance of the clustering monotonically increases as more and more clusters are merged. If two merged clusters are very similar, joining them will not increase the mean internal distance by much. If two very different clusters are merged, as often happens towards the end of the clustering algorithm, there will be a large jump in mean internal distance. We can systematically look for such jumps to find interesting merges.

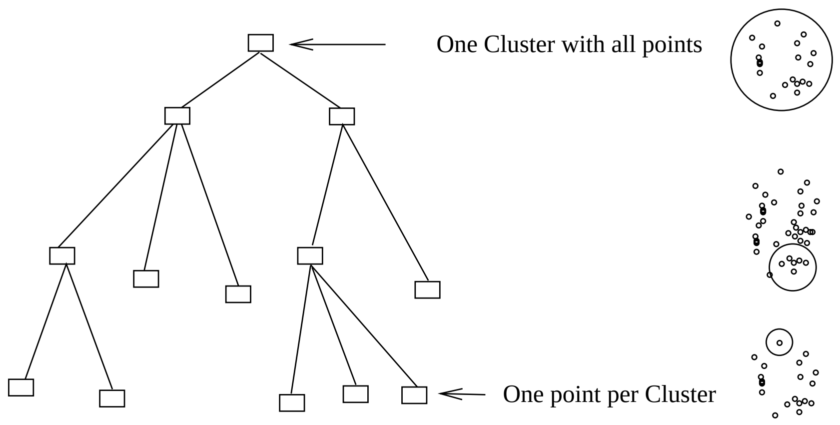

Figure 4: Schematic picture of an agglomerative clustering tree. The leaf nodes correspond to "clusters" consisting of only one data point. The root node represents one single cluster containing all the points.

As long as the increases in mean internal distance are small, we are merging clusters that are very similar, thus the points in the resulting clusters are very similar. During a jump, dissimilar clusters are merged. The resulting cluster is not very homogeneous anymore. It is then interesting to see if the jump in the mean internal distance can be explained by looking at the points that are merged and in what ways they differ. This will give an idea what features are related to the jump.

4 Experiments

We tested our method on a set of protein structures from the globin family. We chose this set of structures, because the globin family is extremely well studied. If our method works, we should be able to rediscover some of the well known properties of helix couples in proteins.

4.1 Globin data set

4.1.1 Extraction of Data

Our data is derived from crystallographic structures publicly available in the Protein Data Bank (PDB) (*http://www.rcsb.org/pdb/*). We used the CATH classification of proteins to find the proteins that belong to the globin family. The access code for globins in CATH is 1.10.490.10. This CATH code can be used to generate a list of access labels for the PDB for the globin domains. With these access labels we then extracted 609 matching structures (i.e.

the atomic coordinates) from the PDB. Following the procedure outlined in section 3.2 we determined the location of all pairs of helices that fit our criteria for helix couples in these proteins. Choosing a length of nine residues as our helix length, this procedure found a set of 3459 instances of helix couples. All the experiments described below were done on this data set.

4.1.2 Labels

There is an extensive literature on the globin which we can use to describe our data set. This was the main reason we chose globins: so much is known about them that they provide an ideal test ground.

The globins have essentially the same overall fold and by extension essentially the same set of helices. They are given specific labels in the globin literature: one of the letters A-H. Figure 2 shows an instance of a globin with labeled helices. Of course for any given structure, we do not necessarily have these labels, but we can easily find them by doing a multiple sequence alignment. Given the alignment, we compare each globin in the data set to a reference structure with known labels to determine the label of each helix. Each helix couple is assigned a pair of letters, corresponding to the labels of the constituent helices. We will refer to this letter pair as 'helix couple label'.

We can also group the helix couples according to the globin from which it came. Using the CATH S- and N-levels, we can label the globins according to sequence similarities. If two sets of equivalent helix couples (same helix couple labels) come from proteins with the same CATH S-level, we know that they must have relatively high degree of sequence similarity. The same is true to an even stronger degree for helix couples with the same CATH N-level.

4.2 Results

4.2.1 Multidimensional Scaling

The results of the multidimensional scaling analysis has been described in a previous paper [6]. They can be summarized as follows: We used MDS to project the pairwise distance matrix into a 10-dimensional euclidian space. Examining the space and a set of auxiliary features using the KNN method showed that the main MDS coordinates[5] are closely connected to the crossing angle and the distance between the two helices in a couple. This is in line with what we expect, since crossing angle and distance between helices are the main features that structural biologists use to characterize a helix couple.

[5] A subset of coordinates that is sufficient to reproduce the distance matrix up to a small error.

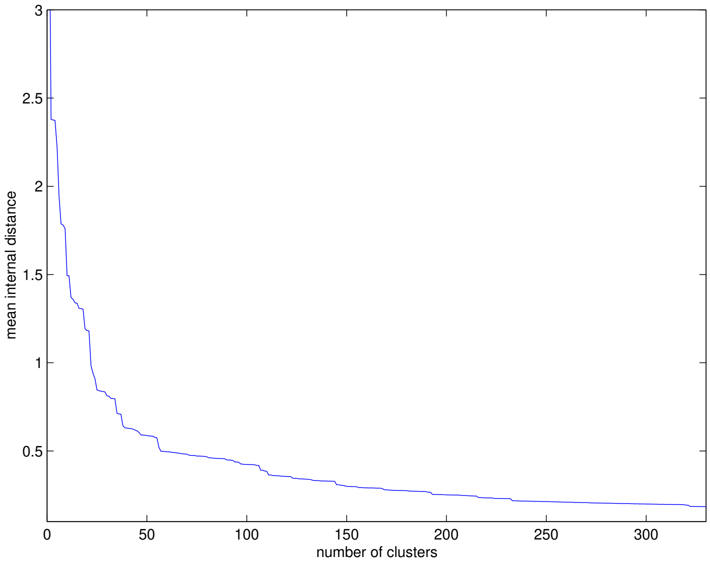

Figure 5: Mean internal distance as function of number of clusters

4.2.2 Mean Internal Distance

Figure 5 shows a plot of the mean internal distance of the clustering as a function of the number of clusters. As described in section 3.6.1, jumps in the mean internal distance indicate that two dissimilar clusters are being merged, which in turn hints at something interesting happening. It is easier to detect jumps in a plot of the difference in mean internal distance, which is shown by the top curve in figure 6. The bottom curve in the same figure shows jumps in a measure of the sequence difference, based on the BLOSUM62 similarity matrix. We can clearly see that jumps in the mean internal distance coincide with jumps in the sequence distance between the underlying helices. This indicates that the clustering in structure space captures important information about the underlying amino acid sequence of the proteins.

4.2.3 Accuracy

To further validate our method we looked at a the following measure of accuracy of the clustering: each cluster is assigned by the label of the majority of the points in the cluster. Points that have the same label as their cluster are counted as correct, whereas points that have a different label are counted as wrong. Obviously, this measure is somewhat biased, since clusters of size one will automatically have the correct label. We can deal with this in two ways: we can either discount small clusters, or classify them as wrong, since small clusters do not give us very much information.

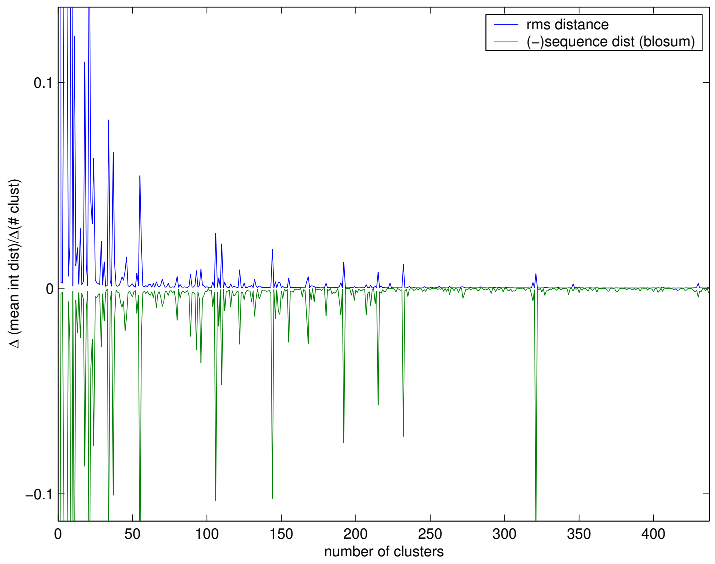

Figure 6: Change in mean internal rms and sequence distance as a function of cluster size

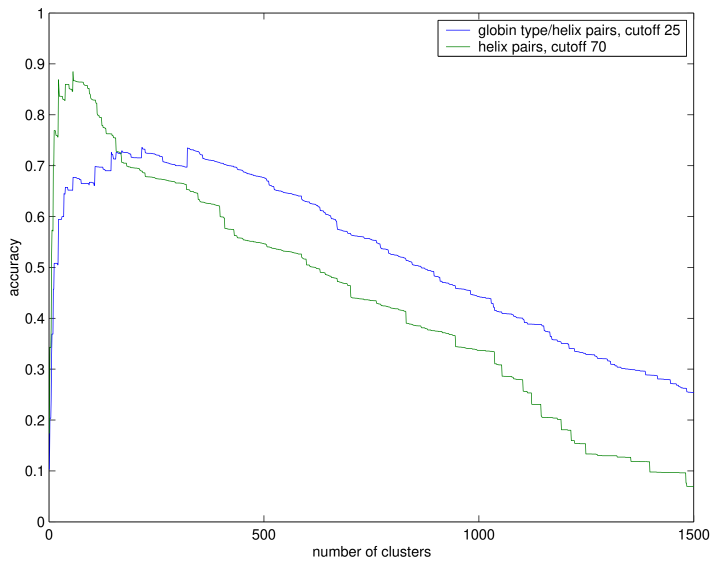

Figure 7: Accuracy as a function of number of clusters. The green curve shows the accuracy for using the helix couple label, and counting clusters with less than 70 points as wrong. The blue curve shows the accuracy for labels describing globin type and helix couple label with a cutoff of 25 points per cluster

We can then average the number of 'correct' labels over all the clusters. This measure of accuracy will vary with the number of clusters we choose for the clustering, and also with what labels we use. Figure 7 shows the accuracy as a function of the number of clusters for two sets of labels. Small clusters were counted as wrong in this case. We can see that the accuracy is fairly high as long as we chose a reasonable number of clusters. We are interested in large clusters, the values the are towards the left of the curve. This is exactly the region where the curves peak. As expected, the two curves peak at different values. If we just use the helix couple labels, we have only 12 different classes, and a small number of clusters suffice to describe the data. We can reach accuracies of over 85%, if we choose between 25 and 100 clusters (random guessing would be 17% accuracy). If we use labels that are a combination of helix couple and globin type, we have 97 different labels. In this case, once we have fewer clusters, we necessarily will start to see the accuracy decrease, since we do not have enough clusters for all the classes. Here the accuracy is greater than 70% if we choose between 150 and 250 clusters (random guess: 10%). This clearly indicates that in both cases, the clustering results in relatively pure clusters in terms of the known classes of helix couple labels and globin types.

4.2.4 Clustering trees

We systematically looked for all merges that caused increases of the mean internal distance of the clustering of more than 1%, and stored all the clusters (before the merge). We used these clusters to build a simplified clustering tree[6]. In order to better understand these clustering trees, we can show the makeup of the clusters before the merge in terms of the various labels.

Figure 8 shows the this reduced clustering tree. Each cluster is indicated by a vertical bar, the edges connecting the bars indicate which clusters merge. The vertical position of the bars indicate the mean internal distance of a cluster, just before merging [7].

In figure 8, the coloring of the bars shows the makeup of the clusters in terms of helix couple labels. The helix couple labels cleanly split up among the different sub-branches of the tree. We can see that below a mean internal distance of about 0.6Å, the clusters are essentially pure: i.e. all the points belonging to a cluster with less than 0.6ÅMID are made up the same helix couple type.

Another way of examining the makeup of the clusters is by looking at the globin types.

[6] We could construct the full clustering tree showing all the merges. But most of the merges are merges of very small clusters, which make the tree very large without providing much additional information.

[7] Note that the parent cluster displayed does not necessarily have the exact same composition as right after the merge that created it. Smaller merges, which did not create a large jump in the mean internal distance might have occurred until the parent cluster is merged with another cluster. It is the composition right before that merge which is displayed

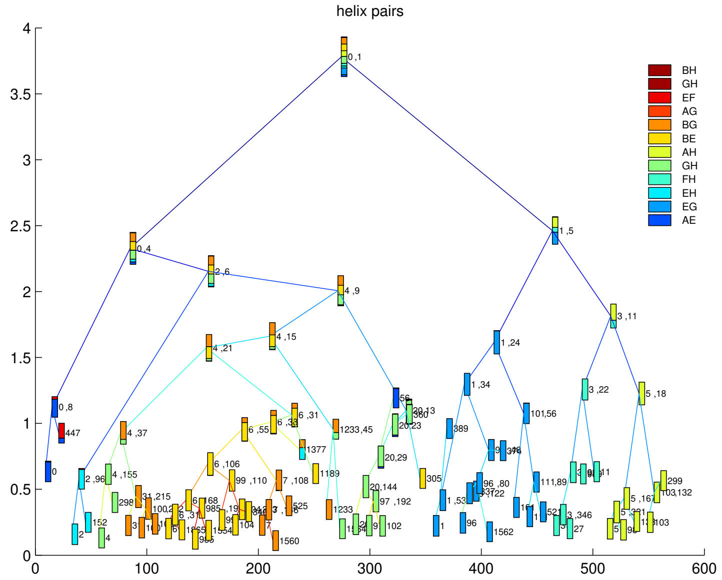

Figure 8: Clustering tree for merges with large change in mean internal distance. The vertical placement of the nodes indicates the mean internal distance of the cluster. The bars indicate the makeup of the clusters in terms of helix couple labels. The numbers next to the nodes give the cluster id and the total number of clusters left after the merge

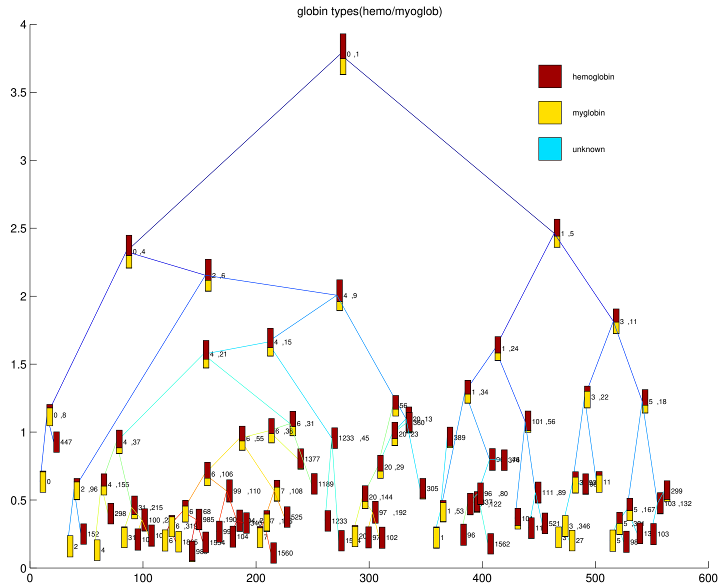

Figure 9: Reduced clustering tree. The bars show the makeup of the clusters in terms of hemoglobins and myglobins

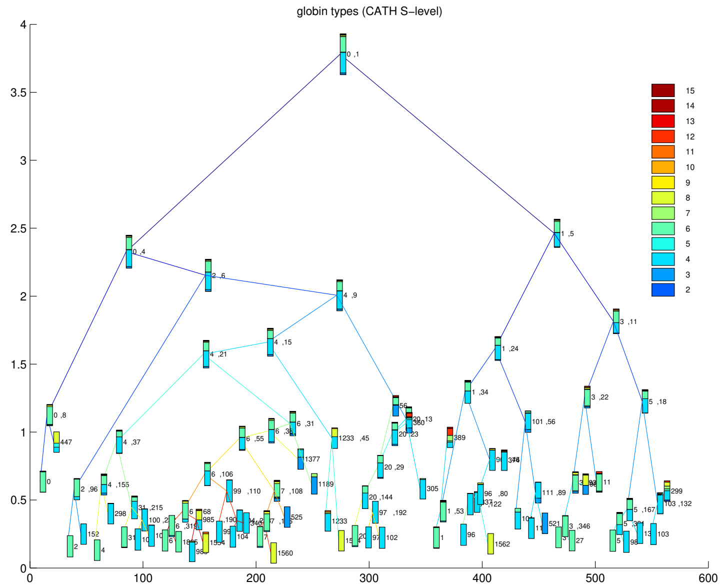

Figure 10: Reduced clustering tree. The bars indicate the makeup of each of the clusters with respect to globin type.

We can do this at different levels. At the coarsest level, we can simply distinguish between hemoglobins and myoglobins. This is shown in figure 9. On this tree, we can clearly see sub-branches of the tree that are predominantly made up of hemoglobins or myoglobins respectively. For example if we look at the child-nodes of the node at coordinates $(150, 0.3)$, labeled with $(6, 106)$ we see that the left subtree consists mainly of helix couples from myglobins whereas the right subtree mostly has helix couples from hemoglobin. By comparing with figure 8, we can see that these helices correspond to the 'BE' helix couple.

Figure 10 again depicts the reduced clustering tree with color showing the globin types as given by the S-level of the CATH taxonomy. The S-level of CATH groups proteins according to sequence homology. If we compare it with the simpler myoglobin-hemoglobins separation on figure 9, we can seen that S-level value 4 corresponds more or less to hemoglobin and S-level value 6 mostly to myoglobins. In contrast to the tree with the helix couple labels, the tree nodes here only become pure for mean internal distances smaller than 0.4Å. This is reassuring, since we expect the difference between different helix couples to be bigger than between equivalent helix couples of different globins.

We can look at an even finer subdivision of the globins in the CATH classification, the N-level. Globin-domains that belong to the same N-level in CATH almost have identical sequences. A reduced clustering tree colored according to the CATH N-level, is displayed in figures 11. For this sub-division, the nodes become only (mostly) pure for mean internal

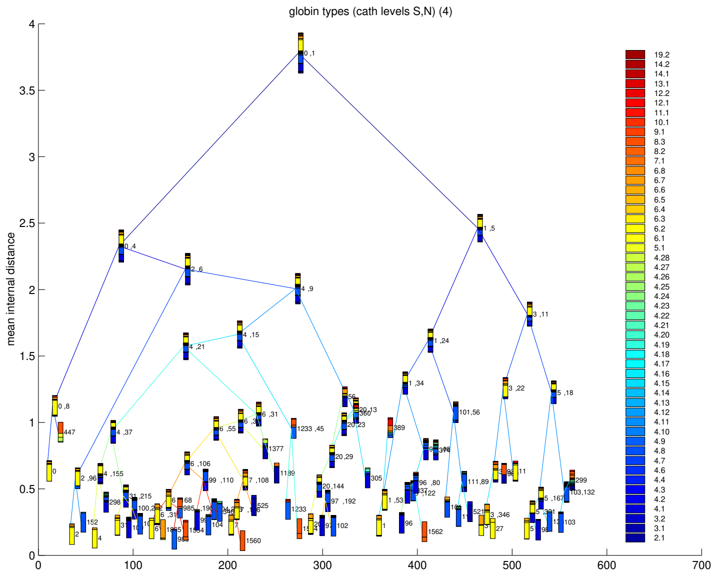

Figure 11: Mean Internal Distance vs cluster size

distance of less than 2 Å.

4.3 Discussion

From our experimental results we can conclude that our method works well. The multidimensional scaling calculation and the analysis with the KNN method show that the rms-distance accurately captures the main descriptive features of our selected motif type, crossing angle and distance. It is important to notice that the rms-distance is a completely general distance metric which can be applied to any kind of motif type. We did not have to specify any particular combination of features to do the clustering. This indicates that the method will work well for other types of motifs, where we do not a priori know which combination of features is most appropriate compare the structures.

As discussed in the introduction, the purpose of extracting these motifs is to find structurally similar sets of motifs, which can the be used for a study of the characteristics of the underlying amino acid sequences. To apply machine learning techniques or statistical methods, we will have to find sufficiently large sets of examples of a given motif that are as pure as possible. This means that we should evaluate the method based the results for the larger clusters. The accuracy data indicates that the hierarchical clustering achieves a quite high degree of purity for larger clusters. We conclude that the method will allow us to find the kind of large, pure clusters that we need for the study of underlying amino acid sequences.

The clustering trees show that the method manages to reproduce important characteristics of the protein hierarchy automatically.

In particular, by looking at specific helix couples, such as the 'BE' or 'EG' helix couples, we can see clear differences between couples coming from hemoglobins or myglobins. This ability the differentiate between myoglobin and hemoglobin is a baseline test for the method, which it clearly passes. It is difficult to relate the differences we see in the clustering tree to specific structural differences between these two types of globins though. The main difference between hemoglobin and myoglobin is that that hemoglobin is a tetramer, made up of two α and β chains, whereas myoglobin is a monomer. We would expect that there might be some differences in the regions where the chains are in contact in the hemoglobin. However, almost all the helix couples in our dataset contain at least one of these helices ('B','G','H' and 'C'), so it is not clear which of the helix couples should display the largest difference.

Another important point to notice is that depending on what level of mean internal focus, we find different levels of structural similarity. Each level of structural similarity can be related to a level of sequence similarity: If we look a clusters with mean internal distance of 0.6Åor less, all the helix couples in a cluster will be identical in terms of the helix couple labels. This means they perform the same function in the protein, and are located in the same range in the sequence. However, they might come from different sub-families of the globin family. This is what we would expect: Different helix couples within the same globin are not expected to have any structural similarity, but for equivalent helix couples from different globins there is such a similarity. If we choose a cluster with mean internal distance of 0.4Åor less, the helix couples in the cluster not only have the same helix couple labels, but they will also belong to the same S-level in CATH, indicating that they have a quite high degree of sequence identity. Finally, all the point in clusters with mean internal distance of 0.2Åor less, will have even more similar sequences, since they will have the same or a very similar N-level classification in CATH. Again, the further down we travel, the more and more similar the helix couples get, not only in terms of structure, as implied in the mean internal distance, but also in terms of underlying amino acid sequence. This clearly shows that our method, using only structural information, allows to find sets of helix couples that have a high degree of similarity in terms of sequence. This is exactly the behavior we are looking for to automatically find sets of data belonging to the same structural motif.

5 Conclusion and Future Work

We have presented a method that allows to automatically extract structural motifs in proteins. The method is based on hierarchical agglomerative clustering and allows to find sets of structurally equivalent points defining such motifs. These motifs can be used for study of the underlying amino acid sequence. The method is tested on set of coupled helical motifs from the globin family of proteins. It rediscovers important features of the well known structural hierarchy of this protein family.

The next steps of the project will involve applying this method to set of general proteins, from all the different levels of CATH. If the method performs as expected, we will be able to find regions of proteins in different types of proteins that all fit the same structural motif. Once we have this data, we will use machine learning methods to look for the rules that characterize the motifs found.

6 Acknowledgments

I would like to thank my advisors, Rich Caruana and John Rosenberg who have guided me through the various stages of this project. Working with them was a truly great experience from which I benefitted a lot! I would also like to thank John Rosenberg for providing financial support for the last two years. Furthermore, I would like to thank Paul Hodor and Bruce Buchanan who worked with us during the early stages of this project. Paul Hodor wrote and helped me use many of the programs I applied during this project. Finally, I would like to thank Hugh Nicholas and John Hempel who helped me greatly with the multiple sequence alignment calculations.

References

[1] P. F. Baldi. *Bioinformatics, the Machine Learning Approach.* MIT Press, Cambridge, 1999.

[2] I. BORG and J. LINGOES. *Multidimensional Similarity Structure Analysis.* Springer-Verlag, New York, 1987.

[3] C. Branden and J. Tooze. *Introduction to Protein Structure, Second ed.* Garland Publishing, New York, 1999.

[4] R. Caruana, P. Artigas, A. Goldenberg, and A. Likhodedov. Meta clustering. *submitted for publication, ICML2001*, 2001.

[5] T. F. Cox and M. A. Cox. *Multidimensional Scaling.* Chapman & Hall, London, 1994.

[6] M. Fasnacht and R. Caruana. A method for automatically finding interpretations of reduced dimension representations. *CMU Technical Report CMU-CALD02-104*, 2002.

[7] L. Streyer. *Biochemistry.* Freeman, 1995.

[8] R. Unger and J. Moult. Finding the lowest free energy conformation of a protein is an np-hard problem: Proof and implications. *Bull. Math. Biol.*, 55:1183, 1993.