Classifying Structural Motifs in Proteins for Rule Learning

Marc Fasnacht MF5K@ANDREW.CMU.EDU
Physics Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract

Protein structure prediction is one of the most important problems in modern biology with direct applications in drug discovery and medicine. A lot of work has gone into predicting the secondary (small scale) structure of proteins from their amino acid sequence. However the spatial relationship between elements of secondary structure is not very well understood.

This project studies a special case of secondary structure interaction, paired helical motifs. The underlying hypothesis of this work is that there are different types of paired 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 paired helical motifs needs to be introduced. This can be achieved by unsupervised learning methods such as clustering. The first step in clustering is to understand the structure of the underlying space. This is the goal of the work presented in this paper. We use fractal dimension, multidimensional scaling and the fastmap algorithm to study this structure. We find that the dimension of the underlying space is between 1.5 and 3. Our results also show that the main features distinguishing two paired helical motifs are the crossing angle and the inter-helix distance. These features correspond to features structural biologists would use to characterize the structures. They are automatically determined by the MDS analysis. This information is very valuable for the next stage of the project. It will simplify the classification of the structures by allowing us to reduce the dimensionality of the problem.

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. (e.g. see fig. 1). 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 this structure determines the function of the protein. (c.f. (Branden & Tooze, 1999)). 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 (Streyer, 1995). Only a small fraction of the total number of known 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 are a large number of known protein structures, one of the main approaches to protein structure prediction is based on machine learning and data mining techniques (Baldi, 1999). Since directly predicting the structure is much too difficult (some studies suggest NP-hard (Unger & Moult, 1993)), it is important to solve intermediate steps such as finding repeating structural motifs that can be predicted from

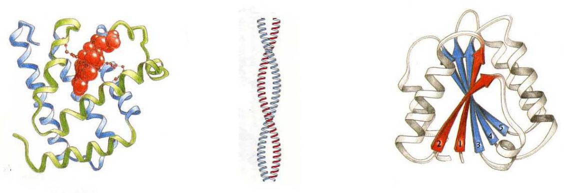

Figure 1. Different schematic representations of protein structures. The protein is represented by the helical thin bands on the first picture from the left. We can see several paired helical motifs (helices crossing at different angles). The middle picture shows a coiled coil (much smaller scale). The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

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.

The long-term goal of this project is the automatic discovery and examination of sequence rules for structural motifs in proteins. In order to do rule learning on the amino acid sequences we need to have a classification of the structural motifs. There are several known examples of such motifs, however currently there is no systematic classification of the motifs. The current set of known protein structures is too large to do this by hand, so automated ways will need to be applied. In order to perform this classification is important to first understand the basic structure of the data. To our knowledge, no systematic study of this has been done yet. This class project examines a particular subset of structural motifs: *paired helical motifs*. We define paired helical motifs as pairs of 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 measure the similarity of two sets of helix pairs, we calculate the root-mean square distance between corresponding atoms in the two structures. Given this distance measure, we examine the structure of the data space by looking at the correlation fractal dimension. This gives information about the underlying dimensionality of the space. We also use multidimensional scaling and fastmap to map the helix pairs into an Euclidian space which allows us to visualize the data. A successful mapping would allow us to relate the coordinates to structural characteristics which can be interpreted. It would also allow us to identify important features.

Studying the distribution of the paired helical motifs in this space is also very interesting from the biological point of view: there are to competing models on the packing of helices in proteins. The model by Chothia et al (Chothia et al., Oct 1977) suggests that the amino acids on the proteins form 'ridges and groves' that lead to preferences in the arrangements of the helices. Bowie (Bowie, 1997) claims that the distribution is essentially continuous. Currently, it is unclear which model is correct. The results of our study should allow a more systematic examination of this question: a very lumpy distribution of the data would favor Chothia's model, whereas a continuous distribution would support Bowie's argument.

Our study on a set of 7681 helix pairs shows that the dimensionality of the data-space is between 1.5 and 3 depending on the scale we are interested in. When mapped into a Euclidian space, the data points come to lie on a cylindrical surface in the subspace of the three most important directions. The axes of this cylinder are strongly correlated to the distance between the helices in the pair and their crossing angle. The MDS analysis automatically finds these relevant features. If this generalizes for other structural motifs, it is an important step in the direction of automatic discovery and examination of sequence rules for structural motifs. The results obtained are also important because they show that we should be able to use a quite small number of features for the classification in the next step of the project. In addition the mapping into Euclidian space allows us to visualize the data, which will be important for evaluating our classification.

The distribution in the Euclidian space we obtain seems to be continuous. However, a more rigorous examination is needed in to determine if the Bowie or Chothia hypothesis are correct. This study will be greatly simplified by the structure discovered in the data.

This project is a preliminary study for my CALD masters program thesis which will be conducted in the framework of a collaboration between researchers from Carnegie Mellon University as well as the University of Pittsburgh. The people involved in this collaboration are Dr. Caruana from CMU, Professors Rosenberg and Buchanan and Dr. Hodor from University of Pittsburgh.

The remainder of this paper is structured as follows: the next section gives a survey of related work. Section 1.2 describes the methods used to characterize our data-space. It is followed by a section describing the results and a conclusion.

1.1 Related Work

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 figure 1). A large body of work has been dedicated to the prediction of secondary structure from the underlying amino acid sequence, which is a first step toward prediction of the full structure.

Early approaches based on very basic statistical analysis and inspection of sequences reached a level of accuracy of about 60 % (Chou & Fasman, 1977). However, even more sophisticated methods based on analysis of adjacent amino acids did not reach levels more than a few percentage points above this. (Rost & Sander, 1993). This indicated that more global information on the sequence is involved in the encoding of secondary structure. In the meantime, newer approaches based on the use of larger databases, evolutionary information and newer methods have increased the level of accuracy of the secondary structure prediction to on the order of 75 % (Karplus et al., 1998),(Brutlag & Sternberg, 1996),(Kleywegt, 1999). Currently it does not seem possible to do much better than this.

Given that secondary structure can be predicted with reasonable accuracy, an interesting question is to go one level beyond secondary structure and to examine the interactions between elements of secondary structure forming more complex patterns. If rules for the relative position of elements of secondary structures can be derived, we can use then as a stepping stone for prediction of the overall structure of the protein. We refer to such structural units involving the interaction of secondary structure elements as "motifs". Very little research has been done on this topic. There are some studies on the structural units involving two elements of secondary structure dealing with the examination of specific **known** motifs. A more general approach as examined in this project has not yet been attempted to the best of our knowledge. One of the main problems is that there is no systematic classification of the different motifs which could be used for prediction. The work in this project provides the groundwork for this classification by studying the underlying structure of the data.

An early example of a successful prediction of a specific known structural motif is the "EF-hand" by Kretsinger et al. (Kretsinger et al., 1972), a motif of two α-helices joined by a specific loop (c.f. (Branden & Tooze, 1999) for a detailed discussion). The method used for the prediction was an analysis of the underlying stereochemical constraints and derivation of a set of rules. An more recent example of a study of a particular motif is the study of coiled coils (a motif of intertwined of helices (see figure 1)) by Berger and Singh et al. (Berger & Singh, 1997 Fall),(Singh et al., March 1998). They showed that iterative learning methods based on pairwise correlation functions can be used for prediction of such motifs. The example of coiled coils is a special case of paired helical motifs. The approach used in (Berger & Singh, 1997 Fall) differs from the research in this project in that it started with a specific known motif, the coiled coil, and learned rules to predict this motif by labeling structures by whether or not they contained a coiled coil. The question asked here, in the context of the overall project, is if such motifs and the underlying rules can be **automatically** identified by looking at a more general data set, such as structures with paired helices. If the approach examined here indeed works, the rules found by Berger and Singh et al. for coiled coils should show up among the general sequence rules found. To our knowledge, no previous research on the automatic discovery of rules for prediction secondary structure motifs has been done.

1.2 Approach

The main methods we use to characterize our dataset are the fractal dimension, multidimensional scaling and fastmap. In the dataset we have, we only know the pairwise distance between structures, but we do not have a coordinate representation. The fractal dimension will allow us the better understand the structure of the space. A mapping of the data into Euclidian space using fastmap and multi-dimensional scaling gives a good starting point for the classification in the next step of the project. In particular, the mapping gives us a greater flexibility in choosing our classification algorithm (methods such as EM and k-means are difficult to use without an explicit coordinate representation of the data). From a more practical point of view having a coordinate representation also greatly reduces the need for storage space, which is an important factor in our project, given the size of the dataset.

1.3 Distance Metric

In order analyze the data-space, we need to have a distance metric to compare pairs of helices. Given two sets of helix pairs, different similarity measures between the pairs 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 helixes are 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. For identical structures it is zero) This minimal rms distance is used as the distance function.

1.4 Calculation of the fractal dimension

The pairwise rms distances between structures can be used to calculate the fractal dimension of the underlying space. The fractal dimension of a set of data points is a useful measure of the deviation from the uniform distribution of the points. It also gives a measure of the intrinsic dimensionality of the data. This is closely related to the number of features needed to describe the data.

Let $N(r)$ be the number of pairs of points that are at most a distance r from each other. Then the fractal dimension can be measured by calculating

$$d = -\frac{\delta log(N(r))}{\delta log(r)} = constant. \tag{1}$$

This can be estimated by measuring the slope of the line $log(N(r))$ vs $log(r)$ for the dataset. The shape of the line allow us to determine if the distribution is continuous (straight line) or lumpy (segments with different slopes and plateaus). It can be used to get an idea of the underlying structure of a dataset (Faloutsos & Kamel, May 1994), (Traina et al., 2000).

1.5 Multi-Dimensional Scaling (MDS)

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 allows a dimensionality reduction of the problem. This allows easier visualization and also application of methods which rely on a coordinate representation (e.g. k-means for clustering). The following description of the methods closely follow the one given by Cox (Cox & Cox, 1994).

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 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

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

f is a monotonic function of the distances. There are several methods of solving this problem. They generally involve minimizing a stress function of the form

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

The specific method used for this project is as follows: Let us define the matrix A as $A_{ij} = -1/2\delta_{ij}^2$, with δ_{ij} as defined above. We define a second matrix $\mathbf{B}$ as $\mathbf{B} = \mathbf{X}\mathbf{X}^\mathbf{T}$ where $\mathbf{X} = [x_1, x_2, \ldots, x_p]^T$. These are the coordinates of the points we want to determine. It can be shown that

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

where H is given by $\mathbf{H} = \mathbf{I} - N^{-1}\mathbf{1}\mathbf{1}^\mathbf{T}$ with the vector $\mathbf{1} = (1, 1, \ldots, 1)^T$ of length N. Since 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 $\mathbf{V} = [v_1, \ldots, v_N]$ is the matrix of eigenvectors. The problem of finding the coordinates x_i therefore reduces to solving the eigenvalue problem 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 eigenvectors found are equivalent to those found by a principal component analysis in the projection space.

1.6 Fastmap

Fastmap (Faloutsos & Lin, June 1995) is another algorithm to map object with given pairwise distances into a p-dimensional space. Given a set of N points and an $N \times N$ distance matrix, it allows to approximate the positions for the points in a k-d Euclidian space such the relative distances are conserved. The basic idea of the algorithm is the following: We can pretend that the points are embedded in an unknown n-dimensional space and try to project them onto k-mutually orthogonal directions. In order to do this, we select two 'anchor' points that are far away from each other. The line between these two points in n-d space is our first direction. If we assume the space is Euclidian and that the triangle inequality holds for the points, we can use the *cosine law* from basic trigonometry to project the other points onto this line. We thus obtain the corresponding coordinates in the first direction. Given these coordinates, we can calculate the distances between all points in the hyperplane orthogonal to the first direction. The procedure is then repeated with these new distances until we have k orthogonal directions. This method is not quite as accurate as MDS,

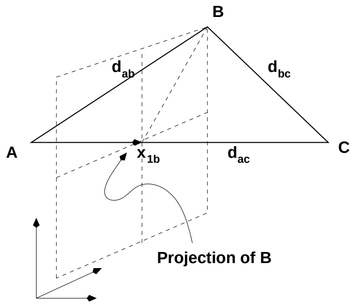

Figure 2. Projection of a point B onto the line between two anchor points A and C)

but the big advantage is that the complexity of the algorithm is $O(Nk)$ whereas MDS is $O(N^2)$.

2. Results

The main goal of our calculations is the exploration of the structure of the dataset in view of applying unsupervised learning methods to obtain a classification. Particularly interesting questions in this respect are

- What is the dimensionality of the data?
- What structural features are most useful in describing the data?
- What does the distribution of the data look like?

These questions were investigated using the methods described in section 1.2.

Our dataset consisted of 7681 pairs of helixes. The data was obtained by extracting the coordinates of a sample of 9739 structures of helical paired motifs by searching through all x-ray structures stored in the Brookhaven Protein Database. The coordinates for the helices were located using the *define_structure* program by Richards and Kundrot (1988). We then calculated the rms distance between all pairs using the ProFit program by Dr. Andrew C. R. Martin (Martin, 1998). For each structure, we also obtained an approximation of the error from the quality measure by Hodor et al. (Hodor et al., 2001). After excluding structures with large errors from the dataset we were left with 7681 structures and the corresponding rms-distance matrix. This was the data used for the calculations below.

2.1 Fractal dimension

Figure 3 shows the plot for the correlation fractal dimension according to equation 1. We can see four different regions: a fairly flat part for distances below 1 Ångström rms-distance, then sections with a slope of about 3 between 1 Å and 2 Å and with a slope of about 1.5 between 2 Å and 6 Å. For distances greater than this, the plot is flat again.

The non-zero flat part in the beginning indicates that there are some structures that are almost identical. This is consistent with what we know about the

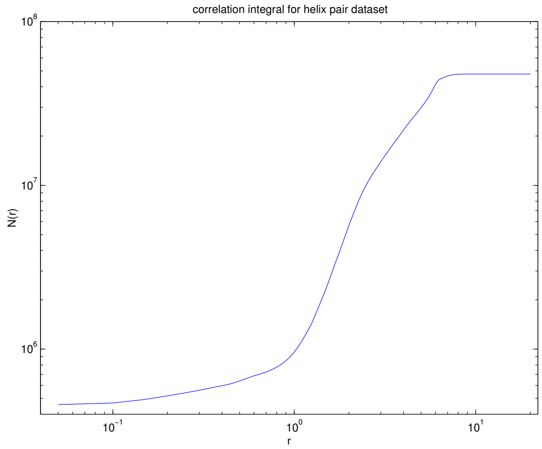

Figure 3. Calculation of correlation integral for paired helix dataset

dataset: there are a lot of redundant structures for well studied proteins such as hemoglobin in the Brookhaven Protein Databank. These proteins have very similar structures for different species, and there are also a lot of proteins that differ only by a small number of mutations in their amino acid sequence, which results in an almost identical structures. So we would expect there to be a flat part in the beginning.

Given that the fractal dimension for distances between 1 and 6 Å, where most of the data is concentrated, is between 1.5 and 3, we can expect that we should be able to characterize the data with very few variables. This is assuming that the hyperplane on which the points lie is not too complicated[1].

The fractal dimension plot does not show any plateau, which would be a strong indication of a very lumpy nature of the data. However, the absence of plateaus does not mean that the space is perfectly continuous.

2.2 Multi-dimensional scaling

For the multi-dimensional scaling calculation, we used a program written by Dr. Caruana and Dr. Hodor. Due to memory limitations on the computers used for the calculations, we could not explore more than 10 dimensions. However, given the results of the fractal

[1]Note, the fractal dimension only gives information about the dimensionality of the hyperplane, not the shape. I.e. it does not have to be flat

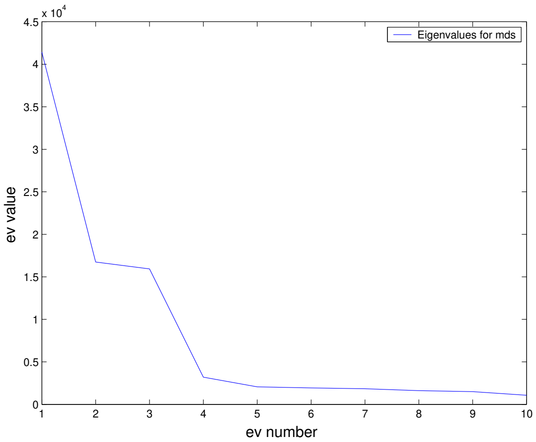

Figure 4. Plot of the first 10 eigenvalues for the MDS calculation

dimension calculations, this does not seem to be a severe limitation. Figure 4 shows a plot of the first 10 eigenvalues resulting from the MDS calculation. We can see that the eigenvalues drop very sharply (from 41434 for the first value to 3204 for the fourth) over the first three dimensions, and then decrease more slowly (to 1075 for eigenvalue 10). The lower, blue curve in figure 5 shows the mean absolute percentage error $< |\Delta d|/\delta >$ in the distance predicated by MDS with respect to the correct distance. We have

$$< |\Delta d|/d >= 1/N \sum_{i,j=1}^{N} |d_{ij} - \delta_{ij}|/\delta_{ij}, \qquad (6)$$

where $N = 7681$, d_{ij} is the predicted distance, and δ_{ij} is the actual distance. Again we can see that the error flattens out after the fourth dimension to a value of about 10%. This seems to be a reasonable approximation given that the original distances are not exact. It is consistent with the fractal dimension calculation. (We note that the error does not go to zero. This is not surprising given that the data is not exact. The underlying x-ray measurements of the structures are limited in precision, so we can't expect the error to disappear.)

Figure 7 shows projections of the data onto the different eigen-directions. The graphs suggest that in the space defined by the first three eigen-directions, the majority of points are located on a cylindrical surface. For the remainder of the report we will refer to the coordinates resulting from MDS as c_i where i is the number of the corresponding eigenvalue in decreasing order. We examined different helix pair features such as crossing angle, distance, offset[2](see figure 6) to find

[2] How far is the center of the helix from the point where the helices are closest?

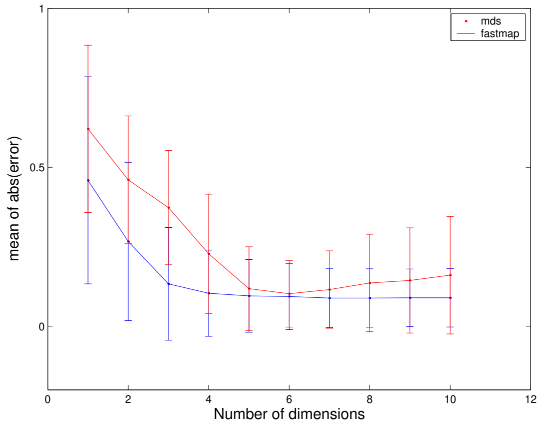

Figure 5. Average percentage error on the distance for the MDS and fastmap methods

an interpretation of the coordinates. The axis of the cylinder, almost parallel to c_2, is strongly correlated to the inter-helix distance. Indeed a linear regression analysis showed an r value of $r = 0.8127$ for a regression of the distance with c_2. The position around the cylinder seems to be strongly correlated to the crossing angle. Since this relation is elliptical rather than linear, a linear regression analysis is not very useful in this case. The lower two plots on figure 7 clearly show the very strong correlation. However, a more rigorous study is needed.

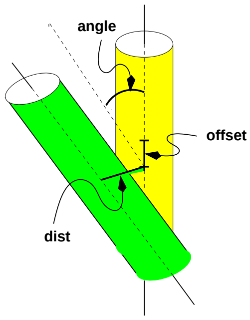

Figure 6. Schematic illustration of angle, distance and offset of two helices)

The other projections showed some interesting patters, but the interpretation of the other directions is not clear at this point in time. In particular, we have not yet found a good interpretation of the radius of the cylinder. A systematic one- and two-value regression analysis between angle, distance and crossing offset by Dr. Caruana showed that the strongest single value correlations was indeed the inter helix distance and the second eigen-direction c_2. A weak correlation ($r = 0.5684$) was found between the crossing angle and the first eigen-direction c_1. All the other r values were below 0.15. Adding a second variable to the regression analysis did not significantly increase the r-values. Any combination of c_2 and another eigen-direction showed a strong correlation to the distance, and combinations involving c_1 showed weak correlations to the angle.

We can therefore conclude that the most relevant features for distinguishing between paired helical motifs are the crossing angle and the distance. It is somewhat surprising that there are no strong correlations between the offset and any of the MDS coordinates. However, the data was chosen such that helix pairs with too big an offset were excluded. So it is possible that the variation in the offset in the dataset is too limited to show up.

The important point to note here is that the relevant coordinates of the MDS correspond to some of features that would be picked intuitively by structural biologists. If this generalizes to other structural motifs, MDS allows us to automatically pick relevant features, which would greatly simplify the study of these motifs.

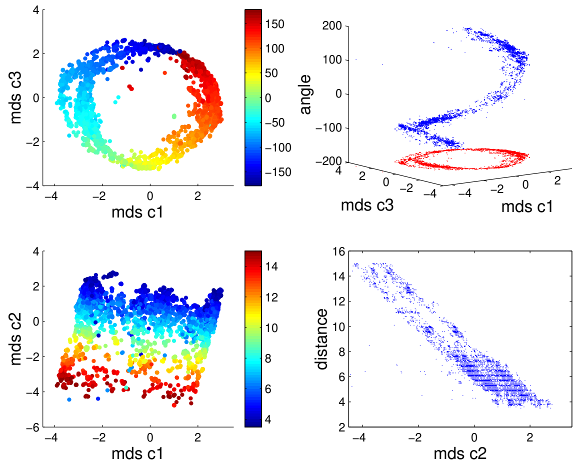

Figure 7. Projections of the data onto first three coordinates provided by the multi-dimensional scaling calculation. The top two graphs show the first and third eigen-directions. The color coding on the left plot shows the crossing angle between the helices. On the left plot the angle is represented by the vertical axis. The plot show a projection of the data into the (c_1, c_3) plane (red) as well as a 3-d plot (blue). The helical shape of the curve indicates a strong correlation with the crossing angle. The bottom two plots clearly show the correlation between the inter-helix distance and the second eigen-direction)

2.3 Fastmap

At the beginning of the analysis we were worried that the MDS calculation might be too slow for our dataset. We therefore did some studies with Fastmap. The code used was provided by Dr. Faloutsos. It turned out that the even though Fastmap was almost a factor of 30 faster than the MDS calculation (2 minutes vs. 1 hour on an Pentium based linux machine), that calculation time for the MDS was reasonable. Since MDS gives chooses the coordinates along principal components, its results were easier to interpret than the fastmap results. Figure 5 also shows that MDS is in general more accurate (though given the standard deviations not significantly).

The results of the fastmap calculations were very similar to those of MDS: Examination with the 3-D plotting program *xgobi* showed that the shape of the data locations in appropriate three dimensional projections was also cylindrical. This is a valuable confirmation of the MDS results. However, since the coordinate directions chosen by fastmap are somewhat random, the axis of the cylinder did not correspond to the coordinate directions as in the case of MDS. This explains why fastmap requires more dimensions (5) to obtain a similar mean error on the distance prediction (see figure 5).

A peculiar characteristic of the fastmap mean error is that it has a minimum for 6 dimensions and then starts to increase. This might be due to the fact that the data is inherently noisy, since it is experimental data. This could result in the triangle inequality not being satisfied for points that are close together. Since fastmap assumes that the triangle inequality holds, it will be prone to errors in this case.

3. Conclusion

This project is part of a study of the application machine learning techniques to the characterization and prediction of interaction of secondary structure in protein. We have performed an analysis of the structure of the data-space defined by paired helical motifs. This analysis should facilitate the classification of the structures into meaningful groups by giving us more freedom on the choice of the classification method and by allowing us to visualize the data. We have used the fractal dimension, multidimensional scaling and the fastmap method to show that different paired helical motifs can be well distinguished by a small number of features. We found that

- The intrinsic dimensionality of the data is between 1.5 and 3.

- The data is well characterized by a small number of features, the most important ones being the the crossing angle of the helices and the inter-helix distance.

- The distribution of the data seems to be continuous. Though this question needs further investigation.

We also found that the application of MDS allows to automatically find the relevant features describing the data. These results are specific to the dataset examined. However, if the approach generalizes to other motifs, it will greatly simplify the study of these motifs. This question warrants further investigation.

There are two main directions in which the research in this project can continue. The first will be to perform unsupervised learning on the paired helical motifs based on the structure of the data discovered during this project. The resulting classification can then be used do rule learning on the amino acid sequence. If completed successful, the research conducted in this project can be applied to other structural motifs. The extracted rules will be interpretable by domain experts. This will improve the understanding of the mechanisms governing the interaction of secondary structure in proteins, which would be an important step in protein structure prediction.

The second direction is to do a statistical study of the distribution of the data in the projection space from MDS. We can do this by projecting the data onto the surface of the cylinder. Given this projection we can statistically test the the predictions of the hypothesis' by Bowie versus those of Chothia and determine which better explains the data. This study is of great interest to the structural biology community.

References

Baldi, P. F. (1999). *Bioinformatics, the machine learning approach.* Cambridge: MIT Press.

Berger, B., & Singh, M. (1997 Fall). An iterative method for improved protein structural motif recognition. *J. Comp. Bio.*, *4(3)*, 261–73.

Bowie, J. U. (1997). Helix packing angle preferences. *Nature Structural Biology*, *4, no. 11*, 915–917.

Branden, C., & Tooze, J. (1999). *Introduction to protein structure, second ed.* New York: Garland Publishing.

Brutlag, D. L., & Sternberg, M. J. (1996). Sequences and topology challenges for algorithms and experts. *Current Opinion in Structural Biology*, *6*, 343–345.

Chothia, C., Levitt, M., & Richardson, D. (Oct. 1977). Structure of proteins: Packing of α-helices and pleated sheets. *Proc. Natl. Acad. Sci. USA*, *74, No. 10*, 4130–4134.

Chou, R., & Fasman, G. (1977). β-turns in proteins. *J. Mol. Biol*, *115*, 135–175.

Cox, T. F., & Cox, M. A. (1994). *Multidimensional scaling.* London: Chapman & Hall.

Faloutsos, C., & Kamel, I. (May 1994). Beyond uniformity and independence: Analysis of r-trees using the concept of fractal dimension. *Proc. ACM SIGACT-SIGMOD-SIGART PODS* (pp. 4–13). Minneapolis, MN.

Faloutsos, C., & Lin, K. (June 1995). Fastmap: A fast algorithm for indexing, data-mining and visualization of tr additional and multimedia datasets. *Proceedings of 1995 ACM SIGMOD, SIGMOD RECORD* (pp. 163–174).

Hodor, P., Caruana, R., Buchanan, B., & Rosenberg, J. (2001). Automated estimation of protein structure quality and its application in the selection of representative structure sets. *to be published.*

Karplus, K., Barrett, C., & Hughey, R. (1998). Hidden markov models for detecting remote protein homologies. *Bioinformatics*, *14*, 846–856.

Kleywegt, G. J. (1999). Recognition of spatial motifs in protein structures. *J. Mol. Biol*, *285*, 1887–1897.

Kretsinger, R., et al. (1972). The structure of a calcium-binding protein from carp muscle. *Cold Spring Harb. Syymp. Quant. Biol*, *36*, 217–20.

Martin (1998). Protein fitting program. *http://www.biochem.ucl.ac.uk/ martin/.*

Rost, B., & Sander, C. (1993). Protein structure perdiction:methods and protocols. In D. Webster (Ed.), *Third generation prediction of secondary structure.* Totowa, NJ: Humana Press.

Singh, M., Berger, B., Kim, P., Berger, J., & Cochran, A. (March 1998). Computational learning reveals coiled coil-line motifs in histidine kinase linker domains. *Proc. Natl. Acad. Sci USA*, *95*, 2738–2743.

Streyer, L. (1995). *Biochemistry.* Freeman.

Traina, C., Traina, A., Wu, L., & Faloutsos, C. (2000). Fast feature selection using fractal dimension. *citeseer.nj.nec.com/372645.html.*

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