Finding suitable clusters for rule learning in protein structure prediction

Marc Fasnacht
mf5k@andrew.cmu.edu
Physics Department
Carnegie Mellon University
Pittsburgh, PA 15213

March 10, 2001

Abstract

We propose research on protein structure prediction. Protein structure prediction in 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. 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, will have to be understood better.

This project is part of a larger research collaboration, which studies the application rule learning methods to the characterization and prediction of interaction of secondary structure in proteins. The particular case studied here is that of 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 restrictions governing the underlying amino acid sequence of the protein. The long-term goal is to automatically discover such rules.

The current project will focus on separating examples of different types paired helical motifs into groups, which will exhibit the kind of rules we are looking for. We will apply agglomerative clustering and the recently introduced framework of meta-clustering in order to find such sets of structures. The classification imposed by the clustering will be used for rule learning based on the amino acid sequence. The extracted rules will be interpretable by domain experts, and can be used to predict the presence of the particular structural motif. If successful, the research conducted in this project can be applied to other structural motifs. This will improve the understanding of the mechanisms governing the interaction of secondary structure in proteins. A set of rules about structural motifs could greatly enhance protein structure prediction.

1 Introduction

1.1 Problem

This proposal addresses the problem of the prediction of protein structure from amino acid sequence. 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. [1]). It is relatively easy to measure the amino acid sequence of a protein experimentally so that most protein sequences are known. The three dimensional structure on the other hand is extremely costly and time intensive to determine [2]. Only a small fraction of the total number of known proteins has been solved.

Proteins are molecules play essential 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 extremely valuable in pharmaceutical research.

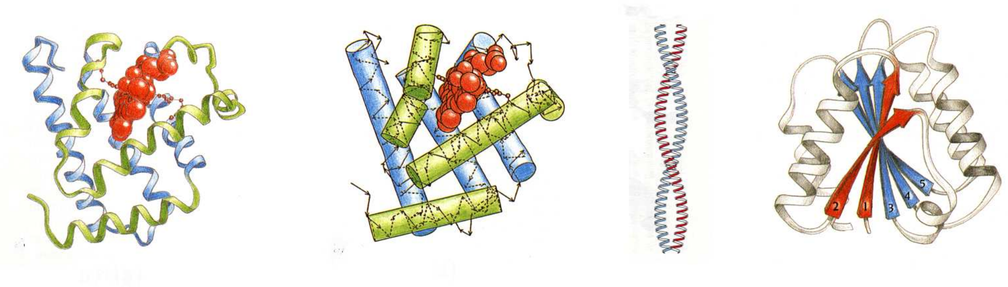

Figure 1: Different schematic representations of protein structures. The protein is represented by the helical thin bands on the first picture from the left. We can see several paired helical motifs (helices crossing at different angles). In the second pictures of the same protein the helices are drawn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

Given that there are a large number of known protein structures, one of the main approaches to protein structure prediction is based on applying machine learning and data mining techniques [3]. Since directly predicting the structure is much too difficult (some studies suggest NP-hard [4]), 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.

This project is part of larger research effort which examines the use of machine learning techniques to automatically find structural motifs and to indetify patterns in their amino acid sequence that allow the prediction of the presence of the motif. There are several known examples of such motifs. The particular structural motifs studied are *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.

1.2 Approach and deliverables

The basic approach is to use a distance metric between sets of paired helical motifs to group the structures into structurally very similar sets (e.g. crossing at a similar angle), using different clustering methods. The classification defined by the clustering will be used to label the helix pairs for rule learning. It should then be possible to learn rules for each of the sets concerning the underlying amino acid sequence that leads to this type of paired helical motif. The main scope of this project is to find good clusterings for the rule learning. We will use meta-clustering to examine this problem. In particular we will

- determine the fractal dimension of the data space underlying the clustering in order to better understand the structure of the data.
- extract a set of promising clusterings and perform rule learning on these clusters.
- try to find rules characterizing the amino acid sequence for the cluster, which can be interpreted by a domain expert.

1.3 Impact

If successful for paired helical motifs, the research conducted in this project can be used as a general tool to find other structural motifs, and to derive rules to predict the occurrence of these motifs. A systematic exploration of such rules should greatly improve the understanding of the mechanisms governing the interaction of secondary structure. In particular a sufficiently large set of such rules might impose sufficient constraints on the structure of a protein to predict it.

2 Related Work

In biology the spacial structure 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 % [5]. However, even more sophisticated methods based on analysis of adjacent amino acids did not reach levels more than a few points above 60 %. [6]. This indicated that more global information on the sequence is involved in the encoding of the 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 % [7]-[10]. 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 addressed in this project, 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 will refer to structural units involving the interaction of secondary structure elements as "motifs". Very little research has been done on this topic. Most deals with the examination of specific known motifs.

An early example of a successful prediction of such a structural motif is the "EF-hand" by Kretsinger et al. [11], a motif of two α-helices joined by a specific loop (c.f. [1] 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) which has been by Berger and Singh et al. in [12] and [13]. 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 [12] 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 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.

3 Approach

This project will be part of my CALD master's 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 main thrust of this ongoing research effort is the automatic discovery and examination of sequence rules for structural motifs in proteins. There are two main problems: the first deals the classification of the structural motifs into groups. Even though some motifs are known, the current dataset is too large to do this by hand in a systematic way. Automated methods will need to be developed. The second problem is the rule-learning problem based on the above-mentioned structural classification for the labeling. The project currently focuses on a specific set of motifs: the case of paired helices (i.e. two helices with a mutual interface). Figure 1 shows and example a protein structure with such paired helical motifs.

3.1 Preliminary Work

Preliminary work in the context of this collaboration has been done by Sassaman [19]. He examined rule learning for a particular paired helical motif, so called 4-4 helix pairs, on a set of 606 unique protein structures. The 4-4 motif is characterized by specific relative position (i.e. angle,distance). The rule learning was done using the symbolic rule learning program RL, developed in Prof. Buchanan's laboratory [18]. We plan to use RL in the later stages of the project. He established that: 1) RL can find rules for both target and non-target classes; 2) The rules found often coincide with statistical amino acid distribution; 3) The rules produces were interpretable by a domain expert.

The study by Sassaman also showed that in order to derive relevant rules is very important to have high quality structures. The structural differences between different paired helical motifs can be small, and can only be detected given accurate structures. For the current project, a better dataset was assembled. The quality the selected structures needed to be assured in two respects:

1. They need to be of high crystallographic quality in the sense that the error on the positions of the atoms needs to be small. For this purpose, a new quality measure was developed by Hodor et al. [20] which uses a weighted combination of crystallographic quality measures.

2. The structures need to be different. Protein structure in the form of atom coordinated can be downloaded from a public database, the Brookhaven Protein Data Bank (PDB) at *http://www.rcsb.org/pdb/*. The PDB contains a lot of almost identical structures. Having such structures in the dataset would be problematic for the rule learning. To ensure independence of the structures they were picked from different levels of CATH ([21]) which is a taxonomy of protein structure according to class, architecture, topology and homology.

641 protein structures were extracted from the PDB by picking the best structure according the new quality measure from each CATH levels.

3.2 Methods

3.2.1 Clustering and Meta-Clustering

The main problem given these structures is to find subset of helix pairs similar enough (i.e. rms distance smaller than 1 Angstrom) to make sure that they belong to the same motif so that the classification makes sense. At the same time we need to retain enough structures in the group to be able to apply machine-learning techniques. We see two main approaches to resolve this tension: the first is to use clustering to group structures with similar motifs together. This will be the main focus of the current stage of the project. The second will be to expand the data set by using proteins with homologous sequences that are known to have a high degree of structural similarity. This will part of a later stage of the project.

Distance Metric In order to do clustering, 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. One could try to extract features such as the crossing angle and the distance.

However, given that the shape of the individual helices can be very complicated (straight, bent, etc) these are ill-defined. A more general and widely used measure 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 distance (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.

Hierarchical Agglomerative Clustering (HAC) Given a distance function defined between two points, HAC starts by putting all the points in the dataset in singleton clusters (i.e. number of clusters is the number of data points). It then proceeds to merge the clusters that have the smallest distance between each point in one cluster and each point in the other. This is repeated until we have one cluster containing all the points. HAC can be described by a dendrogram or clustering tree that can serve to visualize the clustering. For a description of HAC, see [23] or [22]. Figure 2 shows an example.

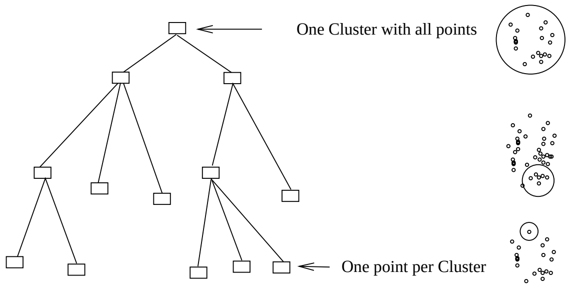

Figure 2: Schematic picture of an agglomerative clustering tree. The leaf nodes correspond to "clusters" consisting of only one data point. The root node represents one single cluster containing all the points.

Meta-Clustering Meta-clustering is a new technique introduced by Caruana et al [22]. The main insight is that there is usually no unique "best" clustering of a dataset. Therefore, the data is clustered in many different ways, for example by varying the distance metric or by using subsets of the data. Given these different clusterings, a distance function between the clusterings themselves is introduced. It puts similar clusterings close together whereas vastly distinct clustering will be far apart. Using this function a new clustering at the meta level is performed. This allows to organize the clusterings of the base-level data into groups of similar clusterings. This meta-clustering can be used to let the user navigate efficiently between different clusterings in order to find an appropriate one.

Fractal Dimension The fractal dimension of a set of data points is a useful measure of the deviation from the uniform distribution of the points. It can be used to get an idea of the underlying structure and dimensionality of a dataset [24], [25]. The fractal dimension of a finite set of data points can be determined by dividing the underlying D-dimensional space of the data into (hyper-cubic) grid cells of side r. Let $N(r)$ be the number of cells that contain one or more data point. Then the fractal dimension can measured by

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

This can be estimated by measuring the slope of the line $log(N(r))$ vs. $log(r)$ for the dataset.

3.3 Proposed Work

Study of Meta-Clustering on data-set In the case of the protein data set, there are numerous ways of grouping similar structures together. It is by no means clear which is the ideal way of clustering the data. The focus of this project is to better understand the clustering with the goal of identifying clusters that represent separate structural motifs. This will then give us a classification for rule learning. The framework of meta-clustering as described above seems promising in this respect. It allows to efficiently examine and compare different types of clusterings. In particular, the meta-clustering should allows to more easily find a good labeling of the data. Meta-clustering was introduced only very recently and though it has been applied to the protein dataset is not very well understood and warrants further study. In order to apply it to the protein data set and to get a better understanding of the structure introduced by meta clustering, we propose to study the fractal dimension of the underlying as well as the meta-clustering space and to compare it to the dimension of the physical space in which the protein structures live. The fractal dimension gives valuable information about the 'lumpiness' of the data and can be used to evaluate the clustering. An interesting question is how the fractal dimension varies with the different metrics used for the meta-clustering.

Finding useful clusters The next step will be to find clusterings that are most suitable for rule learning. The approach we plan to take here is the following: the underlying clustering of the structures is agglomerative so that we can produce a hierarchical tree of the clusters for a given clustering. Different branches of the tree thus represent clusters of different size (c.f. fig 2). We can now apply the rule learning program RL at all different levels of this tree. RL will return a statistical measure of how confident it is that it actually learned something useful. This number can then be used to "paint" the tree, the different colors indicating with what confidence rules can be learned on a particular branches (corresponding to clusters) of the tree. This color code can then be used to visualize and navigate the tree in order to identify clusters which with high likelihood represent interesting structural motifs. The preliminary work by Sassaman showed that such motifs do exist. The coloring should help finding them. This painting can be done for any given agglomerative clusterings. It would take very long to try out all possible clusterings, but the meta-clustering can be used to suggest clusterings that are different enough to yield different interesting rules.

Scope of the KDD-Lab project Given the time scope of this current project, I propose to examine the following subset of questions:

- What is fractal dimension of the data space and how does it relate to the fractal dimension of the meta-clustering space.
- Given meta-clustering, can we pick a few promising clusterings and do rule learning on them as explained above.

Given the results of the results of preliminary studies mentioned above, we expect that this approach should be able to find some interesting structural motifs. The evaluation of the method will be done in two stages: There is a small number on known helical paired motifs. The suitability of the clustering can be examined by checking if the known motifs are among the clusters. The rules found by RL are in a form that can be examined by a domain expert. It is possible to determine whether or not the sequence rules are in agreement with the known stereochemical properties of the encoded amino acids. The evaluation of the rule learning will therefore be done through an examination rules found by the domain expert, Professor Rosenberg.

4 Work Plan

The time line for the proposed work is as follows:

- The first two weeks will be used to familiarize myself with the structure of the data and to learn how to use the different programs (Meta-Clustering, RL)
- The next two weeks will be used for the study of the fractal dimension using the Meta-clustering.
- The remaining time will be used to the painting of the tree using RL. This study will probably not be completed by the April 2nd deadline for this class, but I expect to have preliminary results.

The results of this project will form the basis for continuing research in the framework of my CALD master's thesis. They will be summarized in form of a research paper to be completed by April 2nd.

5 Summary

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. The proposed study of paired helical motifs should allow us to automatically find rules governing the underlying amino acid sequence of the protein which will help in understanding the chemical and geometrical restrictions imposed by the structural motif. Preliminary research has shown that rule learning techniques can indeed be used to find rules in line which expert predictions.

The current project will focus on separating examples of different types paired helical motifs into groups, which will exhibit the kind of rules we are looking for. The framework of meta-clustering and agglomerative clustering should allow a better understanding of the structure of the dataset in order to find such sets of structures. The sets of structures will then be used for rule learning based on the amino acid sequence. The extracted rules will be interpretable by domain experts, and can be used to predict the presence of the particular structural motif. If successful, the research conducted in this project can be applied to other structural motifs. This will improve the understanding of the mechanisms governing the interaction of secondary structure in proteins. A set of rules about structural motifs could greatly enhance protein structure prediction.

References

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

[2] L. Streyer, *Biochemistry*, Freeman (1995)

[3] Pierre F. Baldi, Bioinformatics, the Machine Learning Approach, MIT Press, Cambridge, 1999.

[4] 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).

[5] R.Y. Chou and G.D. Fasman, "β-turns in proteins. J. Mol. Biol 115: 135-175, 1977.

[6] B. Rost and C. Sander, "Third generation prediction of secondary structure" in *Protein structure prediction: methods and protocols*, D. Webster, Eds, Humana Press, Totowa, NJ (1993).

[7] , Karplus, K., Barrett, C. and Hughey, R. (1998). "Hidden Markov models for detecting remote protein homologies. Bioinformatics", 14, 846-856.

[8] Rurkhard Rost, Protein secondary structure prediction continues to rise. http://cubic.bioc.columbia.edu/papers/2000_opinion/paper.html.

[9] Douglas L. Brutlag and Michael JE Sternberg, "Sequences and topology challenges for algorithms and experts", Current Opinion in Structural Biology, 1996, 6:343-345.

[10] Gerard J. Kleywegt, "Recognition of Spatial Motifs in Protein Structures", J. Mol. Biol (1999) 285, 1887-1897.

[11] R.H. Kretsinger, et al. "The structure of a calcium-binding protein from carp muscle." Cold Spring Harb. Symp. Quant. Biol, 1972. 36: p.217-20.

[12] B. Berger, M. Singh, "An iterative method for improved protein structural motif recognition", J. Comp. Bio. 1997 Fall; 4(3):261-73.

[13] M. Singh, B. Berger, P.S. Kim, J.M. Berger and A.G. Cochran, "Computational learning reveals coiled coil-line motifs in histidine kinease linker domains", Proc. Natl. Acad. Sci. USA, Vol. 95, pp 2738-2743, March 1998.

[14] Cyrus Chothia, Michael Levitt and Douglas Richardson, "Structure of proteins: Packing of α-helices and pleated sheets", Proc. Natl. Acad. Sci. USA, Vol 74, No. 10, 4130-4134, Oct. 1977.

[15] James U. Bowie, "Helix Packing angle preferences", Nature Structural Biology 4, no. 11 (1997): 915-917.

[16] Marcel Turcotte, Stephe H. Muggleton and Michael J.E Sternberg, "Automated Discovery of Structural Signatures of Protein Fold and Function. http://citeseer.nj.nec.com/379146.html.

[17] S. Muggleton and L.D. Raedt, "Inductive logic programming: Theory and methods", Journal of Logic Programming, 19/20, 629-679.

[18] F. Provost and B.G. Buchanan, "Inductive Policy, The Pragmatics of Bial Selection", Machine Learning, Vol. 20, Nos. 1/2, pp. 35-61.

[19] Eric Sassaman, "Examination of Sequence Rules for α Helix Packing in Proteins", 1998, Masters Thesis in Computer Science, University of Pittsburgh.

[20] Paul Hodor, Richard Caruana, Bruce G. Buchanan and John M. Rosenberg, , "Automated Estimation of Protein Structure Quality and Its Application in the Selection of Representative Structure Sets", to be published.

[21] C.A. Orengo et al., "CATH– a hierarchic classification of protein domain structures", Structure, 1997.5(8):pp. 1093-1108.

[22] Rich Caruana, Pedro Artigas, Anna Goldenberg, Anton Likhodedov, "Meta Clustering", submitted for publication, ICML2001.

[23] Richard O. Duda,Peter E. Hart and David G. Stork, "Pattern Classification and Scene Analysis: Pattern Classification", Wiley, John & Sons, 2000

[24] Christos Faloutsos and Ibrahim Kamel, "Beyond Uniformity and Independence: Analysis of R-trees Using the Concept of Fractal Dimension", Proc. ACM SIGACT-SIGMOD-SIGART PODS, May 1994, pp. 4-13, Minneapolis, MN.

[25] C. Traina Jr., A Traina, L, Wu, C Faloutsos, "Fast feature selection using fractal dimension, http://www.cs.cmu.edu/afs/cs.cmu.edu/usr/christos/www/PUBLICATIONS/sbbd2k.pdf, 2000.