Finding suitable clusters for rule learning in protein structure prediction

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

February 5, 2001

Abstract

We propose research on protein structure prediction. A lot of work has gone into predicting the secondary 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 particlar case studied here is that of of helix crossings. The underlying hypothesis of this work is that there are different types of helix crossings 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 helix crossings into groups which will exhibit the kind of rules we are looking for. We will apply the framework of meta-clustering and agglomerative clustering in order to find such sets of structures. The clusters of structures will then be used for rule learning based on the amino acid sequence. If 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.

1 Introduction

1.1 Problem

This proposal addresses the problem of the prediction of protein structure from amino acid sequence. Proteins are molecules play essential roles in most biological processes such transportation, storage and immune protection to name only a few examples.

Chemically, proteins are poly-peptides - 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 whereas the three dimensional structure is extremely difficult to determine [2]. For any kind of applications, such as rational design of drugs, it is essential to understand the three dimensional structure, whereas the sequence is of very limited use. Understanding the underlying rules that govern the relationship between amino acid sequence and protein structure is therefore one of the key unsolved problems in molecular biology.

Given that there exists a large set of known protein structures, one of the main approaches protein structure prediction is based on applying machine learning and data mining techniques [3]. Since directly predicting the structure is much too difficult, it is important to solve intermediate steps such as finding repeating structural motifs which can be predicted from the sequence. If a sufficient number of these motifs can be identified for a given sequence, the structure of the protein could be determined from the its amino acid sequence.

This project is part of larger research effort which examines the use of machine learning techniques to identify patterns on the amino acid sequence that allow the prediction of the presence of certain structural motifs from the amino acid sequence of a protein. The particular structural motif studied are helix-helix crossings.

1.2 Approach and deliverables

The basic approach is to use a distance metric between sets of helix-helix crossing structures to group the structures into structurally very similar sets using different clustering methods. If these sets are similar enough so all the members are of the same helix-helix crossing type, it should be possible to learn rules concerning the underlying amino acid sequence of the protein that leads to this type of crossing. 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 clustering of the data.
- extract a set of promising clusterings and preform rule learning on these clusters.
- try to find rules which characterize the amino acid sequence for the cluster which can be interpreted by a domain expert.

1.3 Impact

If successful, the research conducted in this project can be used as a general tool to find other structural motifs. The extracted rules will be interpretable by domain experts. A systematic exploration of such rules should greatly improve the understanding of the mechanisms governing the interaction of secondary structure in proteins, which would be an important step in the direction of protein structure prediction.

2 Prior Work

The focus of this work is to examine if interactions between elements of secondary structure in proteins. Secondary structure refers to the relative spacial arrangement of amino acids that are near each other in the sequence. A typical example of secondary structure are helices (see figure 1 for an example) or sheets. A large body of work has been dedicated to the prediction of secondary structure from the underlying amino acid sequence. Early approaches based on very basic statistical analysis and inspection of sequences reached a level of accuracy of about 60 % [4]. However, even more sophisticated methods based on analysis of adjacent amino acids did not reach levels more than a few points above 60 %. [5]. 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 methods such as Hidden Markov Models [6] have increased the level of accuracy of the secondary structure prediction to on the order of 75 % [7]-[9]. Currently it does not seem possible to do much better than this.

An interesting question is then to go one level beyond secondary structure and to examine the interactions between elements of secondary structure forming more complex patterns and to see if rules for particular instances of the interaction between such elements can be derived. We will refer to structural units involving the interaction of secondary structure elements as "motifs". An early example of a successful prediction of such a structural motif is the "EF-hand" by Kretsinger et al. [10], 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 such the study of coiled coils (a motif of intertwined of helices) which has been by Berger and Singh et al. in [11] and [12]. 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 helix crossings. The approach used in [11] differs from the research in this project in that it started with a specific know 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 crossing 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.

The systematic study of helix crossing is also interesting from a biological point of view. It is not clear if there are discrete sets of angles at which the helices can cross, as suggested by [13] or if there is a continuum, as argued by Bowie [14]. An extension of this project should allow to answer this question.

On an even more global level, Turcotte et al. [15] have applied machine learning techniques to find rules that are common to a particular fold ("fold" refers to a specific overall structure of a proteins, involving the interaction of **all** elements of secondary structure). Their use of Inductive Logic Programming [16] lead to a set of rules which can be shown to coincide with many established classification principles for structures. Because the rule learning is done on groups of proteins with the same fold, this approach will learn rules governing the global structure of these specific protein. This is different from interaction between specific pairs structural elements, which is the focus of the current project. Such global rules are quite specific will be harder to generalize to new protein structures. However, both are extremely important for the understanding of protein folding.

3 Approach

3.1 Prior Work and Background

This project is part of 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 main thrust of this ongoing research effort is the discovery and examination of sequence rules for α-helix packing in proteins. The people involved in this collaboration are Dr. Caruana from CMU, Professors Rosenberg and Buchanen and Dr. Hodor from University of Pittsburgh.

The general approach is to identify motifs in known protein structures which consist of two elements of secondary structure that with a mutual interface, to then find all the independent manifestations of this motif in the Brookhaven Protein Data Bank (PDB), which is a large collection of known protein structures. From this dataset should then allow to find stereochemical constraints or sequence rules concerning the amino acids associated with the motif. The specific example studied is the case of a crossing of two helices. Figure 1 shows and example a protein structure with such helix crossings.

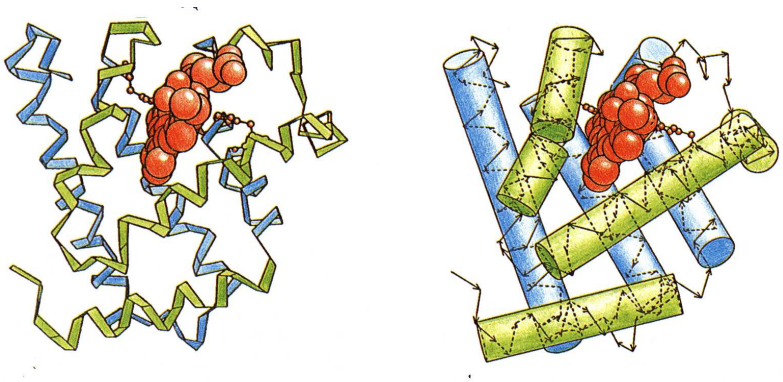

Figure 1: Two different schematic representation of a protein structure with helix crossings. The protein is represented by the helical thin bands on the left picture. In the right picture the helices are drawn as cylinders.

3.1.1 Rule learning

The rule learning in the context of this project will be done using the symbolic rule learning program RL, developed in Prof. Buchanan's laboratory [17]. RL can learn highly non-linear relationships and generate rules which can easily be examined by experts. It is set up to make use of prior information provided by experts as well as semantic relationships among features encoded in the data. Preliminary work using RL on similar data was done by Sassaman [18]. He examined the case of a specific motif, so called 4-4 helix pairs, on a set of 606 unique protein structures. 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.

3.1.2 Selection of dataset

The study by Sassaman also showed that selection of a good dataset for the rule learning is very important. For the current project, a better dataset was assembled. Protein structure can be downloaded from a public database, the Brookhaven Protein Data Bank (PDB) at *http://www.rcsb.org/pdb/*. 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. [19] which uses a weighted combination of crystallographic quality measures.

2. The structures need to be different. 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 ([20]) which is a taxonomy of protein structure according to class, architecture, topology and homology. Members of the same class are expected to be genetically related.

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

3.1.3 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 the the same motif all while retaining enough structures in the group to have a reasonable size set on which 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 which are known to have a high degree of structural similarity. This will part of a later stage of the project.

Distance Metric Given two sets of helix pairs, different distance metrics between the pairs are possible. The distance function currently used is the following: the two pairs of helixes are rotated and translated on top of each other such that the root mean square distance (rms distance) between specific atoms on the backbone is minimized. (This corresponds to finding the orientation of maximal overlap. For identical structures this would be zero) This minimal rms distance is then used as the distance function. This distance function can then be used to structure and cluster the data.

Hierarchical Agglomerative Clustering (HAC) (the following follows a description given in [21]). Given a distance function defined between two point, HAC starts 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 and each points on the other. 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. Figure 2 shows an example.

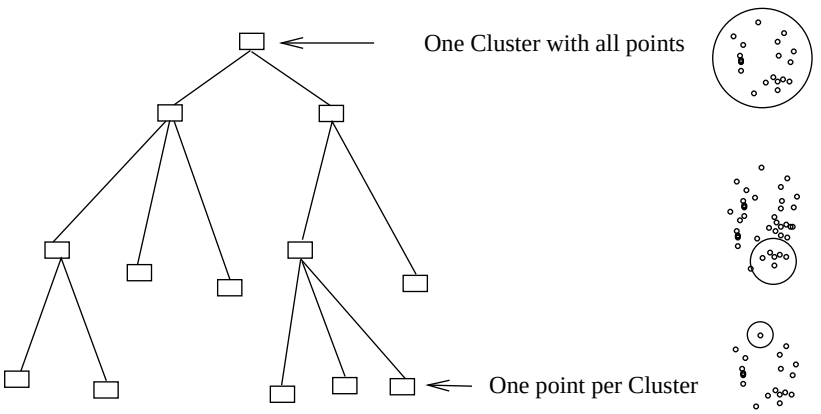

Figure 2: Schematic picture of a agglomerative clustering tree. The leaf nodes correspond to "clusters" consisting of only one data point. The root node represents one single cluster containing all the points.

Meta-Clustering Meta-clustering is a new technique which was introduced by Caruana et al [21]. 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 to organize the clusterings of the base-level data into groups of similar clusterings. Given this Meta-Clustering, the 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 dimensionality of a dataset [22], [23]. The fractal dimension of a finite set of data points can be estimated by calculating by dividing the underlying D-dimensional space of the data into (hyper-cubic) grid cells of side r. Let$N(r)$ be the the number of cells that contain one or more data point. Then the fractal dimension can measured by

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

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

3.2 Proposed Work

Study of Meta-Clustering on data-set In the case of the protein data set , there are numerous ways of grouping structures together and 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 on which we can perform rule learning. Several approaches to analyzing the structure of the data have been tried, among them agglomerative clustering, multidimensional scaling and univariate and multivariate statistical analysis. The framework of Meta-Clustering as described above seems promising in this respect. 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. An interesting question is how the fractal dimension varies with the different metrics used for the meta-clustering and also the relation to an earlier principal component analyis.

Finding useful clusters Once we have a better understanding of the clustering, the next step will be to find clusterings that are particularly 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 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. 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. It will also be interesting to examine how the success of the RL compares with the distance function used for the meta clustering. The last items will part of the main project.

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 prior studies mentioned above, we would expect that this approach should be able to find some interesting structural motifs and possibly. The success of this research can be easily evaluated by looking at the rules found by RL. The rules found can be examined and evaluated by a domain expert, according to whether or not sequence rules are in agreement with the basic properties of the encoded amino acids.

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 masters 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 helix crossings should allow us 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 which are in line which expert predictions.

The current project will focus on separating examples of different types helix crossings into groups which will exhibit the kind of rules we are looking for. The framework of meta-clustering and agglomerative clustering should allow to better understand 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. If 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.

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

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

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

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

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

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

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

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

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

[13] 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.

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

[15] 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.

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

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

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

[19] 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.

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

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

[22] 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.

[23] 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.