Mining Protein Contact Maps

Mohammed J. Zaki*
Computer Science Department
Rensselaer Polytechnic Institute
110 8th Street, Troy, NY 12180-3590
Email: zaki@cs.rpi.edu

Abstract

We discuss some novel mining tasks for protein contact maps (two dimensional representations of the three dimensional structure of proteins). We show that using contact maps and a hybrid mining approach, we can construct "contact rules" to predict the structure of an unknown protein. Furthermore, we mine a model that discriminates physical from non-physical maps using frequent dense patterns and heuristic rules of physicality.

1 Introduction

Bioinformatics is the science of storing, extracting, organizing, analyzing, interpreting, and utilizing information from biological sequences and molecules. It has been mainly fueled by advances in DNA sequencing and genome mapping techniques. The Human Genome Project has resulted in rapidly growing databases of genetic sequences, while the Structural Genomics Initiative is doing the same for the protein structure database. New techniques are needed to analyze, manage and discover sequence, structure and functional patterns or models from these large sequence and structural databases. High performance data analysis algorithms are also becoming central to this task.

Bioinformatics is an emerging field, undergoing rapid and exciting growth. Knowledge discovery and data mining (KDD) techniques will play an increasingly important role in the analysis and discovery of sequence, structure and functional patterns or models from large sequence databases. One of the grand challenges of bioinformatics still remains, namely the protein folding problem.

Proteins fold spontaneously and reproducibly (on a time scale of milliseconds) into complex three-dimensional globules when placed in an aqueous solution, and, the sequence of amino acids making up a protein appears to completely determine its three dimensional structure. Given a protein amino acid sequence (*linear structure*), determining its three dimensional folded shape, (*tertiary structure*), is referred to as the *Structure Prediction Problem*; it is widely acknowledged as an open problem, and a lot of research in the past has focused on it.

Traditional approaches to protein structure prediction have focused on detection of evolutionary homology [2], fold recognition [3, 12], and where those fail, ab initio simulations [13] that

*This work was supported in part by NSF CAREER Award IIS-0092978, DOE Early Career Award DE-FG02-02ER25538, and RPI Exploratory Seed Grant

generally perform a conformational search for the lowest energy state [11]. However, the conformational search space is huge, and, if nature approached the problem using a complete search, a protein would take longer to fold than the age of the universe, while proteins are observed to fold in milliseconds. Thus, a structured folding pathway (time ordered sequence of folding events) must play an important role in this conformational search. Strong experimental evidence for pathway-based models of protein folding has emerged over the years. These pathway models indicate that certain events always occur early in the folding process and certain others always occur later.

It appears that the traditional approaches, while having provided considerable insight into the chemistry and biology of folding have largely hit a brick wall when it comes to structure prediction, hence, a novel approach is called for, for example, using data mining. Mining from examples is a data driven approach that is generally useful when physical models are intractable or unknown, however, data representing the process is available. Thus, our problem appears to be ideally suited to the application of mining – physical models of folding are either intractable or not well understood, and data in the form of a protein data base exists.

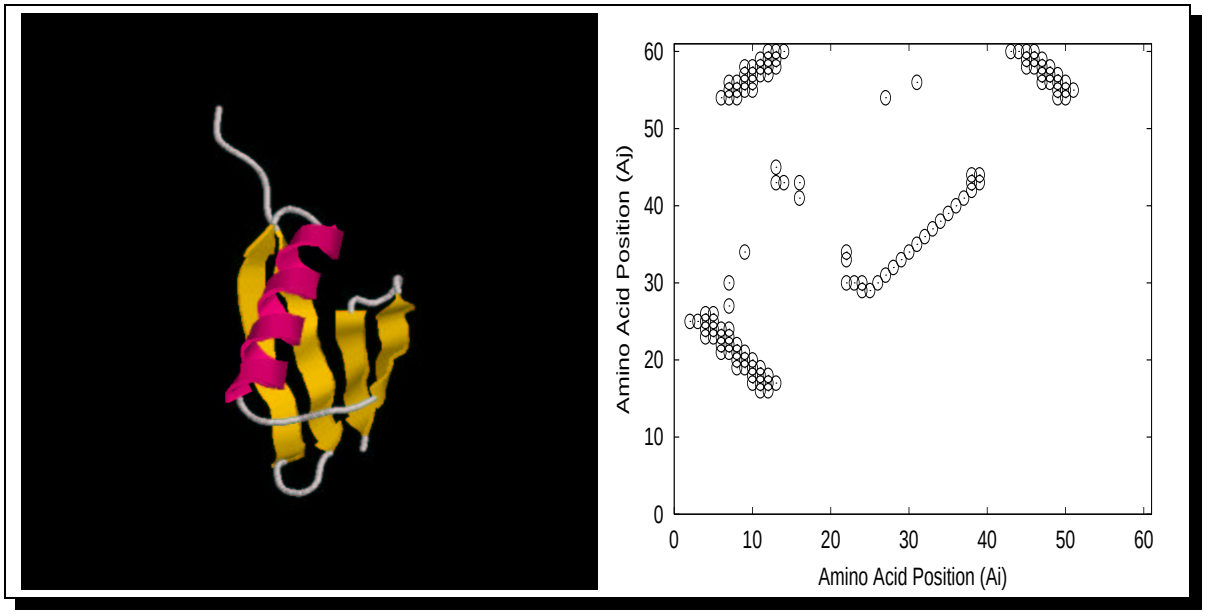

Figure 1: **Left:** 3D structure for protein G (PDB file 2igd, Sequence Length 61). **Right:** Contact Map for protein G – Circles indicate residue contacts, while non-contacts are represented by empty space. Only the upper triangle of the matrix is shown, since the contact map is symmetric. Clusters of circles indicate certain secondary structures, for example, the cluster along the main diagonal is an α-helix, and the clusters parallel and anti-parallel to the diagonal are parallel and anti-parallel β-sheets, respectively.

2 Modeling Protein Folding

It is well known that proteins fold spontaneously and reproducibly to a unique 3D structure in aqueous solution. Despite significant advances in recent years, the goal of predicting the three dimensional structure of a protein from its one-dimensional sequence of amino acids, without the

aid of evolutionary information, remains one of greatest and most elusive challenges in bioinformatics. The current state of the art in structure prediction provides insights that guide further experimentation, but falls far short of replacing those experiments.

Today we are witnessing a paradigm shift in predicting protein structure from its known amino acid sequence $(a_1, a_2, \cdots, a_n)$. The traditional or Ab initio folding method employed first principles to derive the 3D structure of proteins. However, even though considerable progress has been made in understanding the chemistry and biology of folding, the success of ab initio folding has been quite limited.

Instead of simulation studies, an alternative approach is to employ learning from examples using a database of known protein structures. For example, the Protein Data Bank (PDB) records the 3D coordinates of the atoms of thousands of protein structures. Most of these proteins cluster into around 700 fold-families based on their similarity. It is conjectured that there will be on the order of 1000 fold-families for the natural proteins [16]. The PDB thus offers a new paradigm to protein structure prediction by employing data mining methods like clustering, classification, association rules, hidden Markov models, etc.

The ability to predict protein structure from the amino acid sequence will do no less than revolutionize molecular biology. All genes will be interpretable as three-dimensional, not one-dimensional, objects. The task of assigning a predicted function to each of these objects (arguably a simpler problem than protein folding) would then be underway. In the end, combined with proteomics data (i.e. expression arrays), we would have a flexible model for the whole cell, potentially capable of predicting emergent properties of molecular systems, such as signal transduction pathways, cell differentiation, and the immune response.

Protein Contact Maps A *contact map* is a particularly useful two dimensional representation of a protein's tertiary structure. An example is shown in Figure 1. Two residues (or amino acids) a_i and a_j in a protein are in *contact* if the 3D distance is less than some threshold value t (we used $t = 7\mathring{A}$). Using this definition, every pair of amino acids is either in contact or not. Thus, for a protein with N residues, this information can be stored in an NxN binary symmetric matrix C, called the contact map. Each element, C_{ij}, of the contact map is called a contact, and is 1 if residues a_i and a_j are in contact, and 0 otherwise. The contact map provides a host of useful information. For example, clusters of contacts represent certain secondary structures: α-Helices appear as bands along the main diagonal since they involve contacts between one amino acid and its four successors; β-Sheets appear as thick bands parallel or anti-parallel to the main diagonal. Tertiary structure may also be obtained by reverse projecting into 3D space using the MAP algorithm [15].

Sources of Data Since we are using a data driven learning approach, we take a moment to discuss our data sources (see Figure 2). The Protein Data Bank (PDB) records the 3D coordinates of the atoms of thousands of protein structures. The set of all known, globular proteins cluster into around 700 families based on their sequence similarity (PDBselect [6]). It is conjectured that there will be on the order of 1000-2000 fold-families for the natural proteins. Thus from the PDB we can extract a set of proteins along with their known contact maps. These contact maps form the "rule learning data", which will be used to mine for association rules. Part of this data set will be used for learning meaningful patterns, and a part will be set aside for validation. In addition, using HMMSTR and ROSETTA (both to be described later) one can generate additional protein-like structures for which the contact maps can be obtained. These contact maps, in addition to the

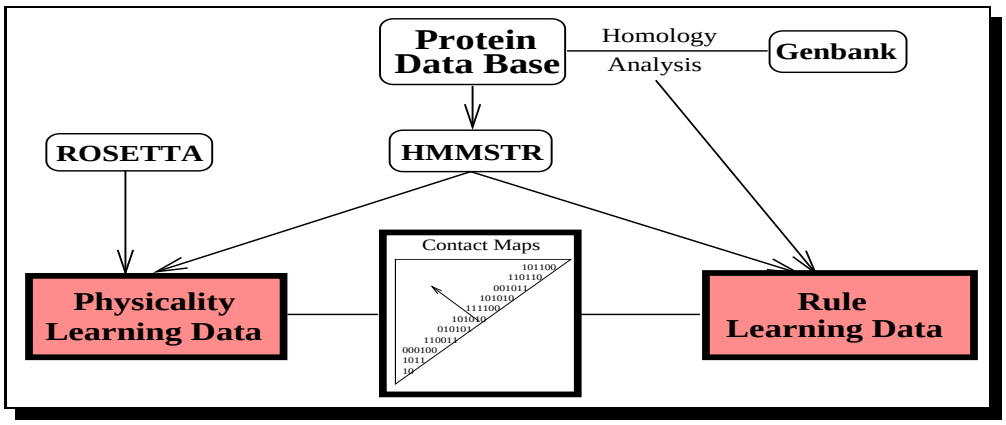

Figure 2: Sources of Data for Mining

rule learning data forms the "physicality learning data", which will be used to learn an model that discriminates physical from non-physical maps.

Below we discuss how data mining can be used to extract valuable information from contact maps. More specifically we focus on two main tasks: 1) Given a database of protein sequences and their 3D structure in the form of contact maps, build a model to predict if pairs of amino acids are likely to be in contact or not. 2) Discover common (non-local) contact patterns or "features" that characterize physical "protein-like" contact maps.

The protein folding problem will be solved gradually, by many investigators who share their results at the bi-annual CASP (Critical Assessment of protein Structure Prediction) meeting [8], which offers a world-wide blind prediction challenge. Here, we will investigate how mining can uncover interesting knowledge from contact maps.

3 Mining Contacts using HMMSTR for Local Structure

We used a generalized hidden Markov model called HMMSTR based on the I-sites library [4] to model statistical interactions between adjacent motifs on the chain, and were thus able to model the local propagation of structure. I-sites (or Initiation-Sites) are local sequence motifs that tend to fold the same way across protein families independent of the context. A rule-based method for predicting tertiary contacts in proteins, using HMMSTR as a preprocessor has already been developed [18] and can be extended to sequentially output probabilities for subsets of contacts.

Super-local Contact Potentials Sequences from the database of all known proteins were pre-processed using HMMSTR. The associated structures were converted to contact maps. The whole dataset was then mined to find common association rules for tertiary contacts. The rules were tested on a subset not used in the data mining.

The database of known proteins was divided into a training set and a test set. Each protein sequence was submitted to Psi-Blast to generate a sequence family, from which a sequence profile was summed, and backbone angles were discretized [4]. For the training set, the amino acid profile and the backbone angle regions were the input data for the forward/backward calculation [10], which produced a "gamma" matrix of position-dependent HMMSTR Markov state probabilities. For the test set, only the amino acid profile was used to generate the gamma matrix. A contact

map was calculated for each member of the training set using a alpha-carbon distance cutoff of 7Å. From these data, a database of "item sets" was constructed. One entry corresponds to an ij residue pair, where $|i - j| > 4$, and consists of the amino acid pair and two sets of Markov state identifiers. Markov state identifiers were included as "items" associated with the ij contact if the position-dependent probability of the state was greater than twice the *a priori* probability of that state. Each entry had a label "1" meaning i and j were in contact in the structure, or "0" if they were not. All items are discrete symbols.

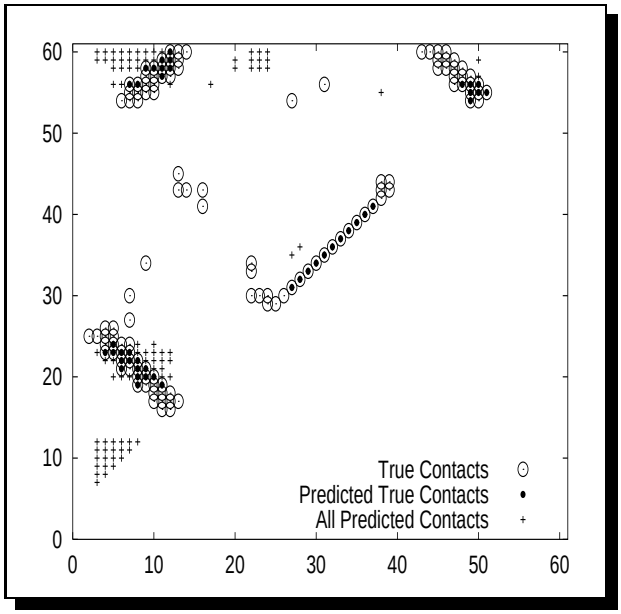

Figure 3: Predicted Contact Map (PDB protein 2igd). We were able to predict parts of the major structures.

Association rule data mining [1, 17] was applied to the database of item sets to extract rules which were predictive of contacts. A rule has the form: "item1" + "item2" + ... $\Rightarrow C$, where C is 1 or 0. Items may be amino acids, predicted or observed secondary structure symbols, predicted or observed Markov state identifiers, or sequence separation ranges. Prediction of contacts in the test set was carried out by comparing the rule support for contact prediction with the rule support for non-contact prediction. The ratio of these values for all ij pairs in a given protein was sorted, and the top N pairs were predicted to be in contact, where N is the expected number of contacts, which depends on the sequence length. Figure 3 shows an example contact map prediction. Previous work on contact prediction has employed Neural Networks [5], and statistical techniques based on correlated mutations [9, 14]. Recent work by Vendruscolo [15] has also shown that it is possible to recover the 3D structure from even corrupted contact maps. Our recent results [18] show that our model obtains around 20% accuracy and coverage over the set of all proteins; the model is also 5.2 times better than a random predictor. We can significantly enhance coverage to over 40% if we sacrifice accuracy (13%). For short proteins (length< 100) we get 30% accuracy and coverage (4.5 times better than random); if we lower accuracy to 26% we can get coverage up to 63%. While these results are better than (or equal to) those reported previously, we have still a long way to go before the goal of protein structure prediction is fully realized. Generating three-dimensional structure from the predicted contact maps is now a subject of our investigations.

4 Predicting Physical Contact Maps

The output of the method described in the previous section is a set of rules $R^1, \ldots, R^K$. We can convert these rules into energies for each contact ij using a log linear transformation $E_{ij} = -\log(c_1 R^1_{ij} + \ldots + c_K R^K_{ij})$, which is just a number associated with each unassigned contact, related to energy of that contact, were it to be present. This stage is sufficiently fast, but it ignores the geometric constraints on the system.

Proteins are self-avoiding, globular chains. A contact map, if it truly represents a self-avoiding and compact chain, can be readily translated back to the three-dimensional structure from which it came. But, in general, a symmetric matrix of ones and zeros does not have this property. Only a small subset of all such matrices correspond to three-dimensional self-avoiding chains. The task is to output a contact map, given the matrix of probabilities above, that both satisfies the geometrical constraints and is likely to represent a low energy structure and protein folding pathway. Interactions between different subsequences of a protein are constrained by a variety of factors. The interactions may be initiated at several short peptides (initiation sites) and propagate into higher-order intra or inter-molecular interactions. The properties of such interactions depend on (1) the amino acid sequence corresponding to the interactions, (2) the physical geometry of all interacting groups in three dimensions, and (3) the immediate contexts (linear, and secondary components for tertiary structural motifs) within which such interactions occur.

4.1 Physicality Model: Incorporating Geometric Constraints

This task is to learn a method for predicting whether a given contact map is physical or not. In order to apply mining techniques, to learn what a physical map is, one needs a data set, a set of physical and non-physical contact maps along with which maps are physical. The PDB database will serve as the source for physically possible contact maps (i.e., the contact maps corresponding to real proteins whose structure is known). In order to generate nonphysical/unprotein-like contact maps, one can randomly perturb the physical maps in ways that are known to be nonphysical in addition to generating random contact maps and testing their physicality using a program called ROSETTA. Additionally since random maps are extremely likely to be nonphysical, we could also generate maps at random, labeling them to be non-physical.

By mining from this data set, we can learn a fast model for determining which two-dimensional patterns of contacts correspond to three-dimensional self-avoiding protein chains. By learning from more data, the learned model can be made more accurate. In particular, we mine frequent dense patterns or structural motifs in contact maps. These motifs represent the typical non-local structures that appear in physical protein-like contact maps, and which can be used to improve the quality of contact map prediction by eliminating impossible contacts (those that never occur in real proteins). Briefly, there are three major stages for the approach: (1) Data generation, which involves creating of a large set of physical and non-physical contact maps, (2) Mining, which involves computation of all the frequent dense patterns, and (3) Pruning mined frequent patterns and integration of these patterns with biological data.

4.1.1 Mining Frequent Dense Patterns

To enumerate all the frequent 2D dense patterns we scanned a database of 12524 contact maps with a 2D sliding window of a user specified size [7]. For all structures, any sub-matrix under the window that had a minimum "density" (the number of '1's or contacts) was captured. For a $N \times N$

contact map, using a 2D $W \times W$ window, there are $(N-W) \times (N-W)/2$ possible submatrices. We have to tabulate those which are dense, using different window sizes. We chose window sizes from 5 to 10, to capture denser contacts close to the diagonal (i.e., short-range interactions), as well as the sparser contacts far from the diagonal (i.e., long-range interactions).

As we slide the $W \times W$ window, the sub-matrix under the window will be added to a dense pattern list if its density exceeds the *min_d* threshold. However, we are interested in those dense patterns that are frequent, i.e., when adding a new pattern to the list of dense patterns we need to check if it already exists in the list. If yes, we increase the frequency of the pattern by one, and if not, we add it to the list initialized with a count of one.

After obtaining mined patterns that are frequent and are relatively dense, we pruned them using a number of heuristics in order to extract biologically meaningful structural motifs. All the potential structure motifs fall into two categories: that of secondary structures which primarily consist of alpha helices or beta sheets (parallel or anti-parallel), and that of tertiary structures which involve interactions between secondary structure components. For example, alpha helices, beta sheets and beta turn regions can have multiple contacts between them, such that components that are farther away in linear sequence could be brought together to form functional groups. These tertiary structures are particularly important for biological processes such as high-specificity binding of ligands and receptors.

By applying above method, two major groups of patterns were isolated, as described in the next section: one group being major secondary structure components such as alpha and beta sheets, another group being tertiary structural motifs involving two secondary structural components.

0 0 0 0 0 0 0 0	0 1 1 1 0 0 0 0	1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0	1 1 1 0 0 0 0 0	0 1 0 0 0 0 0 0
0 0 0 0 0 0 0 0	1 1 0 0 0 0 0 0	0 0 1 0 0 0 0 0
0 0 0 0 0 0 0 0	1 0 0 0 0 0 0 0	0 0 0 1 0 0 0 0
0 0 0 0 0 0 0 1	0 0 0 0 0 0 0 0	0 0 0 0 1 0 0 0
0 0 0 0 0 0 1 1	0 0 0 0 0 0 0 0	0 0 0 0 0 1 0 0
0 0 0 0 0 1 1 1	0 0 0 0 0 0 0 0	0 0 0 0 0 0 1 0
0 0 0 0 1 1 1 0	0 0 0 0 0 0 0 0	0 0 0 0 0 0 0 1
sup = 2.0%	sup = 2.2%	sup=1.9%
anti-parallel beta	anti-parallel beta	parallel beta

Table 1: Frequent Dense Submatrices

We discovered frequent submatrix patterns whose supports reach 1 to 2% when using a sliding window of size 8 and a minimum density of 0.125. These patterns turned out to correspond to beta sheets secondary structures (parallel and antiparallel '1's in the binary contact map). Examples of the most frequent dense patterns are given in Table 1. While these patterns are not novel, they serve to validate our approach, i.e., the most frequent structures we expected were in fact discovered by the mining process.

The second class of patterns involve interactions between different secondary structures. The kinds of the interactions revealed with the frequent dense patterns differ in terms of the involved secondary structure components, multiplicity of interacting atoms, and the contexts (linear and secondary) surrounding such interactions. Roughly, based on the type of secondary components that are involved, we observed that the mined tertiary structure motifs can be categorized into the following classes:

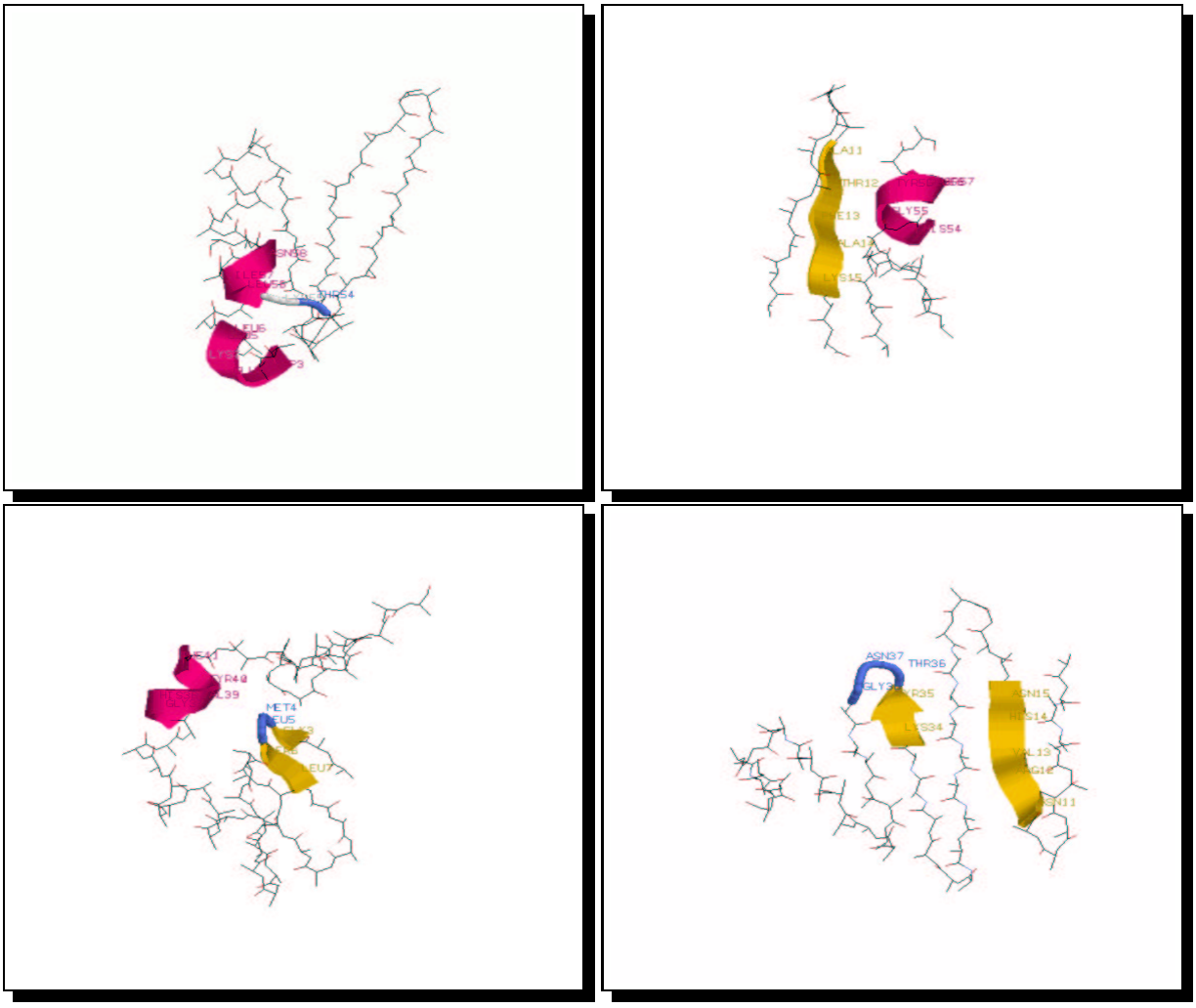

Figure 4: Frequent Patterns between Secondary Structures: 1) Alpha Helix - Alpha Helix 2) Alpha Helix - Beta Sheet, 3) Alpha Helix - Beta Turn, 4) Beta Sheet - Beta Turn

1. Alpha helix a_1 :: Alpha helix a_2
2. Alpha helix a :: Beta sheet b
3. Alpha helix a :: Beta turn bt
4. Beta sheet b :: Beta turn bt

Figure 4 shows an example of each of these four types of interactions. The above classes can be further divided into sub-classes according to the number of contacts involved in each component, multiplicity of interacting atoms (one to one, one to many, or many to many), sequence specificities, and the linear/secondary structural contexts of the interaction. More details on mining frequent dense patterns appear in [7]. We are currently creating a library of all possible non-local interactions in "real" contact maps.

5 Future Work

As part of future we hope to compile a library of such non-local interactions between different secondary structures. This library would be analogous to the I-sites library, but while the I-sites library records the common motifs for short contiguous segments (3-19 residues), the new library will record interactions between non-contiguous segments.

Mining Rules for "Physicality" Simple geometric considerations may be encoded into heuristics that recognize physically possible and protein-like patterns within contact maps, C. For example, we may consider the following to be rules that are never broken in true protein structures:
If $C(i,j) = 1$ and $C(i+2, j+2) = 1$, then $C(i, j+2) = 0$, and $C(i+2, j) = 0$.
If $C(i+2, j) = 1$ and $C(i, j+2) = 1$), then $C(i,j) = 0$, and $C(i+2, j+2) = 0$.
These rules encode the observation that a beta sheet (contacts in a diagonal row) is either parallel or anti-parallel, but not both.

Another example may be drawn from contacts with alpha helices: If $C(i, i+4) = 1$ and $C(i,j) = 1$ and $C(i+4, j) = 1$, then $C(i+2, j) = 0$. This follows from the fact that $i+2$ lies on the opposite side of the helix from i and $i+4$, and therefore cannot share contacts with non-local residue j. Local structure may be used in the definition of the heuristics. For example, if an unbroken set of $C(i, i+4) = 1$ exists, the local structure is a helix, and therefore, for all $|j-i| > 4$ in that segment, $C(i,j) = 0$. The question is whether one can mine these rules automatically.

Consider the contact map for the parallel beta sheet shown in Table 1. We can discover "positional" rules, i.e., the heuristic geometric rules by considering an appropriate neighborhood around each contact $C(i,j)$ and noting down the relative coordinates of the other contacts and non-contacts in the neighborhood, conditional on the local structure type(s). Consider a lower 1-layer (denoted LL1) neighborhood for a given point, $C(i,j)$. This includes all the coordinates within $i+1$ and $j+1$, i.e. each point has 3 other points in its LL1 neighborhood, namely $C(i, j+1)$, $C(i+1, j)$ and $C(i+1, j+1)$. If we repeat this process for each point we obtain a database which can be mined for frequent combinations.

For instance looking at the parallel beta sheet submatrix in Table 1 we would get the set $C(i,j) = 1$, $C(i+1, j) = 0$, $C(i, j+1) = 0$, $C(i+1, j+1) = 1$ for each contact. If one were to do this for many other proteins one would find the pattern "If $(Ci, j) = 1$ and $C(i+1, j+1) = 1$, then $C(i, j+1) = 0$ and $C(i+1, j) = 0$," among several others. This is the same rule deduced by hand above.

Other patterns can be found by defining an appropriate neighborhood, which can be t layers thick (where t is the maximum coordinate difference between points), and can encompass all points within t or some subset of that region. From each of these we can construct examples which can be mined for frequent patterns to obtain heuristic contact rules. We can also incorporate sequence information to mine more complicated patterns. We propose to develop techniques to mine such heuristic rules of contact automatically.

References

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Inkeri Verkamo. Fast discovery of association rules. In U. Fayyad and et al, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. AAAI Press, Menlo Park, CA, 1996.

[2] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17), 3389-402, 1997.

[3] S. Bryant. Evaluation of threading specificity and accuracy. *Proteins*, 26(2), 172-85, 1996.

[4] C. Bystroff, V. Thorsson, and D. Baker. HMMSTR: A hidden markov model for local sequence-structure correlations in proteins. *Journal of Molecular Biology*, (to appear), 2000.

[5] P. Fariselli and R. Casadio. A neural network based predictor of residue contacts in proteins. *Protein Engineering*, 12(1), 15-21, 1999.

[6] U. Hobohm and C. Sander. Enlarged representative set of protein structures. *Protein Science*, 3(3), 522-524, 1994.

[7] J. Hu, X. Shen, Y. Shao, C. Bystroff, and M.J. Zaki. Mining protein contact maps. In *2nd BIOKDD Workshop on Data Mining in Bioinformatics*, July 2002.

[8] J. Moult, J. Pedersen, R. Judson, and K. Fidelis. A large-scale experiment to assess protein structure prediction methods. *Proteins*, 23(3), ii-v, 1995.

[9] O. Olmea and A. Valencia. Improving contact predictions by the combination of correlated mutations and other sources of sequence information. *Folding & Design*, 2, S25-S32, June 1997.

[10] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257-86, 1989.

[11] K.T. Simons, C. Kooperberg, E. Huang, and D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. *Journal of Molecular Biology*, 268(1), 209-25, 1997.

[12] M. Sippl. Helmholtz free energy of peptide hydrogen bonds in proteins. *J. Mol. Biology*, 260(5), 644-8, 1996.

[13] J. Skolnick, A. Kolinski, and A. Ortiz. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. *Proteins*, 38(1), 3-16, 2000.

[14] D. Thomas, G. Casari, and C. Sander. The prediction of protein contacts from multiple sequence aligments. *Protein Engineering*, 9(11):941-48, 1996.

[15] M. Vendruscolo, E. Kussell, and E. Domany. Recovery of protein structure from contact maps. *Folding & Design*, 2(5), 295-306, September 1997.

[16] Y. I. Wolf, N. V. Grishin, and E. V. Koonin. Estimating the number of protein folds and families from complete genome data. *Journal of Molecular Biology*, 299(4), 897-905, 2000.

[17] M. J. Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372-390, May-June 2000.

[18] M. J. Zaki, S. Jin, and C. Bystroff. Mining residue contacts in proteins using local structure predictions. In *IEEE International Symposium on Bioinformatics and Biomedical Engineering*, November 2000.