WAAMD 2007
III Workshop em Algoritmos e Aplicações de Mineração de Dados

Mining structural signatures of proteins

R.C. Melo[12], J.S. Gomide[1], P.S.L. Dias[1],W. Meira Jr.[1], M.M. Santoro[2]

[1]Departamento de Ciência da Computação - Universidade Federal de Minas Gerais

[2]Departamento de Bioquímica e Imunologia - Universidade Federal de Minas Gerais

{raquelcm,janaina,samer,meira}@dcc.ufmg.br, santoro@icb.ufmg.br

Abstract. *Proteins are the most important macromolecules in living systems. It is well known that their function is totally dependent on their structure. However, identical structures can be formed by very dissimilar amino acid sequences and little is known about how so different sequences fold into identical structures and functions. In this work, we propose a clustering approach to obtain patterns of amino acid chemical interactions in protein families that compose a structural signature for each family.*

1. Introduction

Bioinformatics is an emerging field undergoing rapid growth. This has been fueled by advances in DNA sequencing techniques. The Human Genome Project resulted in an exponentially growth in the database of genetic sequences and Structural Genomics Initiative is doing the same for Protein Data Bank (PDB) [Berman et al. 2000]. One of the most active research areas is protein topology analysis. Proteins are the most versatile macromolecules in living systems serving crucial functions in all biological processes, such as catalysts, transporters, and mechanical support. They are composed by a sequence of amino acids which is called *primary structure*. Different regions of the sequence form regular *secondary structures* such as α-helices or $beta$-sheets. The *tertiary structure*, which is the 3D structure of the protein, is formed by packing such structural elements into compact globular units called domains. The functional properties of proteins depend upon their 3D structures that arises because a particular chain of amino acids folds to generate domains with specific 3D structures. It is known that the chain completely determines the structure of a protein. However, there may be several proteins with the same structure (or family), and the same function, but with very different sequences and many variations in secondary structure sizes and positions. Hence, the study of protein topology is very important because the topology determines protein function.

PDB has on its archives approximately 45,000 proteins and this number has been increasing year after year. Even thought it is a huge data set, significant knowledge still needs to be extracted from it. As a modest step towards this ambitious long-term goal, in this work we use a data mining approach to analyze similarity of proteins and to extract conserved information on dissimilar sequences of proteins of the same family. These patterns are part of what we call *structural signature* of a protein family. A structural signature is a set of characteristics that can univocally identify a family, and thus structure and function proteins. It could additionally give clues about the nature of interactions that a protein can establish with ligands and other proteins. In the present work, we use a database that contains information about chemical interactions within proteins, represented by contact maps, exploiting the spatial co-location of interactions as evidence towards defining a protein signature.

Contact maps encode long-range interactions within proteins in a compact way and have been used in the literature as two-dimensional representations of proteins' 3D topology. We present some examples of contact maps in Figure 1. In fact, a chain folds into a 3D structure because of chemical interactions between its amino acid residues. These interactions are indispensable for the action of proteins, being of interest to study the similarity of proteins based on their chemical interactions patterns. The 3 most important kinds of interactions are *hydrophobic*, *hydrogen bonds* and *electrostatic*. The *hydrophobic contacts* arise because of the molecules water aversion. This effect makes apolar parts of chain closer in the 3D structure. The *hydrogen bonds* are short-distance interactions between residues that share a common hydrogen atom. They are essential in the stabilization of secondary strutures. The *electrostatic interactions* are the attraction or repulsion of charged amino acids, which occur far apart in structure. They are not being considered in this work because the occurence of charge clusters are very rare in proteins. In [Agarwal et al. 2007], the authors prove that contact map overlap problem is NP-Hard and describe an algorithm which solves the problem in polynomial time for some particular cases. The advantage of our proposal is that we do not overlap contact maps but signatures which are much more concise representations of them.

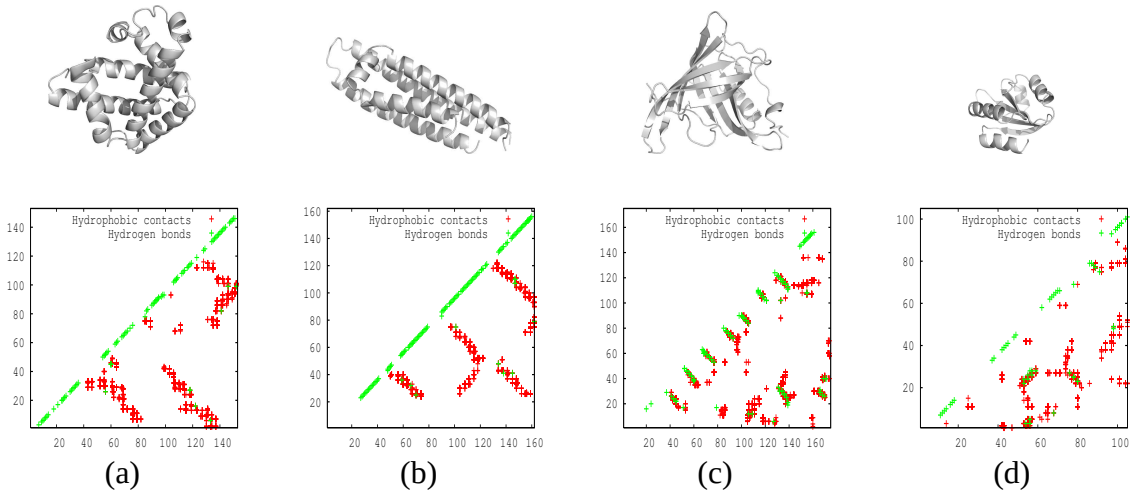
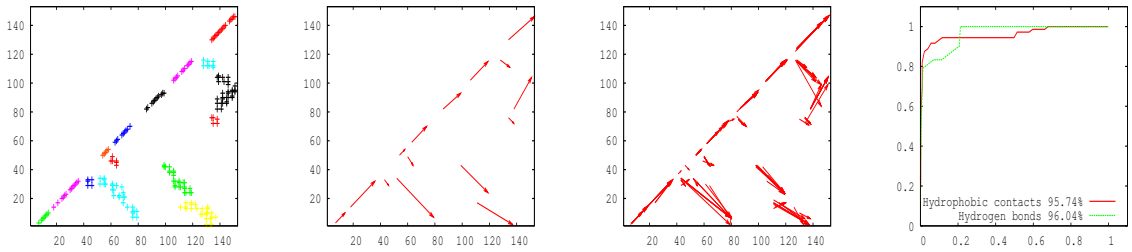


Figure 1. Protein families and their contact maps. In (a), human being Myoglobin, an oxigen carrier in the muscles; in (b) an Apolipoprotein from human being, a lipid transporter; in (c) a Retinol Binding Protein (R.B.P.) from pig plasma; in (d) a human being Thioredoxin, an electron transporter.

2. Mining structural signatures

In this section we present both our mining strategy and its application to samples of Myoglobins, Apolipoproteins (helical proteins similar to Myoglobins), Plastocyanins, Retinol Binding Proteins (R.B.P.s) and Thioredoxins from varied animal species selected from PDB according to SCOP manually curated classification [Murzin et al. 1995]. The initial step was to compute contacts given a set of atomic coordinates from a PDB file which is partially based on [Sobolev et al. 1999] and [Mancini et al. 2004]. We analyze each atom of the 20 most commonly found residues as in [Sobolev et al. 1999] and attribute for each one a class: acceptor and donnor (for hydrogen bond determination) and hydrophobic (for hydrophobic interactions) and use the same distance thresholds of [Mancini et al. 2004].



(a) (b) (c) (d)

Figure 2. We present the clusters defined by DBSCAN of hydrophobic interactions and and hydrogen bonds for the Myoglobin of Figure 1 in (a). In (b), we have the the vectors that represent these clusters, i.e., the structural signature of the protein. In (c), we have Myoglobin family structural signature and in (d), precision of the classification of Myoglobins according to protein structural signature.

Contact maps of protein families present conserved clusters, which we detect using a density-based clustering algorithm, DBSCAN [Ester et al. 1996]. This algorithm is able to handle an important characteristic of the clusters of contacts: they present a linear shape and are always parallel or anti-parallel to the map diagonal. The parallel clusters indicate that, numbering the sequence from the beginning to the end, we have two increasing parts of the chain close w.r.t. the structure establishing chemical interactions, while the anti-parallel means the reverse. The parameters of the algorithm were determined usign the heuristic suggested by the authors. In Figure 2.(a), we can see some examples of clusters identified for the Myoglobin presented in Figure 1. Notice that some of the clusters are not linear, but are detected by DBSCAN. As an example, we can see a ”L”-shape hydropobic cluster in Figure 2.(a). Such clusters represent that a secondary structure is in contact with two or more secondary structures.

The next step of our strategy is to determine lines that characterize the clusters. We use the Hought Transform [Illingworth and Kittler 1988] to detect the single or multiple lines that characterize a cluster. In Figure 2.(b) we can see the representative lines for the clusters of Figure 2.(a). The representation of cluster using lines makes easier to recognize relevant patterns and to detect the lines (and thus clusters) that are conserved among several proteins. In Figure 2.(c), we plot the representative lines for the clusters of 9 Myoglobin samples. They are from human being, whale, horse, pig, elephant, turtle, tuna, seal, and mollusc. The diagonal clusters (hydrogen bond clusters) mark the helix occurence and are very well conserved. Proteins of Myoglobin family are composed by, generally, 6 to 8 helices. They are named from A to H. Helices C and D are very small and in many samples of family they do not occur. There are 5 hydrophobic clusters conserved in all Myoglobin samples. They are probably an important component of the structural signature we are looking for. They represent contacts between parts of the chain in helices AB with E, AB with G, AB with H, G with H and F with H. According to [Hughson et al. 1997], the Myoglobin folding intermediate contains the A, G, and H, i.e., these helices are the first to fold and their contacts are important to the stabiization of the molecule. Particularly, the A and G helices interact to stabilize each other, as shown by the effect of mutations in the G helix on the unfolding of the A helix. Notice that helices A, G and H are present in all the conserved hydrophobic clusters of the proposed

structural signature.

After determining the structural signatures for the various families of proteins, we used the signature vectors to classify proteins, i.e., determine their families. As we wanted to determine the matches of vectors which minimize the cost of mutate the map of a protein into the map of another, we modeled the problem of comparing two sets of vectors in a 2D space (each set is a protein structure signature) as a Transportation Problem. The dissimilarity of the maps is measured by the minimum cost of moving all the origin and destination points of the vectors from a signature to the vectors from another signature. Comparing the same 9 Myoglobins of different animal species against 62 proteins of different classes (Apolipoproteins, Plastocyanins, R.B.P.s and Thioredoxins), i.e., classifying a protein as Myoglobin or non-Myoglobin, we achieved a 95% precision (measured as the area under the ROC curves) using hydrophobic contacts and hydrogen bonds, as we can see in Figure 2.(d). The classification methodology consist of comparing all against all proteins, ranking them according to the dissimilarity index given by the cost of transporting vectors of one protein to the other and selecting the k most similar as Myoglobin and the others as non-Myoglobin. By cutting these ranks in different ks and computeing the false positive rate and true positive rate, we obtain the curves of Figure 2.(d).

3. Conclusions

In this paper, we presented a novel methodology for understanding contact maps and through data mining tecnhiques extract patterns that we call protein structural signatures. This is a new concept in molecular biology and can be helpful in the solution of the widely studied and still open *Protein Folding Problem.* We also show that the structural signatures can be used to classify proteins according to their structure and function.

References

Agarwal, P. et al. (2007). Fast molecular shape matching using contact maps. *J. Comput. Biol.*, 14(2):131–143.

Bourne, P. et al. (2000). The protein data bank. *Nucleic Acid Research*, 28:235–242.

Xu, X. et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 226–231.

Baldwin, R. et al. (1997). Cooperativity of folding of the apomyoglobin ph 4 intermediate studied by glycine and proline mutations. *Nature Structural Biology*, 4:925–930.

Kittler, J. et al. (1988). A survey of the hough transform. *Computer Vision, Graphics, and Image Processing*, pages 87–116.

Neshich, G. et al. (2004). Sting contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. *Bioinformatics*, 20(13):2145–2147.

Chothia, C. et al. (1995). Scop: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540.

Sobolev, V. et al. (1999). Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15:327–332.