Mining Structural Signatures in Proteins using Intrachain Interactions

R.C. Melo[12], J.S. Gomide[1], W. Meira Jr.[1], J.C.D. Lopes[3], G. Neshich[4] and 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
[3]Departamento de Química – Universidade Federal de Minas Gerais
[4]Grupo de Bioinformática Estrutural - CNPTIA/EMBRAPA
{raquelcm, janaina, samer, meira}@dcc.ufmg.br, santoro@icb.ufmg.br

Abstract

Proteins are the most versatile macromolecules in living systems serving crucial functions 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 structures. Hence, the study of protein topology is very important because the topology determines protein function.

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. 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 types of interactions are *hydrophobic*, *hydrogen bonds* and *electrostatic. Electrostatic interactions* are are not being considered in this work because the occurence of charge clusters are very rare in proteins.

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 familywhich is a set of characteristics that can univocally identify the structure and thus the function of proteins. 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. The initial step was to compute contacts given a set of atomic coordinates from a PDB file which is partially based on [1] and [2]. Contact maps of protein families present conserved clusters, which we detect using a density-based clustering algorithm, DBSCAN [3]. 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 amino acids 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 next step of our strategy is to determine lines that characterize the clusters. We use the Hought Transform [4] to detect the single or multiple lines that characterize a cluster. The representation of cluster using lines makes easier to recognize relevant patterns and to detect the lines conserved among several proteins. Finally, We used the signature vectors to classify proteins, i.e., determine their families. We modeled the problem of comparing two sets of vectors in a 2D space (structural signatures) 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. Retrieving 9 Myoglobins of different animal species from a dataset with 62 proteins of different classes (Apolipoproteins, Plastocyanins, R.B.P.s and Thioredoxins), we achieved a 95% precision.

References

[1] V. Sobolev, A. Sorokine, J. Prilusky, E. Abola and M. Edelman. Automated analysis of interatomic contacts in proteins, Bioinformatics, 15:327-332, 1999.

[2] A. Mancini, R. Higa, A. Oliveira, F. Dominiquini, P. Kuser, M. Yamagishi, R. Togawa and G. Neshich. STING Contacts: a web-based application for identification and analysis of amino acid contacts within protein structures and accross protein interfaces, Bioinformatics, 20(13):2145-2147, 2004.

[3] M. Ester, H. Kriegel, J. Sander and X. Xu. 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, 226–231, 1996.

[4] J. Illingworth and J. Kittler. A survey of the hough transform. Computer Vision, Graphics, and Image Processing, 87–116, 1988.