**A hybrid method for identification of structural domains(2014)**

Hua et al[1]. used a hybrid method by combining two algorithms: Support Vector Machine(SVM) and K-means to identify the structural domains. Their algorithm takes into account, the density, length and dispersion of a domain in order to train SVM. The method works in two steps, the first step is where single and multi domain chains are classified by splitting a chain into two by using k-means and then computing the above mentioned properties of the two clusters and using them to train the SVM and then classify them.

Once all the single-domain chains are filtered out, in the second step similar process is repeated to classify two-domain proteins from the rest of the multi-domain proteins. Furthermore, three and four domain proteins are identified based on the modularity of a cluster. The modularity contains intra-cluster and inter-cluster cohesiveness. Optimizing modularity by the following formulas can identify the number of domains.

(1.1)

(1.2)

(1.3)

The ck, ni, center are the kth-cluster’s center, ith node, the center of the whole chain

respectively. The d(x,y) represents the Euclidean distance between x and y.

**References**

[1]: Hua, Y., Zhu, M., Wang, Y., Xie, Z. & Li, M. A hybrid method for identification of structural domains. Sci. Rep. 4, 7476; DOI:10.1038/srep07476 (2014).

**DomainParser**

DomainParser uses a top-down graph theoretical approach for domain decomposition with a rigorous post processing step. In the first step, the protein structure is modelled as a graph by considering residues as nodes and the edges between them is drawn based on their proximity. The edge capacity(the maximum flow of an edge) is assigned keeping in mind that inter-domain interactions should be less than intra-domain or in other words the network is to be divided into two portions such that the edge capacity across the portions is minimum. Also, further rules like each domain must not have many discontinuous segments, splitting of 𝛽-sheet should be avoided while decomposing a protein and a 𝛽-strand should not be cut are also incorporated. Thus, edge capacity turns out to be a function of (a) the number of atom-atom contacts between residues; (b) the number of backbone-backbone contacts; (c) interactions across a 𝛽-sheet and within a 𝛽-strand. These parameters are optimized during the training stage of the algorithm. Then, the Ford-Fulkerson algorithm[1] for finding minimum cut/maximum flow is applied to bipartition the graph. Briefly, the algorithm works by artificially adding a source and a sink node to the graph. A minimum cut is then calculated by finding the maximum flow from source to sink. A set of critical edges in the graph are identified by gradually increasing the flow of all edges and the ones with the least capacity are the ones which form the bottleneck. The removal of these edges stops the flow from source to sink. Thus, the nodes connected to the source form one of the possible domain, while the others connected to sink forms the other domain. This process is repeated a number of times by attaching the source and the sink to different nodes of the network. A set of minimum cuts is obtained and are evaluated by applying a lot post-processing steps. The algorithm is repeated recursively on the partitions obtained till the size of any one of the obtained domain falls below 80 or any of the obtained cuts stop meeting the criteria of the definition of a domain. The post-processing steps are used to evaluated the minimum cuts obtained and to refine the ones which are accepted. Some of the characteristic properties of domains which are used to evaluate the obtained partitions are (a) a domain should have at least 40 residues; (b) 𝛽-sheets are intact; (c) a domain should be compact; (d) the interface between two domains should be small; (e) the number of segments in a domain must be small.

**References:**

[1]: Ford, L. R.; Fulkerson, D. R. (1956). "Maximal flow through a network" (PDF). Canadian Journal of Mathematics. 8: 399–404. doi:10.4153/CJM-1956-045-5

**CATH**

CATH is a semi-automatic method to classify protein domains in a hierarchical manner. The four main levels of its classification are class(C), architecture(A), topology(T) and homology(H). Class is the simplest level and defines the secondary structure composition of the protein. Topology captures the sequential arrangement, while architecture reflects the shape and the orientation of the protein. Homology puts together those domains which share a common topology and a similar functionality. Briefly, the database is created by the following steps. First, only well resolved crystal structures(3.0 Å resolution or better) and NMR structures from the Protein Data Bank(PDB) are used. Next, pairwise comparisons between the sequences of all the proteins selected for CATH are performed using Needleman & Wunsch algorithm[1]. This is done as nearly three-quarters of the structures have identical sequences and are thus clubbed together. Completely identical proteins(100% sequence similarity and 100% overlap of structures) are grouped together into identical(I) families. Near-identical(N) families are subsequently created(>95% sequence similarity, at least 85% of larger protein equivalent to smaller).The S level grouping is created by clustering proteins having 35% or more sequence identity. The best resolved crystal structure of each family is used as the representative of that family. In the next step, domain boundaries for multi-domain proteins are assigned by using a consensus based approach between DOMAK[2], DETECTIVE[3] & PUU[4] with the threshold set as having at least 85% overlap. Once the domains are identified, Class, Homology and Topology are assigned to each of the families in the I, N & S class using automatic procedures[5][6]. Finally, Architecture is assigned by manual inspection.

**References:**

[1]: Needleman, S.B. & Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443.

[2]: Siddiqui, A.S. & Barton, G.J. (1995). Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. Protein Sci. 4, 872–884.

[3]: Swindells, M.B. (1995). A procedure for detecting structural domains in proteins. Protein Sci. 4, 103–112.

[4]: Holm, L & Sander, C. (1993). Parser for protein folding units. Proteins 19, 256–268.

[5]: Michie, A.D., Orengo, C.A. & Thornton, J.M. (1996). Analysis of domain structural class using an automated class assignment protocol. J. Mol. Biol. 262, 168–185.

[6]: Orengo, C.A., Brown, N.P. & Taylor, W.R. (1992). Fast structure alignment for protein databank searching. Proteins 14, 139–167.

**SCOP**

The Structural Classification of Proteins(SCOP) is a manually created database which consider domains as the classifying unit. It classifies in a hierarchical manner and have four levels of classification. (a) Family: All proteins which either have a sequence similarity greater than or equal to 30% or those which have lower sequence similarities but are functionally similar are classified into the same family; (b) Superfamily: Those proteins which have low sequence similarity but their structure or function suggest a common evolutionary origin are placed under the same superfamily; (c) Common fold: Two or more proteins are said to be having a common fold if they have the same major secondary structure arrangement and also share similar topological connections; (d) Class: Based on the type of secondary structure a protein has, it can belong to one of the 5 structural classes: all-α, all-𝜷, α/𝜷, α+𝜷 & multi-domain proteins who have domains of different fold and for which no known homologues are present.