**EVEREST**

EVEREST(EVolutionary Ensembles of REcurrent SegmenTs)is a domain family identification method which assumes that a domain is a continuous sequence of amino acids that recurs(non trivially) in the protein space. Thus, domains are those segments of proteins that are conserved and reused throughout evolution. Based on this working principle, EVEREST uses a rigorous process to identify domain families. It begins by constructing a library of protein segments that emerge in all vs. all pairwise sequence alignment using BLAST[1]. Next, it cluster these segments into putative domain families by using the average linkage clustering algorithm. mention about regression tool. A HMM is then chosen not chosen a HMM profile is constructed for each of the putative families, and the procedure is iterated by using the HMM profiles to recreate the putative domains database and repeat the procedure. Mention why is iterative process is carried out.

Refer to caption of Figure for the outline of the algorithm.

**References:**

[1]: Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410.

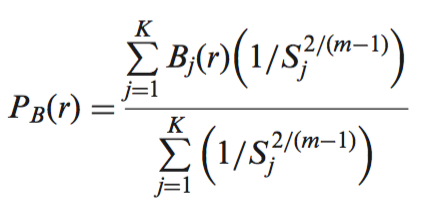
Please note that the method is a very elaborate one and I have removed all the details and the pecularities. But I’m not sure that this write up is sufficient, please let me know if I need to add more.

**FIEFDom**

FIEFDom(Fuzzy Integration of Extracted Fragments for Domains) is a method to predict domain boundaries of proteins from a given sequences, its sequence profile using Fuzzy nearest neighbour algorithm[1]. A three step procedure is used to predict the boundaries. First, a Position Specific Scoring Matrix (PSSM) [2] for the query sequence is generated using PSI-BLAST program and searching against a non-redundant protein sequence database containing information about domain boundaries, e.g., CATH [3], SCOP [4]. PSSM is a 2D matrix which represents the likelihood of each amino acid occurring at every position along the protein sequence. The generated profile is then used to search for similar fragments in the database by doing profile-sequence alignment between the query profile and the proteins in the database using PSI-BLAST program. The expectation value (e-value) is set to 10,000 in this step to ensure that both large and small sized fragments are retrieved. These alignments obtained are parsed and scored using the following scheme [5]:

eqn not typed!

where S is a dissimilarity measure. Thus, the sequence fragments in the database that have high sequence similarity and high statistical significance (or low e-value) with the subsequences of the query protein have low scores. Finally, the domain boundaries (if any) are predicted using the scored fragments. For each residue, the PB is calculated from the domain boundary memberships (B) of the residues in the fragments that are aligned with the current residue. The PB of the query protein is calculated using the following expression for the Fuzzy Nearest Neighbour algorithm:



where, r is the current residue identifier, K is the number of fragments that have a residue aligned with the current residue r, Bj(r) ϵ (0 if the residue lies in the domain and 1 if the residue lies on the domain boundary) is the domain boundary membership of the residue in the jth fragment that has a residue aligned with the current residue r, Sj is the score for the jth fragment defined in the first equation, and m is a fuzzifier [1] that controls the weight of the dissimilarity measure, S.

**References:**

[1]: Keller,J.M., Gray,M.R. and Given,J.A. (1985) A Fuzzy K-Nearest Neighbor Algorithm. IEEE Trans. Syst. Man Cybernetics., 15, 580–585.

[2]: Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389–3402.

[3]: CATH: comprehensive structural and functional annotations for genome sequences.

Sillitoe I, Lewis, TE, Cuff AL, Das S, Ashford P, Dawson NL, Furnham N, Laskowski RA, Lee D, Lees J, Lehtinen S, Studer R, Thornton JM, Orengo CA. Nucleic Acids Res. 2015 Jan, doi: 10.1093/nar/gku947

[4]: Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995)

SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol., 247, 536–540.

[5]: Bondugula,R. and Xu,D. (2007) MUPRED: a tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. Proteins, 66, 664–670.

Ma’am please note that I was having some issue with writing the second equation. Thus for now, I have pasted the image of the equation. I’ll correct it in the final draft

**STRUDL**

STRUDL(STRUctural Domain Limits) unlike many of it’s predecessor methods doesn’t take into account any information on secondary structures and handles any number of domains made up of contiguous or non-contiguous chain segments. The core of the algorithm works on partitioning the 3D structure of a protein into sets of residues such that the interactions between the sets is minimum. This is a graph partition problem which is known to be NP-hard, but the type of graphs representing residue interactions can be analyzed by graph heuristic, which gives an approximate solution. The authors have used Kernighan-Lin graph heuristic algorithm [1] to partition the protein structure into sets or residues with minimum interaction between them. The contact area between atoms is defined as the area of intersection of the Van Der Waals sphere around each atom and the faces of its weighted Voronoi polyhedron[2]. Using the contact-area measure and the Kernighan-Lin heuristic, the procedure identifies the partition with minimum contact area. The partition is then accepted or rejected based on various additional criteria pertaining to the expected properties displayed by a structural domain like size and compactness. When a partition is accepted, the procedure is repeated recursively to generate sub-structures until no further splits are viable.

about minimum contact profile.

When you move on to the review of structure based methods, first give the general approach / assumption on which these methods are based.

**References:**

[1]: Kernighan, B. W.; Lin, Shen (1970). "An efficient heuristic procedure for partitioning graphs". Bell System Technical Journal 49: 291–307.doi:10.1002/j.1538-7305.1970.tb01770.x

[2]: Edelsbrunner H. Algorithms in combinatorial geometry. Berlin Heidelberg: Springer; 1987. 423 p.