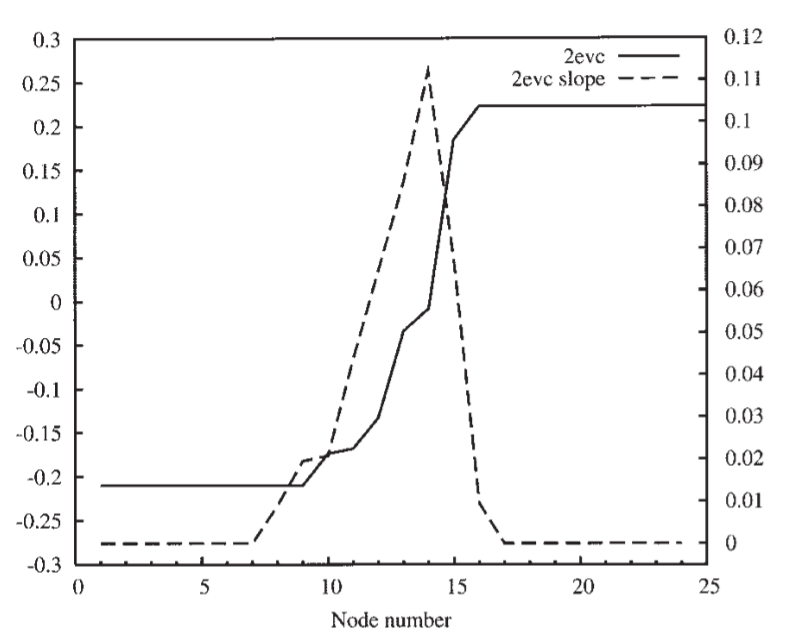
**Vishveshwara’s method**

This method employs graph theoretic technique which takes into account the overall connectivity and topology of the protein structure. It is based on the assumption that intra-domain interactions are more than inter-domain interactions. The method begins by considering the given protein molecule as a graph of connected amino acid residues. Two types of definitions are employed to construct this graph. The first definition constructs by connecting all the alpha carbons which are less than 6.5 Å apart and this graph is labeled as a Protein Backbone Graph(PBG), while the other one constructs by connecting all the side chain carbons, the beta carbons, which are less than 4.5 Å apart and the resulting graph is called Protein Side-Chain Graph(PScG). In the next step, Laplacian matrices are constructed for both the graphs and are diagonalized to obtain the eigen spectra. Using the fact that eigenvector components corresponding to the second lowest eigenvalue(2evc) contain information about clusters present in the graph[1]. Nodes having similar magnitude of second lowest eigenvalue belong to the same cluster. - repetitive An example graph of 25 nodes is plotted with node number on the x-axis and 2evc on the y-axis. how does this graph look like?



The plateaus on the graph represent the nodes which belong to the same cluster while the nodes on the slope are the ones which acts as the interface for their respective cluster. These obtained clusters are the identified domains in the protein. Not at all clear what you are trying to say. Write explicitly how does one identify domain boundary by SV's approach? Any limitations on this approach?

**References:**

[1]: Hall KM. An r-dimensional quadratic placement algorithm. Manage Sci 1970;17:219–229. what is this reference, any standard book on graphs would also have sufficed.

No ref to Sv's paper!

**PDP**

PDP(Protein Domain Parser) attempts to identify domain by hierarchical decomposition of a protein into smaller fragments based on the idea that inter-domain contacts are much more than intra-domain contacts. The program starts off by considering the whole protein chain as one domain consisting of one continuous fragment. At each step, PDP cuts a domain into two domains by two ways: (1) By a single cut in all possible sites in the polypeptide chain, or (2) by a double cut in spatially close (distance between Cα-atoms is <8 Å), but are sequentially distant (more than 35 residues apart). After each attempt the number of contacts *nc(i,j)* between two new formed domains is counted and normalized by the size of the domains, and referred as nnc (normalized number of contacts).

nnc(i, j) = nc(i, j)/(|i|α.|j|α)

Where |i| is the size of the domain i, α = 0.43. The assumption on which this formula is based is that the expected number of contacts between two domains depends on their surface areas, which is proportional to *n*2/3 for a spherical domain of *n* amino acids.

If the minimum of the normalized contacts is less than the threshold, the cut is implemented and the recursive step is repeated for the two new domains. The threshold is computed specifically for each given domain and is equal to one half of the average contact density for the whole domain. After all cuts are made, the contacts between all domains are checked again and domains with the large number of contacts (the normalized number of contacts is greater than two) are combined into one larger domain. At the last step PDP filters out all tiny domains (less than 30 amino acids)

Cut-and-paste from the PDP paper!

Ref?

**DDomain**

DDomain divides a structure into domains using a normalized contact-based domain-domain interaction profile. The working assumption is that each structural domain corresponds to a continuous segment of its amino acid sequence, and the interaction between the domains is the weakest under a correct domain assignment. Domain-domain interaction is estimated by counting the number of contacts between the domains, where a contact is defined by the distance between two residue side-chain centers of mass within a distance cutoff of 6.5 Å. In order to facilitate comparison among domains, the interaction energy is normalized by the size of the individual domains. Thus obtained energy is termed as the interaction profile between the two domain candidates, with one domain being defined from residue 1 to i and another from i+1 to Nr, where Nr is the total number of residues for the given protein. The lowest interaction profile EProfile(Imin) is selected as it implies the weakest interaction between the two domain and is subjected to the following criteria, where Imin is the location for two separated domains:

for a continuous

segment of length > Lcut in both proposed domains (1:Imin and Imin+1:Nr).

Here, Elowcutoff is the maximum allowed profile energy for a residue designated to be domain boundary, Ecutoffexcess is the minimum profile energy that is above the profile energy at the domain boundary, and Lcut is the minimum length of a continuous segment that satisfies the above condition. Finally, if two domains are found, the algorithm is repeated on each of the domain to see if further splits are possible. The above algorithm assumes that a domain is at least 40 residues long. Further, the three parameters Elowcutoff, Ecutoffexcess & Lcut are determined by obtaining various data sets like CATH & SCOP and dividing it into two, training and testing data sets. For a given data set, the three parameters are obtained by optimizing the agreement between number of domains predicted and annotated in a given training set. Optimization is performed by simple grid search in step size of 0.01 for Elowcutoff & Ecutoffexcess and 1 for Lcut.

Fig 1 of the paper might help explain the terms used above.