**Identification of Single & Multi Domain Proteins**

Identification of the number of domains is a crucial problem and has been a topic of research for over 40 years now. In order to identify the number of domains, we first try to segregate single domain proteins from multi domain proteins based on their physical properties. Once we have successfully separated single domain from multi domain proteins, we can move on to identify the number of domains in multi domain proteins. Some of the important physical properties are discussed in the following paragraphs.

The **Length** of a protein can serve as an indicator of the number of domains that the protein might have. A large protein, typically of length more than 200, is expected to have more than one domain. **Interaction Energy** of a protein was calculated by first splitting the protein into two clusters by k-means and then by calculating the total number of pairwise interactions(within 7 Å) between the two clusters and normalizing it by the size of the protein as per equation 1.1. Here Nxy is the total number of pairwise interactions between residues of clusters x & y. Nx & Ny are the number of residues in clusters x & y respectively.

(1.1)

**Radius Of Gyration** by definition is the distribution of components around an axis. A single domain protein is expected to be more compact as compared to a multi domain one. Hence, multi domain proteins are expected to have a higher radius of gyration as compared to single domain proteins. This property was calculated by following the works of Lobanov et al.[1].

Similarly, **Density** of a protein structure can reflect its compactness and thus, single domain proteins are expected to show higher density as compared to multi domain proteins. The density of a protein was calculated by dividing the length of the protein by its approximate volume[2].

We analyzed the above four properties by taking two at a time and found that the combination of Length and Interaction Energy worked best in separating single domains from multi. By choosing a length of 150 and energy 0.25, we achieved the highest accuracy of 83% each in predicting single domain and multi domain proteins. Figure 1.2 shows the distribution of single and multi domain proteins based with respect to their length and interaction energy.

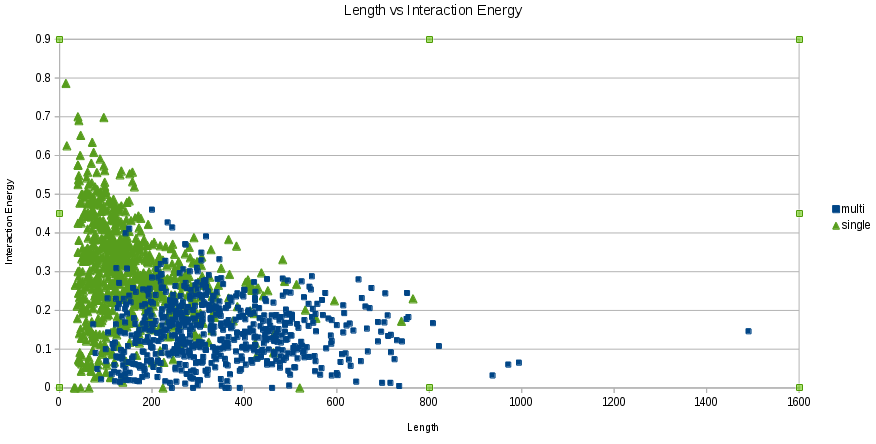


Figure 1.2: Distribution of single and multi domain proteins w.r.t. Length & Interaction Energy

By keeping these two properties as base, we tried to improve the accuracy of prediction by combining the values of Radius of Gyration and Density but observed no significant increase. Thus, only length and energy were chosen as useful parameters to segregate single and multi domain proteins. The computed values of all the four properties on our dataset can be found in appendix 1.

References

1: Lobanov, M.Y., N.S. Bogatyreva and O.V. Galzitskaya, 2008. Radius of gyration as an indicator of protein structure compactness. Mol. Biol., 42: 701-706.

2: 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).