**1.3.1 Sequence Based Methods**

Predicting domains based on the sequence information only and without the knowledge of their 3D structure is a tough task. Some of the early approaches to the problem ranged from

assembling secondary structure elements into domains[1] to identifying domains as those areas having high residue contact density[2]. A more recent approach guesses the number of domains based on their size based distribution[3]. Unfortunately, these methods had poor results and were unreliable.

Current methods rely more on the fact that a domain is a continuous sequence of amino acids that recurs in the protein space. Thus, domains are evolutionary in nature and are those segments of protein that are conserved and reused throughout evolution. Hence, it is observed that sequences which have a substantial sequence similarity(> 30%) share common domains that possesses a common fold and thus usually share similarity in function[4]. Such methods employ an alignment approach where domains are identified by aligning the target sequence against sequences present in the database with known boundaries[5]. This method is efficient but relies on the existence of homology. Also, this method fails in identifying non-contiguous domains as it assigns each conserved segment to a separate domain. In addition to sequence alignments, some methods employ machine learning to further enhance their prediction. Some of the methods using the above techniques are discussed below.

**References**

[1]: Busetta B, Barrans Y. The prediction of protein domains. Biochim Biophys Acta 1984;790:117–124.

[2]: Kikuchi T, Ne’methy, G, Scheraga HA. Prediction of the location of structural domains in globular proteins. J. Protein Chem. 1988;7:427– 471.

[3]: Wheelan SJ, Marchler-Bauer A, Bryant SH. Domain size distributions can predict domain boundaries. Bioinformatics 2000;16:613–618.

[4]: Doolittle, R. F., (1995), The Multiplicity of Domains in Proteins, Annu Rev Biochem, 64, 287- 314.

[5]: Marchler-Bauer, A., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Lu, S., Marchler, G.H., Mullokandov, M., Song, J.S., Tasneem, A., Thanki, N., Yamashita, R.A., Zhang, D., Zhang, N., Bryant, S.H., (2005), CDD: CDD: a Conserved Domain Database for Protein Classification, Nucleic Acids Research, 33, 192-196.