Sequence space graph and some applications

Proteins are major constituents of all cells and central to our understanding of molecular biology. The past and present genome projects have provided us with an exponentially growing wealth of protein sequences from many diverse organisms. However, this wealth creates a new problem. The majority of protein sequences have not been studied experimentally in the laboratory and our first-hand knowledge about them is minimal. The biological function of a novel protein sequence can often be inferred by studying similar protein sequences in other organisms. Just as whole organisms are related to each other by evolutionary descent, so are the proteins between organisms related by inheritance and mutation. We study similarity relationships between proteins using all versus all sequence comparisons. Graph representations and clustering of the similarity relationships at whole protein, domain and residue level will be presented. In particular, the residue level analysis generates a feature space which is useful for extracting functionally divergent protein subgroups and their function specific residues.