**Course: Advanced Bioinformatics**

**Module title: Stage 2 Definitions**

**Module no. : 56**

In this module we provide some basic definitions which are required for Multiple Sequence Alignment and then phylogenetic analysis.

**Motif:** A short conserved region in DNA, RNA or PS which corresponds to structural or functional feature in proteins and shared by several sequences and can be generated by MSA. It can be represented using position-specific scoring matrices

A position-specific scoring matrix, or matrix of scores representing a motif 22 columns, one for each of the 20 amino acids, 2 for the penalties of opening and extending gaps

**Profile:** The rows of the profile: aligned amino acid residues of a group of sequences where residues with the highest scores define a consensus

**Protein family:** A set of evolutionarily related proteins whose members may range from very similar to quite diverse. They often share domains.

Domain is a part of a protein (greater than a motif) that can fold and

carry out a function independently.

**Motif /domain DBs:** Secondary databases, smaller than primary DBs and contain representations of conserved sequences shared by a sequence family Primarily used for annotation of unknown sequences.

Examples: Pfam, Blocks, PRINTS, Prodom, PROSITE

**Protein FAMily (Pfam):** large collection of MSAs of sequence motifs or domains

Pfam-A: curated DB of gapped profiles

Pfam-B: generated automatically from seqs. It originates from the Prodom DB, no overlap with Pfam-A. Hidden Markov Models are used to define domains or to align a set of sequences.