**Course: Advanced Bioinformatics**

**Module title: Sequence Acquisition**

**Module no. : 54**

**FIVE STAGES OF PHYLOGENETIC ANALYSIS:**

Molecular phylogenetic analyses can be divided into ﬁve stages: (1) selection of sequences for analysis, (2) multiple sequence alignment of homologous protein or nucleic acid sequences, (3) speciﬁcation of a statistical model of nucleotide or amino acid evolution, (4) tree building, and (5) tree evaluation. The module sheds light on the first stage.

**Stage 1: Sequence Acquisition**

We can acquire the sequences from many sources, including the following.

1. HomoloGene at NCBI includes thousands of eurkaryotic protein families. HomoloGene entries can be viewed as sequences in the fasta format (or as a multiple sequence alignment).
2. Results from the BLAST family of proteins can be selected, viewed in Entrez. Protein or Entrez Nucleotide, and formatted in the fasta format.
3. Sequences from a large variety of databases can be output in the fasta format (or as multiple sequence alignments). For RNA, these databases include Rfam and the Ribosomal Database. For proteins, these databases include Pfam and InterPro. For viruses, examples include reference databases for human immune deﬁciency virus and hepatitis C virus.

**Types of DBs:** Broadly, we can categorize biological DBs into five broad categories. Here we enlist each of these databases with some of their examples.

1. Sequence Databases
2. Bibliographic Databases
3. Clinical Databases
4. Integrated Databases
5. Structural Databases
6. **Sequence Databases**

DDBJ: DNA Data Bank of Japan

EMBL: European Molecular Biology Laboratory

GenBank

SwissProt: Swiss Protein

1. **Bibliographic DBs**

Used for searching for reference articles

For all (loosely) medically related papers, PubMed from the NCBI

Currently holds over 12 million MEDLINE entries.

Web of Sciences: Free to academics, but requires username and password

PubCrawler; Free to academics, will search journals and sequences daily, weekly or monthly and alert the user when results are found corresponding to their search

1. **Clinical Databases**

Information from Human.

Human Gene Mutation DB, Cardiff, UK:

Registers known mutations in the human genome and the diseases they cause. dbSNP, Bethesda, USA: Largest DB for single nucleotide polymorphisms. Accession numbers in dbSNP are not compatible with other SNP DBs

1. **Integrated Databases**

Overview information from a variety of different DBs, offer links to further information

GeneCards: Extremely thorough overview of a particular gene, with links to various other integrated and clinical databases.

Interpro: Integration of individual protein

Resources; i.e. PRINTS, PROSITE, SMART, ProDom, Pfam, TIGRfam into one DB.

A search will scan entries of each and output results.