Databases

PDB (Protein Data Bank). Experimentally determined three-dimensional structures of proteins and nucleic acids, obtained by X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy,

AlphaFold DB (AlphaFold Database). Computationally determined protein structure predictions for the human proteome and 20 other key organisms.

NCBI (National Center for Biotechnology Information). A collection of databases relevant to biotechnology including (i) GenBank for DNA sequences, (i) PubMed, a bibliographic database for biomedical literature, (iii) NCBI Epigenomics database, (iii) Gene, Online Mendelian Inheritance in Man, the Molecular Modeling Database (3D protein structures), (iv) dbSNP (a database of single-nucleotide polymorphisms), (v) Reference Sequence Collection, a map of the human genome, (vi) a taxonomy browser, (v) coordinates with the National Cancer Institute to provide the Cancer Genome Anatomy Project. The NCBI assigns a unique identifier (taxonomy ID number) to each species of organism. The NCBI hosts **BLAST** a sequence similarity searching program that can quickly and easily do sequence comparisons against the GenBank DNA database.

KEGG (Kyoto Encyclopedia of Genes and Genomes). A collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances. KEGG is utilized for bioinformatics research and education, including data analysis in genomics, metagenomics, metabolomics and other omics studies, modeling and simulation in systems biology, and translational research in drug development.

UniProt (UniProtKB/Swiss-Prot). A manually annotated, non-redundant protein sequence database that combines information from scientific literature and biocurator-evaluated analysis. The aim of UniProt is to mskr accessible all known relevant information about a particular protein. The manual annotation of an entry involves analysis of protein sequence and of the scientific literature.

Classification Databases

ECOD (Evolutionary Classification of Protein Domains). A hierarchical classification of protein domains by evolutionary relationships. Proteins with experimentally determined structures from the PDB database are classified in ECOD. Compared with other classification databases, such as SCOP and CATH, ECOD emphasizes evolutionary relationships.

SCOP (Structural Classification of Proteins). The SCOP database provides structural and evolutionary relationships between all proteins whose structure is known. It provides a broad survey of known protein folds, detailed information about the close structural relatives of any particular protein.

Pfam (Protein Families). A large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs). Pfam identifies domains and generates higher-level groupings, known as clans, and are collections ntries related by similarity of sequence, structure or profile-HMM (hidden Markov models).