**ENZYME DATABASES**

**BRENDA**

* **Enzymes**
* Largest + most diverse group among proteins
* Play essential role in metabolism
* Catalyze + regulate all chemical reactions and metabolic steps within a cell.
* **BRENDA**
* “BRaunschweig ENzyme DAtabase”
* Represents a relational function database containing – comprehensive enzymatic and metabolic data, extracted, continuously updated and evaluated from primary literature.
* **Key Developments –**

1. Conversion of : DB 🡪 Organism – specific information system
2. Improvement of validation and correction of data
3. Standardization of entries – to create prerequisites for a systematic access and analysis

* **Contents of BRENDA**

This DB covers organism – specific information on functional and molecular properties based on –

|  |  |  |
| --- | --- | --- |
| Nomenclature | Reaction and Specificity | Enzyme structure |
| Stability | Application and Engineering | Organism |
| Ligands | Literature references | Links to other databases |

1. **Enzymes –**

* BRENDA contains all enzymes classified according to the system of EC numbers – implemented by IUBMB [International Union of Biochemistry and Molecular Biology]. Nomenclature = based on reaction the enzyme catalyzes ; NOT ON INDIVIDUAL ENZYME MOLECULE.
* **For enzymes having same EC number –** data is periodically updated by – manual extraction of parameters from the literature references accessible via literature databases

**GENOME DATABASES**

**CATH**

**(Class, Architecture, Topology and Homologous Superfamily)**

* **Def –** hierarchical classification of protein domains (sub – sequences of proteins that may fold, evolve and function independently of the rest of the proteins), based on –

sequence information + structural and functional properties

* **4 Main Levels of Classification in a Hierarchical scheme –**

|  |  |  |
| --- | --- | --- |
| **Category** | **Full form** | **Grouping of domains according to** |
| C – level | Class | Secondary structure content  4 categories –   1. Mainly alpha 2. Mainly beta 3. Mixed alpha – beta 4. Domains with only few secondary structures |
| A – level | Architecture | General orientation of their secondary structures |
| T – level | Topology | Connectivity (i.e., the order) of the secondary structures |
| H – level | Homology superfamily | Combination of both –   1. Sequence similarity 2. Measure of structural similarity obtained from SSAP (dynamic programming algorithm). |
| **5 more levels of classification – S, O, L, I, D** | | |
| S, O, L, I – level | --- | Increasing sequence overlap and similarity |
| D – level | --- | Assigns a unique identifier to every domain, thus ensuring that no 2 domains have exactly the same CATHSOLID classification. |

* **Data Accessibility – provides a convenient way to locate and compare similar structures**
* Quick Search Box
* Links provided for –

1. Search by keyword or domain ID
2. Search using a sequence in FASTA format
3. Browse the database from the top of the hierarchy
4. Download datasets

* **CATH DB Construction –**

1. **Criteria for adding new structures –**

* **Data obtained from –** PDB
* **Resolution –** 4 ˚A or better
* **Length –** minimum 40 residues with 70% or more of the side chains resolved

1. **Steps involved in adding new structures –**
2. Submitted protein chains are chopped to obtain domains.
3. Classifications are assigned to the resulting domains.

* **IMPORTANT TERMS –**

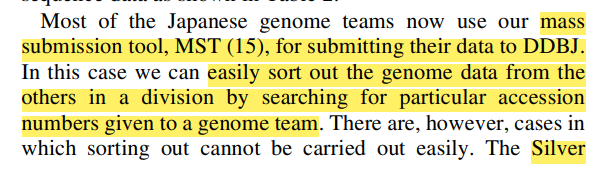
1. FunFams (Functional Families)
2. Structural Clusters (SC)

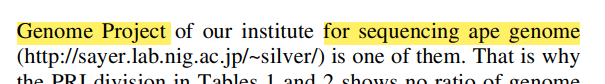
**DDBJ (DNA DATA BANK OF JAPAN)**

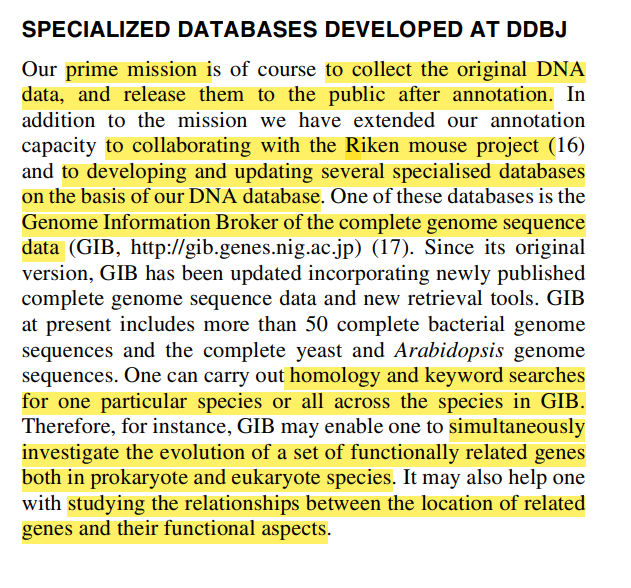
* Genes in an individual organism do not function independently but do interdependently in complex functional networks. Hence study extended to genome.
* **INSD (International Nucleotide Sequence Databases) – composed of –** DDBJ, EMBL Bank and GenBank. {**INSDC (International Nucleotide Sequence Databases Collaboration)}**
* **Aim of DDBJ –** TO COLLECT THE ORIGINAL DNA SEQUENCE DATA, AND RELEASE THEM TO TH EPUBLIC AFTER ANNOTATION.
* As part of INSDC, data thus releases is exchanged with the EMBL Bank and GenBank on a **DAILY BASIS.** This practice allows the 3 data banks to serve users worldwide with the same quality and quantity of data.
* **TAXONOMIC DIVISIONS FROM DDBJ SUBMISSION –**

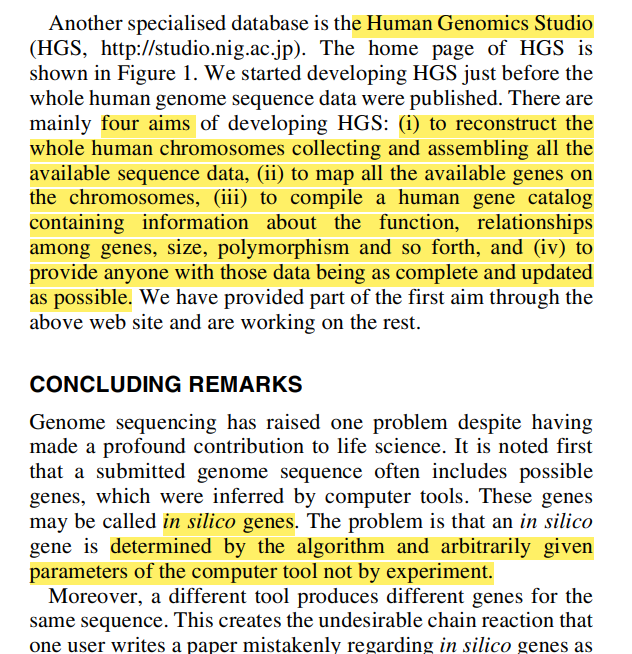
|  |  |
| --- | --- |
| HUM | Human |
| PRI | Primate |
| ROD | Rodent |
| MAM | Mammal |
| VRT | Vertebrate |
| INV | Invertebrate |
| PLN | Plant and fungi |
| BCT | Bacteria |
| VRL | Virus |

Number in parentheses = ratio of data submitted from the Japanese human genome teams in percentage









**MSA**

* Def – alignment of 3 or more biological sequences (protein or NA) of similar length.
* **Uses of MSA –**

1. To detect key functional residues
2. Predict secondary / tertiary structures
3. Infer evolutionary history of a protein family
4. Identify conserved regions in sequences that provide insights into the function of the sequences being studied
5. Identify new members of a protein family by comparing them with similar sequences by aligning and displaying homology

* 3 main tools of MSA – CLUSTAL OMEGA, T – COFFEE, MUSCLE – accessible through EMBL – EBI Bioinformatics Web and Programmatic tools framework.

|  |  |  |  |
| --- | --- | --- | --- |
| **MSA Tool** | **Description** | **Method** | **Special features** |
| **CLUSTAL OMEGA**  **SEQUENCE LIMIT –** 4000 sequences + 4MB of data | Can align 3 or more sequences together – in a computationally efficient and accurate manner.  It uses – **seeded guide trees + HMM (Hidden Markov Models) profile – profile techniques –** to generate alignments between sequences. | **1. PROGRESSIVE APPROACH –** builds alignment step – by – step, starting with 2 most similar sequences, progressively adding others.  **2. Algorithm employs guide trees –** to determine order of sequence alignment - ↑ efficiency  **3.** Uses a combination of **pairwise + MSA –** to achieve final alignment. | **1. Speed –** tool designed to be fast and scalable – suitable for large – scale sequence alignments.  **2. Accuracy**  **3. User – friendly Web interface –** accessible to users unfamiliar with command – line tools |
|  | | | |
| **T – COFFEE**  **(Tree – based Consistency Objective Function**  **For**  **Alignment Evaluation)**  **SEQUENCE LIMIT –** 500 sequences or a maximum file size of 1MB | Can align protein, RNA and DNA sequences – using structural information + homology extension. | **1. Consistency – based MSA program / approach –** integrates O/P of various MSA methods – CLUSTAL, MAFFT, PROBCONS, MUSCLE – into 1 alignment  **2. Library – specific scoring –** builds a library of pairwise alignments and constructs a library – specific scoring function to evaluate consistency of each PA with the MSA.  **3.** Final alignment produced – by optimizing the objective function – based on **consistency scores.** | **1. Versatility –** tool can incorporate alignment from various sources (structure – based + profile – profile alignments) – to ↑accuracy.  **2. Available through Web server –** accessible for users who prefer a graphical user interface.  **3. Consistency –** uses consistency scores to produce more accurate alignments, especially in regions of high variability. |
|  | | | |
| **MUSCLE**  **(MUltiple**  **Sequence**  **Comparison by**  **Log – Expectation)**  **SEQUENCE LIMIT – 5**00 sequences + 1MB of data |  | **1. Progressive approach –** similar to CLUSTAL OMEGA  (WRITE SAME)  **2. Log – expectation scoring scheme –** considers likelihood of observing the observed residues in the sequences – given their evolutionary relationships.  **3.** Allows for **Refinement Iterations ­–** to improve initial alignment. | **1. Speed –** Not as fast as CLUSTAL W ; still efficient and can handle large datasets with good performance.  **2. Accuracy –** tool known for producing highly accurate alignments ; often used when high precision is crucial.  **3. Command – line and Web interface** |

* **MSA Significance** **–**

1. **Homology Inference –**

* MSA – crucial for inferring homology between biological sequences
* Identifying conserved regions through alignment 🡪 indicative of functional and structural importance + evolutionary relationships and functional implications

1. **New member identification –**

* MSA – compares similar sequences 🡪 accurate alignments facilitate recognition of homologous sequence and new members of protein families.

1. **Evolutionary analysis –**

* MSA – considers evolutionary events – mutations, insertions, deletions, rearrangements
* Studying evolutionary history and relationships between sequences 🡪 provides valuable insights into genetic and functional evolution of biological entities.

1. **Benchmarking and efficiency –**
2. **Handling large data sets –**

* MSA – enable alignment of numerous sequences 🡪 allowing for comprehensive comparative analyses and evolutionary studies.
* **REPRESENTATIONS –**

\* = positions having single, fully conserved residue

: = conservation between groups of strongly similar properties / residues

; = -------------------------“------------ weakly --------------“-------------------

**GENOME BROWSERS**

|  |  |  |
| --- | --- | --- |
| **Name** | **Description** | **Key Features** |
| NCBI Taxonomy | * Maintained by NCBI * Comprehensive DB system + standardized classification system * Organizes and categorizes biological species into a hierarchical structure 🡪 provides framework to understand evolutionary relationships | **1. Hierarchical Organization –** Taxonomic hierarchy ; tree – like structure ; Branches –  Root node 🡪 kingdoms 🡪 phyla 🡪 classes 🡪 orders 🡪 families 🡪 genera 🡪 species  **2. Assigns and uses Binomial scientific nomenclature –** ensures standardized naming conventions  **3. Taxonomic Identifier (taxid) –** unique numerical identifier assigned for easy reference and retrieval of information.  **4. Reflects evolutionary relationships –** based on scientific evidence from fields such as genetics, phylogenetics, morphology etc.  **5. Integration with NCBI databases –** via links to access data such as sequences, literature etc.  **6. Taxonomic Browser and Search tools –** provides user – friendly interface  **7. Serves as Standard Reference –** to maintain consistency in naming and classification |
|  | | |
| UCSC | * Web – based browser * Developed by the University of California, Santa Cruz * To visualize and analyze genomic data * Explore genomes * Compare genetic sequences * Study various genomic annotations and data tracks | **1. Genome Visualization –** provides graphical representation of genomes, DNA sequences, genes, exons, introns, regulatory elements etc.  **2. Multiple genome assembly support –** for different species – human, mouse, fruit fly etc.  **3. Data tracks –** tracks for references of gene expression data, chromatin accessibility, epigenetic modifications, evolutionary conservation etc.  **4. Custom tracks –** users may upload own experimental / computational data 🡪 enabling personalized analysis and visulaization  **5. Tools and utilities –** for searching genes / sequences / specific genomic regions ; data retrieval, analysis, export  **6. Continuous updates**  **7. User – friendly interface**  **8. Educational Resources –** tutorials, guides and documentation |
|  | | |
| ENSEMBL | * Developed by the Ensembl project (collaborative effort between EBI + the Wellcome Sanger Institute) * Web – based platform * Comprehensive and user – friendly interface for exploring and analyzing genome assemblies, gene annotations, comparative genomics data etc. | 1. **Genome annotations** 2. **Multiple Genome Alignments** 3. **Variation data** 4. **Customizable data views** 5. **Gene trees and homology** 6. **Functional analysis tools** 7. **User – friendly interface** 8. **Regular updates** |
|  | | |
| GOLD |  |  |
| MBGD |  |  |
| ICTVdb |  |  |