The notion of homology

The notion that in the study of the relationships and order of organisms characters are not equally important predates evolutionary theory. However, in those pre-Darwinian and pre-Linnaean periods, there was no clear idea of what actually constituted the basis for the natural order of taxa. In order to determine the natural order of taxa, comparative biologists tried to discover the essential characteristics of organisms that could reveal this order (except for, of course, the nominalists, for whom all groupings were artifacts of the taxonomist's mind). However, it remained unclear which features or organ systems could be considered as essential characters and there was no consensus about a methodology with which such characters could be discovered. Important advances were made in the theory and method of comparative biology when pre-evolutionary biologists developed the concept of homology, implying equivalence or sameness of organismic parts. Homologous features indicated relations of similarity between structures and disclosed the natural order of nature. Subsequently, evolutionary theory finally provided a causal explanation for homologous similarity, viz. common descent. As a consequence, homology now basically implies equivalence or sameness of organismic parts due to common ancestry. Although the concept of homology attained a central role in post-Darwinian biology, it is also a concept that continues to elicit discussions about its definition, recognition, and causation, as is illustrated by the recent publication of an entire volume devoted to the concept of homology in comparative biology.

Biological homology concept the biological concept of homology is concerned with shared developmental constraints and attempts to causally explain patterns of conservatism in the evolution of morphological characters; it is specifically not concerned with the recognition of phylogenetic relationships and the identification of taxa. This approach to homology falls within the transformational approach in comparative biology because it examines processes of change and stasis. A definition of homology that has been suggested within the context of biological homology is that of Wagner (1989a): 'Structures from two individuals or from the same individual are homologous if they share a set of developmental constraints, caused by locally acting self-regulatory mechanisms of organ differentiation. These structures are thus developmentally individualized parts of the phenotype.'

Types of homology :- Besides concepts of homology, which differ in their explanation of the phenomenon, one can also distinguish between different types of homology. Four types of homology have been distinguished, depending on the kind of comparisons being made.

1. Iterative homology (correspondence between characters in the same individual at the same time)

2. Ontogenetic homology (correspondence between characters of the same individual at different times)

3. Polymorphic homology (correspondence between characters of different individuals of the same species)

4. Supraspecific homology (correspondence between characters of different species or higher taxa)

Early in the days of protein and gene sequence analysis, it was discovered that the sequences from related proteins or genes were similar, in the sense that one could align the sequences so that many corresponding residues match. This discovery was very important, since strong similarity between two genes is a strong argument for their homology. Bioinformatics is based on it.

EUROPEAN MOLECULAR BIOLOGY LABORATORY (EMBL)

It is a molecule-based biology research foundation, maintained by 21 member states and formed in the year 1974. It stores and makes available raw nucleotide sequences. It is situated in UK. European Bioinformatics Institute (EBI) maintains EMBL nucleotide sequence database.

European Bioinformatics Institute (EBI)

The various aims of the organization are as follows:

- To provide freely available data and bioinformatics services to all facets of the scientific community.
- To contribute to the advancement of biology through basic investigator-driv- en research.
- To provide advanced bioinformatics training to scientists at all levels.
- To help disseminate cutting-edge technologies to industry.
- To coordinate biological data provision throughout Europe (www.ebi.ac.uk/).

European Nucleotide Archive (ENA)

Free as well as unrestricted information access on DNA and RNA sequences is provided by ENA. This archive is created using three databases which are Sequence Read Archive, Trace Archive and EMBL Nucleotide Sequence Database (www.ebi.ac.uk/ena). The information in ENA can be extracted manually or programmatically and resultant files can be obtained in various formats like XML, HTML, FASTA and FASTQ. Using accession numbers and other specific text queries the users can obtain individual archives.

EMBL Nucleotide Sequence Database

It contains the high-level genome assembly data of sequences and their functional annotation. The data is store in flat file format.

Data Classes

The different data classes of sequences are

Data Class	Definition
EST	Raw expressed sequence tags without sequence quality information
WGS	Genomic contigs
GSS	Genome survey sequence; single pass, single direction sequence
HTC	High throughput assembled transcriptomic sequence and optional annotation
HTG	High throughput assembled genomic sequence and optional annotation
STD	Assembled and annotated sequences
CON	Scaffolds build from genomic or transcriptomic contigs
STS	Sequence tagged site
PAT	Patent sequences
TSA	Transcriptomic contigs
CDS	Coding sequences

Database entry structure

The EMBL flat-file comprises of a series of strictly controlled line types that are presented in a tabular manner and consists of four major blocks of data.

• Descriptions and identifiers. Entry name, confidential status, molecule type, taxonomic division, and total sequence length found in the **ID** line;

accession number (**AC**);

sequence version (**SV**);

date of creation and last update (**DT**);

brief description of the sequence (**DE**);

keywords (**KW**);

taxonomic classification (**OS, OC**) and links to related database entries (**DR**).

• Citations. The citation details (**RX, RA, RT** and **RL**) of the associated publication and the name (**RA**) and contact details (**RL**) of the original submitter.

• Features. Detailed source information, biological features, feature locations and feature qualifiers (multiple **FT** lines).

• Sequence. Total sequence length, base composition (**SQ**) and sequence.

GenBank (www.ncbi.nlm.nih.gov/genbank/)

It is located in the USA. NCBI since 1992 has provided access to GenBank DNA sequence database through NCBI gateway server freely. The three nucleotide sequence databases GenBank, European Molecular Biology Laboratory (EMBL) and DNA Data Bank of Japan (DDBJ) coordinate among themselves so that all three of them are updated with the latest findings.

A detailed structure of a nucleotide sequence file format in this database includes the following:

1. **Locus:** This can be defined as a title given by GenBank itself to name the sequence entry. It includes the following:
   a. **Locus Name:** Similar to accession number for the sequence.
   b. **Sequence Length:** Tells the number of bases existing in the sequence.
   c. **Molecule-Type:** Identifies the type of nucleic acid sequence. The various types are mRNA (which is present as cDNA), rRNA, snRNA, and DNA.
   d. **GB Division:** Postulates class of the data according to classification criteria of GenBank.
   e. **Modification Date:** The date on which the record was modified.
2. **Definition:** This denotes the name of the nucleotide sequence.
3. **Accession:** This covers accession number, accession version, and GI number. Accession number can be defined as the unique identifier associated with each nucleotide sequence present in the database. If more than one record is created for a particular sequence then it will have the same accession number but all records will have different versions associated with that accession number.
4. **Keyword:** Defined words that were used to index the entries.
5. **The Source:** This describes organism from which sequences have been obtained. The accepted common name is mentioned first and then the scientific name is mentioned. In the end, the taxonomic lineage according to GenBank is specified.
6. **The Citation:** Includes the journal from which with the sequence was derived as initially the sequences were obtained only from published literature.
7. **Features:** These consist of the information derived from the sequence such as biological source, coding region, exon, intron, promoters, alternate splice patterns, mutations, etc.
8. **Sequence:** Contains the following:
   a. Count of presence of each nucleotide in the sequence,
   b. Whole nucleotide sequence,
   c. Beginning of sequence is determined by keyword “ORIGIN”, and
   d. End is marked as “\\”.

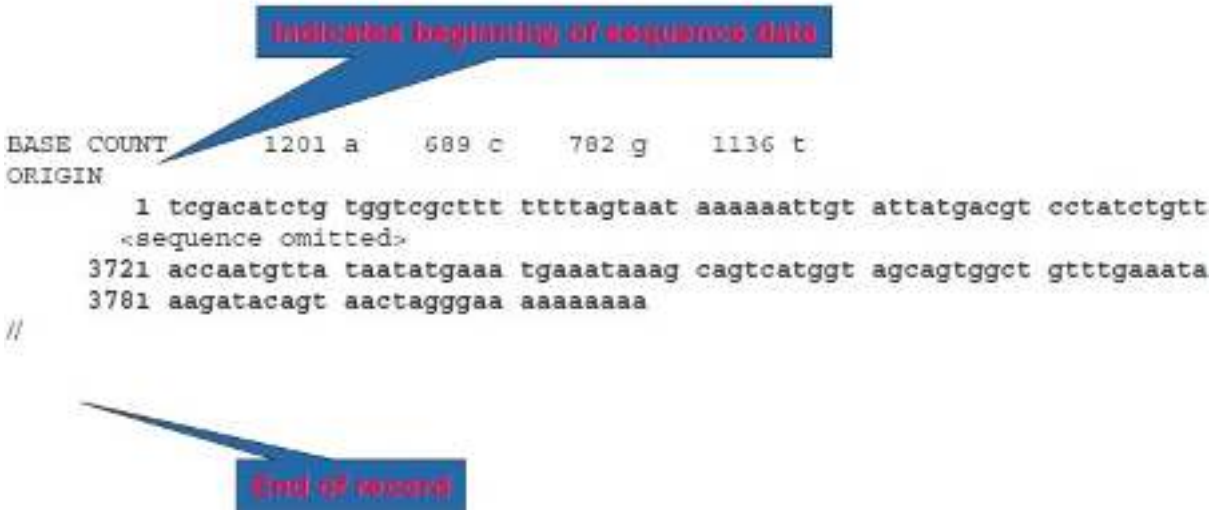
There are many techniques for retrieving and searching data from GenBank.

1 The sequence identifiers can be searched in GenBank along with Entrez Nucleotide.

2 Using BLAST search and then aligning nucleotide sequences to the query sequence.

3 To search the appropriate link and then download nucleotide sequences. (www.ncbi.nlm.nih.gov/genbank/).

Record structure: sequence



Databases

Genbank divisions

PRI: primate sequences
ROD: rodent sequences
MAM: other mammalian sequences
VRT: other vertbrate sequences
INV: invertebrate sequences
PLN: plant, fungal and algal sequences
BCT: bacterial sequences
VRL: viral sequences
PHG: bacteriophage sequences
SYN: synthetic sequences
UNA: unannotated sequences
EST: expressed sequence tags
PAT: patent sequences
STS: sequence tag sites
GSS: genome survey sequences
HTC: high throughput cDNA sequences
HTG: high throughput genomic sequences

DNA DATA BANK OF JAPAN (DDBJ) (www. ddbj.nig.ac.jp)

This biological database resource belongs to **National Institute of Genetics** (NIG) in Japan. DDBJ is the only nucleotide sequence data bank currently present in Asia. Although DDBJ essentially has Japanese researchers as contributors but it also accepts the data from researchers of other countries. It is an associate of the **International Nucleotide Sequence Database Collaboration** (INSDC). The major driving force behind DDBJ operations is the advancement of the quality of INSD as the nucleotide sequence accounts organism development more directly than other biological constituents.

Key tasks of DDBJ Center are as follows:

1. Construction and operation of INSDC which offers nucleotide and amino acid sequence data along with the patent request.
2. Provides searching and analysis of biological data.
3. Training course and journal.

DDBJ Flat File Format

The data submitted in DDBJ is managed and retrieved according to the DDBJ format (flat file). The flat file includes the sequence and the information of who submitted the data, references, source organisms, and information about the feature, etc.

Entrez (http://www.ncbi.nlm.nih.gov/Entrez/.)

Entrez (GQuery) is a user-friendly, versatile, text-based search and retrieval system developed by the NCBI. It searches linked databases using a single word or combination of words entered as search term. Thus, Entrez provides a global query system and forms a web of connections with the databases (nodes in the web of connections).

The search at the NCBI can be performed either using a specific database, or using Entrez across databases simultaneously. Databases that can be selected from the drop-down menu on the NCBI home page, and then the search term can be entered in the space shown. Hitting the "search" button will usually return a number of entries.

Depending on the database selected for search and retrieval, the primary source of some of the retrieved entries may be other related but specialized databases. For example, the Nucleotide, RefSeq, EST, GSS, and Gene databases all have entries on the same nucleotide sequence or part thereof, under database-specific accession numbers and descriptors.

Because all these databases are linked, selecting the Nucleotide database for searching a sequence will retrieve all entries related to the sequence from other related and specialized databases as well. However, selecting a specialized database will retrieve a smaller number of entries.

Alternatively, the user can access the Entrez home page and perform a search across all databases simultaneously by entering the search term in the space shown. Hitting "Search" will return the number of entries available in each database, which is displayed next to the database name.

Try searching

Mus musculus Slco1a6.

Taxonomy database

(contains the names of all organisms that are represented by nucleotide or protein sequences)

UniGene database

(contains non-redundant information on computationally identified transcripts from the same locus across species; described above)

Epigenomics database

(a relatively new database that provides epigenomic data in the context of biological sample information).

UniGene is an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location.

In addition to sequences of well-characterized genes, hundreds of thousands novel expressed sequence tag (EST) sequences have been included. Consequently, the collection may be of use to the community as a resource for gene discovery.

Or

UniGene is a largely automated analytical system for producing an organized view of the transcriptome. By analysing sequences known to be expressed, and the libraries or samples from which they were derived, it is possible to organize the data into gene-specific clusters, and, in some cases, evaluate the patterns of expression by tissue, health status, and age.

transcriptome

The transcriptome refers to the full set of transcripts in a cell assembled by a method called RNA-seq in which RNA from cells is collected, sampled, and sequenced. It includes alternative splice variants, variants created by alternative transcription initiation and alternative transcription termination, and noncoding RNA genes.

Data Model

The data model for UniGene is straightforward. Identify sequences of RNA molecules, the source of those sequences (species, tissue, age, health status), compute when independent sequences are derived from the same gene based on sequence similarity, and report the results. Historically this computation was based on ESTs (Extended Sequence Tags), but now the vast majority of sequences are either full-length clones or RNAseq data.

Access

The UniGene web pages are retired at the end of July, 2019. But one can download the final UniGene builds from ftp website (https://ftp.ncbi.nlm.nih.gov/repository/UniGene/) and also one can able to match UniGene cluster numbers to Gene records by searching Gene with UniGene cluster numbers(https://www.ncbi.nlm.nih.gov/gene/advanced).