EE550
Computational Biology

Week 5 Course Notes

Instructor: Bilge Karaçalı, PhD

Topics

- Database systems
  - Database management systems
  - Data models
    - Relational data model
    - Semistructured data model
  - Integration of information
- Gene databases
- Protein databases

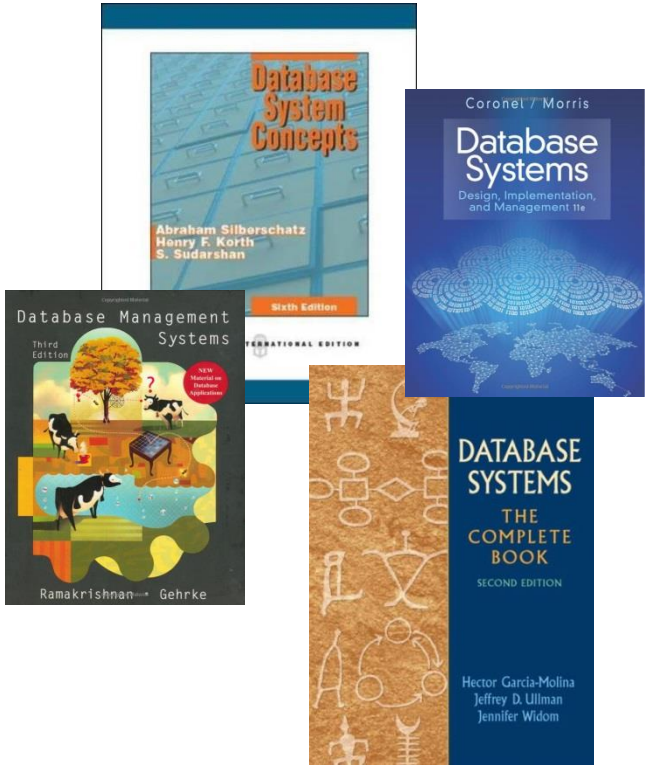
Database Systems and Computational Biology

- Computational biology owes its success to rapid and wide dissemination of quantitative data on molecular biology
  - Large amounts of data are generated by high-throughput techniques
  - The generated data are analyzed using computational algorithms whereby knowledge about the biomolecular machinery of the living systems are obtained
  - Both the data and the results of the analyses are made available to all
- Storing and managing large amounts of biomolecular data **efficiently** is of paramount importance

➔ Database management systems

Database Management Systems

- Databases are collections of information
  - Banking systems
  - Store inventories
  - Student records
  - Gene and protein databases
- A database management system allows:
  - Creating and structuring new databases
  - Accessing/querying/mining the data
  - Storing and expanding large amounts of data
  - Recovering the data in case of errors or failures
  - Controlling and monitoring access to the data
- Database management systems are maintained by system administrators that allow users to access and/or to modify the data

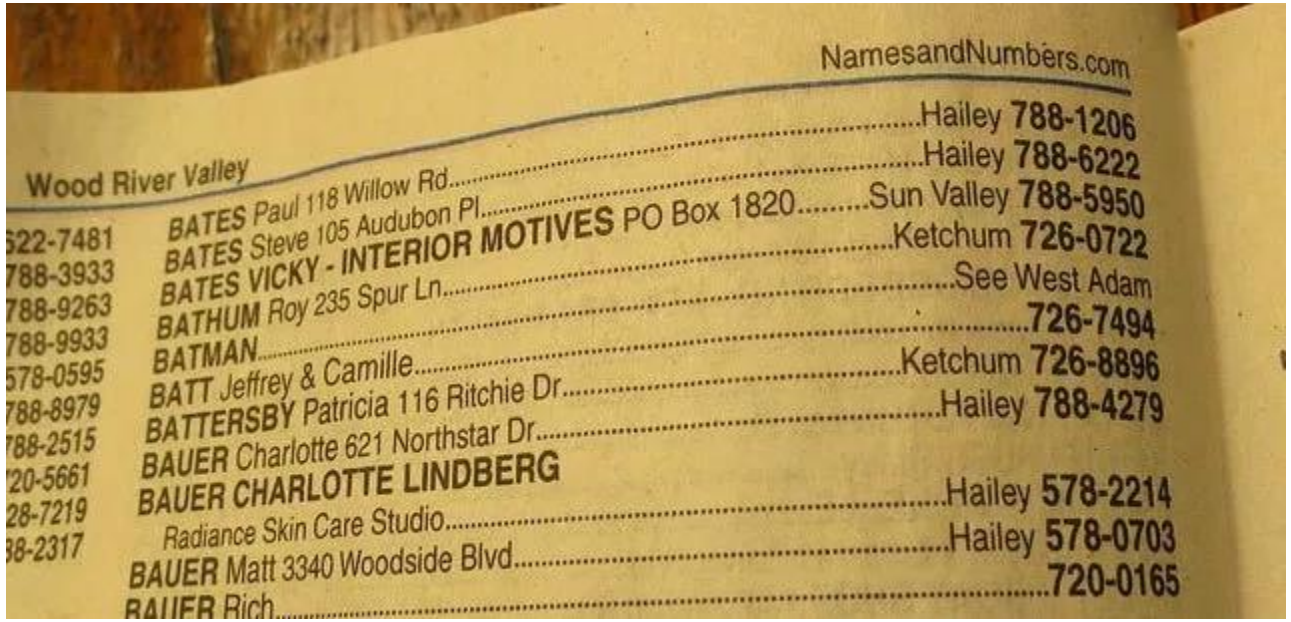


Data Models

- A database is an organized collection of information
- The manner in which the information is stored and accessed is structured by data models
- Data models address three basic requirements:
  - specify the structure of the stored data
    - Names are strings of characters
    - Dates are formatted DD/MM/YYYY
    - …
  - establish the acceptable values for the data
    - Grades can only be AA, BA, BB, CB, …
    - Days of the months must be integer numbers between 1 and 31
    - …
  - define what operations can be performed on the data
    - Modifications
    - Searches

The Relational Data Model

- The data is organized around a common “relation”
  - In a phonebook, names are “related” with phone numbers and addresses
  - The database is created in the form of a table where the information is stored in columns
    - Each row in a phonebook carries a name, an address, and a phone number



Source: https://boingboing.net/2017/06/10/how-adam-west-played-a-prank-u.html

The Relational Data Model

- Elements of a relational model
  - Relation determines the nature of the information
  - Attributes are information fields to be filled for each entry
    - Name, address, phone number
  - Schema specifies the structure of the database
    - Phonebook(Name,address,phone_number)
  - Entries (tuples) are the contents of the database (i.e., the data)
    - Abogu Sname, 123 Fake St., 2321234567
    - …
  - Domains are the data types for the attribute values
    - Name:string, address:string, phone_number:integer

The Relational Data Model

- Remarks:
  - A relation is a set, not a list
    - The order in which the entries are presented does not matter
    - Whichever order is presumed, the relation is still the same
  - The order of the attributes in the schema can be changed
    - The order of the attribute values has to change accordingly
  - A relation is to be dynamically updated as new information becomes available
    - New data is made available continuously
    - The new data is to be incorporated into the relation
    - This also covers the instances where corrections or revisions on the existing entries are required
  - Certain keys are enforced into the relations to prevent multiple entries for the same information
    - “If it looks like an apple and tastes like an apple, it must be an apple”

The Relational Algebra

- The set of operations that can be performed on a relation is called the relational algebra
- The relational algebra allows constructing new relations from existing relations for rapid and easy access to stored information
- An algebra consists of operations and operands
  - Algebra of mathematics uses ‘+’, ‘-’, ‘×’, and ‘÷’ operations on operands that are numbers (real or complex)
- The operations in a relational algebra serve to search or manipulate the data in the database
  - Usual set operations: union, intersection, and difference
  - Restriction operations: selection (for rows), and projection (for columns)
  - Combination operations
  - Renaming operations (to alter the schema)
- Queries are relational algebra expressions

The Semistructured Data Model

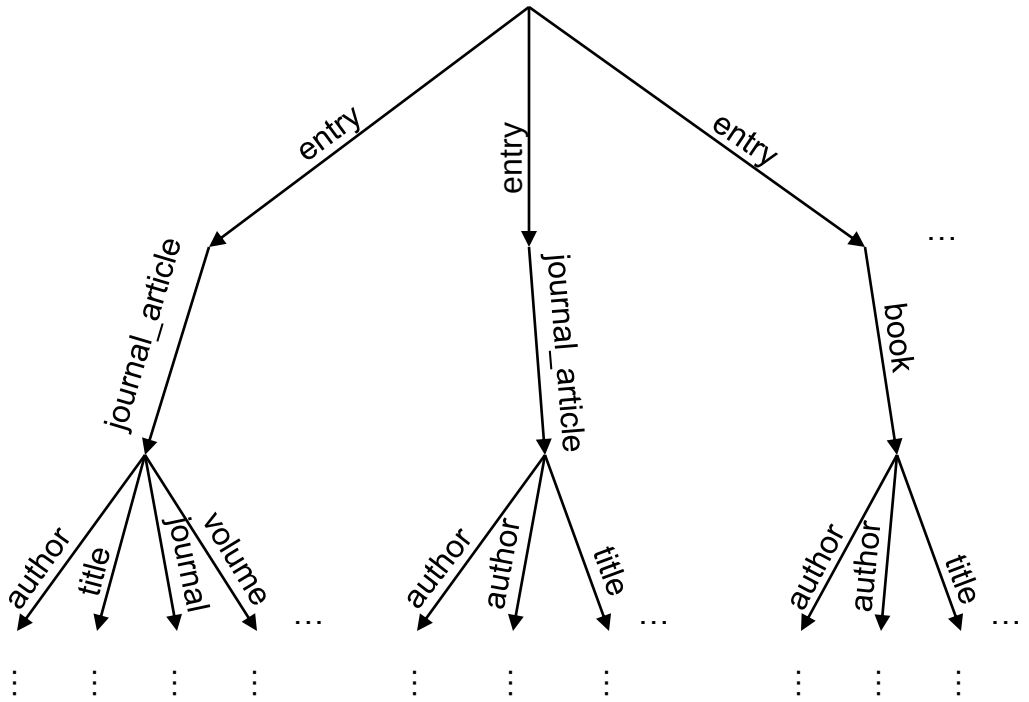
- The necessity for an all-encompassing schema imposes constraints on the data to be stored in a database
- In applications where such constraints are not suitable, the database structure needs to be loosened to allow flexibility to incorporate data
  - The world-wide-web (HTML)
  - Bioinformatics data
- The semistructured data model goes around this problem by representing the data in labels giving meaning to the underlying database structure
  - Self-referencing strategy where the schema of the database is contained within the data itself
  - Extensible mark-up language (XML)

The Semistructured Data Model

- The underlying strategy in the semistructured data model is to represent the information organized in trees
  - Branches are labeled with symbols
  - Internal nodes do not necessarily contain meaningful data
  - The data is contained in the leaf nodes
- Access to the data via browsing or querying is carried out by moving up and down the hierarchies

The Semistructured Data Model

```
<Bibliography>
    <journal_article>
        <author>…</author>
        <title>…</title>
        <journal>…</journal>
        <volume>…</volume>
        <issue>…</issue>
        <pages>…</pages>
        <year>…</year>
    </journal_article>
    <journal_article>
        <author>…</author>
        <author>…</author>
        <title>…</title>
        <journal>…</journal>
        <volume>…</volume>
        <issue>…</issue>
        <pages>…</pages>
        <year>…</year>
    </journal_article>
    <book>
        <author>…</author>
        <author>…</author>
        <title>…</title>
        <year>…</year>
    </book>
</Bibliography>
```



Integration of Information

- Data of interest usually resides in many different sites
  - In general, each site can
    - employ a different file format
    - store different parts of the desired information
- Fusing all these different data banks together into one functional database requires resolving several issues of
  - Compatibility
    - “address” or “addr”
  - Equivalence of entries
    - “MUC4” or “MUC-4”
  - Inconsistencies
    - “phone_number=2321234567” or “phone_number=2321234576”
  - Continuation of existing services and applications
    - Legacy databases support many running applications
    - The operation of existing applications are **not to be disturbed**

Integration of Information

- The first task for information integration is to establish **wrapper applications** that can translate the same bit of data into a form understandable to any of the involved data banks and back
  - Translation of stored information from one format to another
  - Translation of queries to the format of each databank and as well as translation of query results
- The integration of data from several different sites can be carried out using three main strategies
  - Federation:
    - All sites employ wrappers to communicate with every other site
  - Warehouse:
    - Individual databank contents are translated into a unified global database with its own structure
  - Mediator:
    - Acts as a virtual warehouse
    - Receives the requests, distributes it to all sites, and organizes the results

Gene Databases

- Sequence repositories initially stored the sequenced gene information
- Eventually, the available information in gene sequences was structured in publicly available annotated databases
- These databases also provide functionality for
  - Searching through the sequence data
  - Carrying out comparative evaluations of sequences or sets of sequences
  - Identifying the common or distinguishing elements across sequence groups associated with different conditions
    - Human vs. chimpanzee
    - Normal cells vs. cancer
    - Cell lines vs. cancer
    - Breast cancer vs. prostate cancer
    - …

Gene Databases

- Gene databases usually store data in plain text files
  - The information is presented as text typed into the file
  - The contents of the file can be displayed and read across all computing hardware platforms → parsing algorithms!!
    - Text files are sequences of ASCII characters
    - No encoding or sophisticated formatting takes place
- Even then, a basic understanding of the data format (within the text file) is required
  - This implies
    - organizing the stored information in a manner that is well-suited for the intended applications, and
    - making the format explicit so that all applications can access the information correctly
      - via dedicated wrappers and regardless of their own data representations

Gene Databases

- Compliance with the announced data format is paramount for reliable use of the presented data
  - Large data quantities prevent manual validation of each entry and small errors may go unnoticed
  - Such small errors would
    - Corrupt the data included in the study
    - Produce erroneous results
- Cross-checking measures are usually implemented in order to minimize the occurrence of such errors
  - New entries are not made available for general purpose operations
  - They undergo a series of independent validation steps before they are fully incorporated into the main data bank

Gene Databases

- A common format for sequence data is the FASTA format:

```
>ref|NC_000003.10|NC_000003:c197023545-196959307 Homo sapiens …
TACAGCCCCAAGGTCGCTCCCTCTGGGGCCCTTTCTTCCCCATTCTTCCCAGCAGCCCAAAGCTCTGGTG
GGACAGGGGCAGCCCCTGGGGAGGGAGGAGAGGACCCAGGAACCCGGCTAGGAGGGTGGCCCACCCATTT
CCAGTGTGACCTGTTCCCATTCCCCCATGTCTCCTCCCATCCCTCCCGCCACTCAGCTCAGGCTGATGAG
AAGCAGAGCAACGGGTGTATCGGTGTTTTCTTTCCTGGTGGGGTAGTGGGGTGGGGCTGAGGAGAGAAAA
```

…

  - The entries following ‘>’ denote the attributes of the sequence including
    - the accession number
    - the name
    - a brief description
  - The description is followed by the sequence data typed using the 4 nucleotide codes
    - 60 characters per line except perhaps the last line

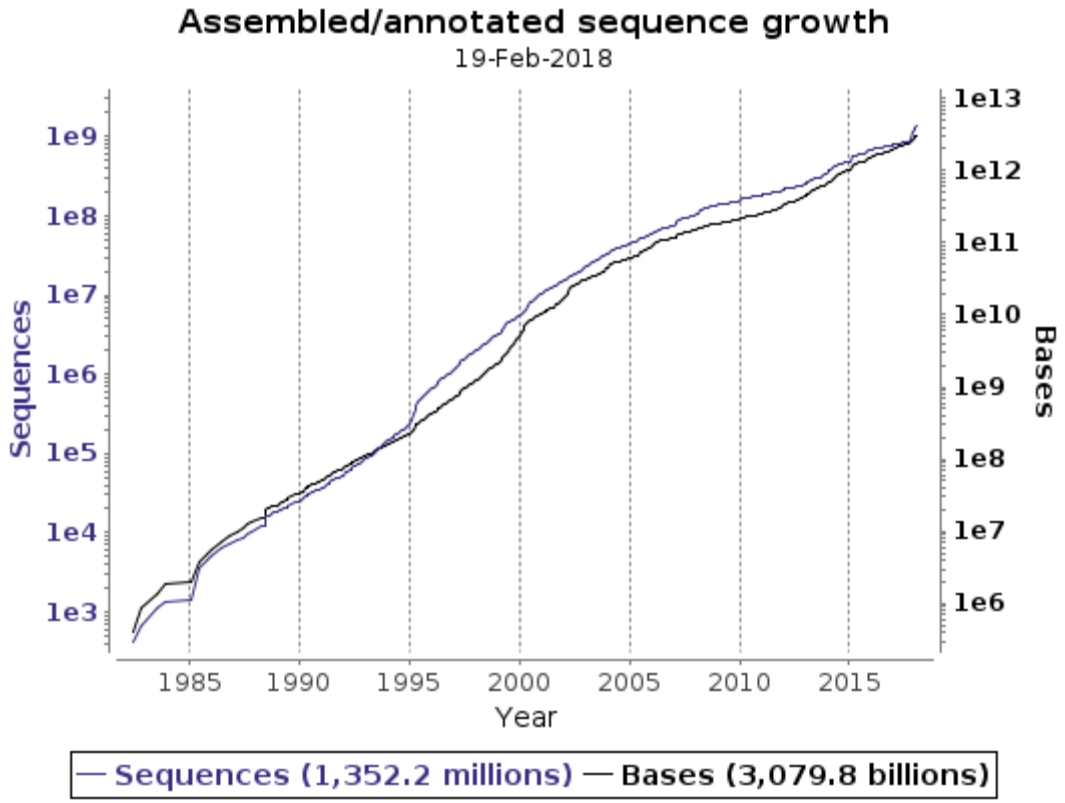
Gene Databases

- Members of International Nucleotide Sequence Database Collaboration
  - European Molecular Biology Laboratory nucleotide sequence database – European Nucleotide Archive
    - URL: http://www.ebi.ac.uk/ena
    - Maintained by the European Bioinformatics Institute (UK)
  - GenBank
    - URL: http://www.ncbi.nlm.nih.gov/Database/
    - Maintained by the National Center of Biotechnology Information (USA)
  - DNA Databank of Japan
    - URL: http://www.ddbj.nig.ac.jp/
    - Maintained by the National Institute of Genetics (Japan)

Gene Databases: ENA

- The primary collection of nucleic acid sequences in Europe
  - URL: http://www.ebi.ac.uk/ena
- Comprises nucleic acid sequences collected from a variety of sources
  - Published sequences
  - Individual scientists
  - Research groups
  - European Patent Office
  - Partner institutions
  - …
- Currently contains
  - 1,157,925,701 entries with
  - 2,700,988,919,811 nucleotides from
  - more than 150 species

Gene Databases: ENA



Source: http://www.ebi.ac.uk/ena/about/statistics/

Gene Databases: ENA

- Data structure:
  - The entries are both human and computer readable
  - The information fields (i.e., the attributes) are arranged systematically for easy access
    - Queries need not go through the whole file to find what they are looking for
  - Some of the attributes are:
    - ID: Entry name, taxonomic division, and the sequence length
    - AC: Unique accession number
    - DE: Description of the entry containing information on the genes encoded by the sequence, the region of the genome from where it is derived, …
    - RN, RP, …: Literature references
    - DR: Cross-references to other nucleic acid sequence and protein databases
    - CC: General comments
    - KW: Keywords
    - FT: Feature table containing information on sequence characteristics and annotations
    - …

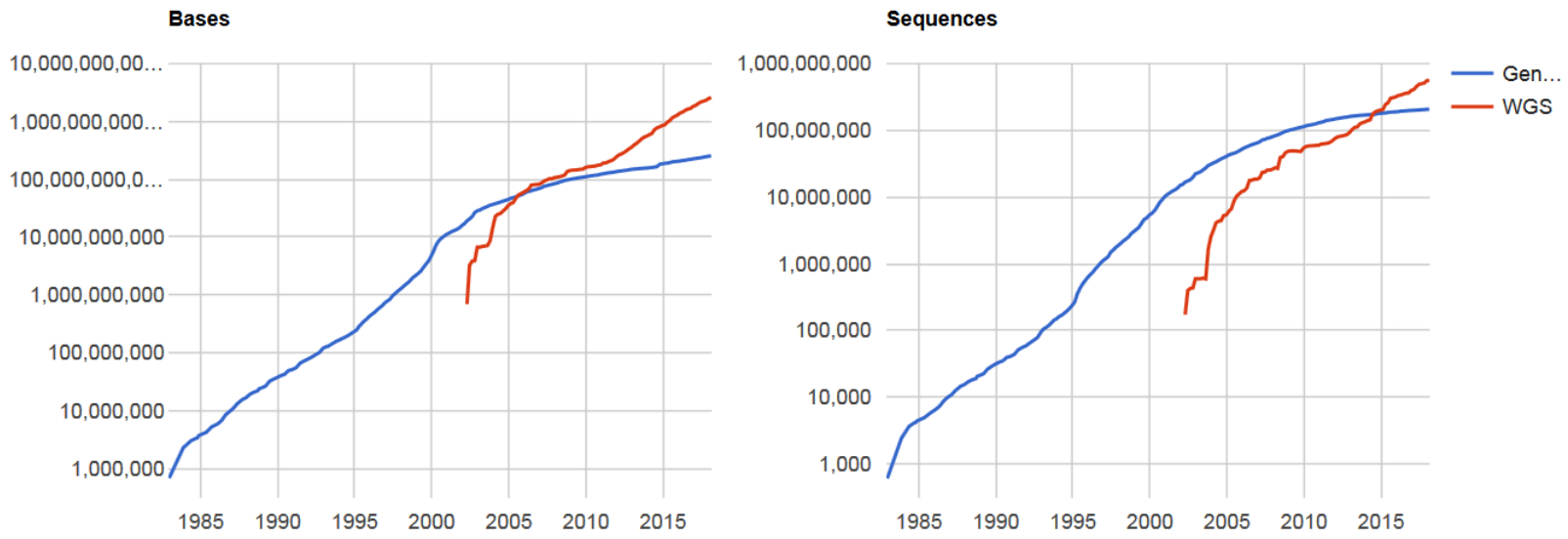
Gene Databases: ENA

```
ID   AJ400850; SV 1; linear; mRNA; STD; HUM; 6575 BP.
XX
AC   AJ400850;
XX
DT   20-OCT-2000 (Rel. 65, Created)
DT   07-OCT-2008 (Rel. 97, Last updated, Version 2)
XX
DE   Homo sapiens mRNA for MUC4 protein splice variant sv13 (MUC4 gene)
XX
KW   alternative splicing; MUC4 gene.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
XX
RN   [1]
RP   1-6575
RA   Choudhury A.;
RT   ;
RL   Submitted (19-APR-2000) to the EMBL/GenBank/DDBJ databases.
RL   Choudhury A., Department of Biochemistry and Molecular Biology, University
RL   of Nebraska Medical Center, 984525 Nebraska Medical Center, NE, 68198-
       4525,
RL   USA.
XX
RN   [2]
RA   Choudhury A., Moniaux N., Hollingsworth M.A., Aubert J.P., Batra S.K.;
RT   "Human MUC4 mucin splice variants in pancreatic adenocarcinoma";
RL   Unpublished.
XX
DR   Ensembl-Gn; ENSG00000145113; Homo_sapiens.
DR   Ensembl-Tr; ENST00000308466; Homo_sapiens.
…
DR   Ensembl-Tr; ENST00000405167; Homo_sapiens.
DR   H-InvDB; HIT000247383.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..6575
FT                   /organism="Homo sapiens"
FT                   /mol_type="mRNA"
FT                   /tissue_type="pancreatic tumour"
FT                   /db_xref="taxon:9606"
FT   CDS             72..3545
FT                   /gene="MUC4"
FT                   /product="MUC4 protein splice variant sv13"
FT                   /db_xref="GOA:Q99102"
FT                   /db_xref="HGNC:7514"
FT                   /db_xref="InterPro:IPR000742"
FT                   /db_xref="InterPro:IPR001846"
…
```

Gene Databases: GenBank

- The primary genomic sequence database in the USA – maintained by the NIH
  - URL: http://www.ncbi.nlm.nih.gov/genbank
- First released in 1982 with 606 entries, now accepts sequence information from
  - Direct author submissions
  - Large scale sequencing projects funded by the US Federal Government
  - Genome Sequence Data Base
  - US Patent and Trademark Office
  - Partner institutions
  - …
- Currently contains
  - 207,040,555 entries with
  - 253,630,708,098 bases primarily from human, mouse and rat genomes

Gene Databases: GenBank



Source: http://www.ncbi.nlm.nih.gov/genbank/statistics

Gene Databases: GenBank

- The database is linked to all the other bioinformatics resources made available by the NIH NCBI's Entrez retrieval system
  - DNA and protein databases
  - Genome mapping
  - Phylogenetic, gene expression, protein structure information
  - Literature through MEDLINE (PubMed)
  - Sequence similarity search through BLAST
  - …

Gene Databases: GenBank

- Data structure:
  - The main files in the database are the sequence files containing the sequence and the associated annotation
  - The attributes for each entry are organized with several keywords summarizing the information
    - LOCUS: Labels the entry
    - DEFINITION: Describes the sequence in human terms
    - ACCESSION: Unique number assigned to the sequence
    - VERSION: Indicates the revision number
    - KEYWORDS: Lists several short phrases associated with the entry
    - SOURCE: Indicates the source (organism) from which the sequence was derived
    - REFERENCE: Literature on the sequence
    - COMMENT: Any additional comments on the entry
    - FEATURES: Feature table with details like sequence properties, coordinates of its coding sequence, etc…
    - After the attributes, the sequence itself follows
  - The entry ends with a line containing “//”

Gene Databases: GenBank

LOCUS NM_004532 4415 bp mRNA linear PRI 14-SEP-2008
DEFINITION Homo sapiens mucin 4, cell surface associated (MUC4), transcript variant 4, mRNA.
ACCESSION NM_004532
VERSION NM_004532.4 GI:167736352
KEYWORDS .
SOURCE Homo sapiens (human)
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 4415)
AUTHORS Ponnusamy,M.P., Singh,A.P., Jain,M., Chakraborty,S., Moniaux,N. and Batra,S.K.
TITLE MUC4 activates HER2 signalling and enhances the motility of human ovarian cancer cells
JOURNAL Br. J. Cancer 99 (3), 520-526 (2008)
PUBMED 18665193
REMARK GeneRIF: findings demonstrate that MUC4 plays a role in ovarian cancer cell motility, in part, by altering actin arrangement and potentiating HER2 downstream signalling in these cells
COMMENT REVIEWED REFSEQ: This record has been curated by NCBI staff. The reference sequence was derived from AJ000281.1, AC069513.28, AJ242549.1, EF091824.1, DB227644.1, BC131770.1 and AW190850.1.
On Feb 13, 2008 this sequence version replaced gi:112382232.
…

Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Entrez Gene record to access additional publications.
COMPLETENESS: complete on the 3' end.

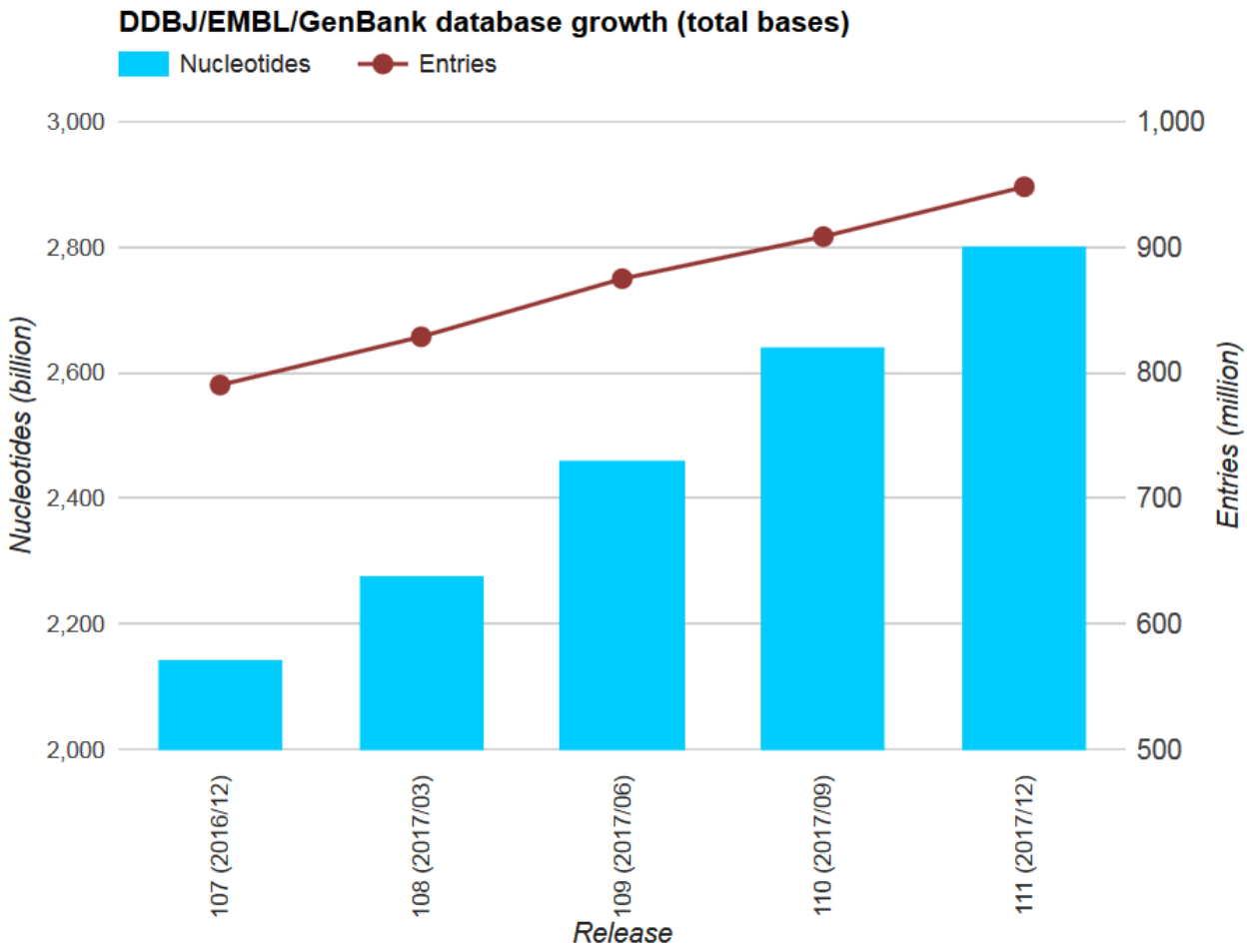
PRIMARY	REFSEQ_SPAN	PRIMARY_IDENTIFIER	PRIMARY_SPAN	COMP
1-27	AJ000281.1	2-28		
28-28	AC069513.28	15243-15243		
29-420	AJ000281.1	29-420		
421-722	AJ242549.1	1-302		
723-1062	EF091824.1	3188-3527		
1063-1093	DB227644.1	353-383		
1094-3321	AJ242549.1	674-2901		
3322-3322	AC069513.28	73834-73834		
3323-3596	AJ242549.1	2903-3176		
3597-3878	BC131770.1	121-402		c
3879-4139	AJ242549.1	3459-3719		
4140-4251	AW190850.1	158-269		c
4252-4415	AJ242549.1	3827-3990		

FEATURES Location/Qualifiers
source 1..4415
/organism="Homo sapiens"
/mol_type="mRNA"
/db_xref="taxon:9606"
/chromosome="3"
/map="3q29"
gene 1..4415
/gene="MUC4"
/synonym="HSA276359“

Gene Databases: DDBJ

- Acts as the primary genomic database in Japan
  - URL: http://www.ddbj.nig.ac.jp/
- First released in 1986, receives its content from
  - Mass submissions from genome sequencing projects
  - Patent offices
  - Partner institutions
  - …
- Parallels EMBL and GenBank in that it contains
  - 948,165,315 entries with
  - 2,802,943,314,196 bases

Gene Databases: DDBJ



Source: http://www.ddbj.nig.ac.jp/breakdown_stats/dbgrowth-e.html#dbgrowth-graph

Gene Databases: DDBJ

- Data structure:
  - The organization of the sequence data along with the associations mirrors that used by the GenBank with the attributes
    - LOCUS: Labels the entry
    - DEFINITION: Describes the sequence in human terms
    - ACCESSION: Unique number assigned to the sequence
    - VERSION: Indicates the revision number
    - KEYWORDS: Lists several short phrases associated with the entry
    - REFERENCE: Literature on the sequence
    - FEATURES: Feature table with details like sequence properties, coordinates of its coding sequence, etc…
    - …

Gene Databases: DDBJ

```
LOCUS       AJ000281               3595 bp    mRNA    linear   HUM 07-OCT-2008
DEFINITION  Homo sapiens mRNA for mucin protein, MUC4.
ACCESSION   AJ000281
VERSION     AJ000281.1
KEYWORDS    MUC4 gene; mucin.
SOURCE      Homo sapiens
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1
  AUTHORS   Aubert,J.P.
  JOURNAL   Submitted (04-JUL-1997) to the EMBL/GenBank/DDBJ databases. Aubert
            J.P., Biologie et Physiopathologie des Cellules Mucipares, INSERM
            U377, Place de Verdun, 59045 lille cedex, FRANCE.
  REMARK    Revised by [3]
…
FEATURES             Location/Qualifiers
     source          1..3595
                     /db_xref="H-InvDB:HIT000243800"
                     /organism="Homo sapiens"
                     /chromosome="3"
                     /map="q29"
                     /mol_type="mRNA"
                     /clone_lib="lambda gt10"
                     /clone="JER103"
                     /tissue_type="colon mucosa"
                     /db_xref="taxon:9606"
     sig_peptide     461..541
                     /gene="MUC4"
…
BASE COUNT         973 a         1212 c         732 g         678 t
ORIGIN
        1 gtacagcccc aaggtcgctc cctctgggcc ctttcttccc cattcttccc agcagcccaa
       61 agctctggtg ggacaggggc agcccctggg gagggaggag aggacccagg aacccggcta
      121 ggagggtggc ccacccattt ccagtgtgac ctgttcccat tcccccatgt ctcctcccat
      181 ccctcccgcc actcagctca ggctgatgag aagcagagca acgggtgtat cggtgttttc
      241 tttcctggtg gggtagtggg gtggggctga ggagagaaaa gggtgattag cgtggggccc
      301 cgccctcttt tgtcctcttc ccaggttccc tggccccttc ggagaaacgc acttggttcg
      361 ggccagccgc ctgaggggac gggctcacgt ctgctcctca cactgcagct gctgggccgt
      421 ggagcttccc cagggagcca gggggacttt tgccgcagcc atgaaggggg cacgctggag
      481 gagggtcccc tgggtgtccc tgagctgcct gtgtctctgc ctccttccgc atgtggtccc
      541 aggaaccaca gaggacacat taataactgg aagtaaaact cctgccccag tcacctcaac
      601 aggctcaaca acagcgacac tagagggaca atcaactgca gcttcttcaa ggacctctaa
      661 tcaggacata tcagcttcat ctcagaacca ccagactaag agcacggaga ccaccagcaa
      721 agctcaaacc gacaccctca cgcagatgat gacatcaact ctttttctt ccccaagtgt
      781 acacaatgtg atggagactg ttacgcagga gacagctcct ccagatgaaa tgaccacatc
      841 atttccctcc agtgtcacca acacactcat gatgacatca aagactataa caatgacaac
      901 ctccacagac tccactcttg gaaacacaga agagacatca acagcaggaa ctgaaagttc
      961 taccccagtg acctcagcag tctcaataac agctggacag gaaggacaat cacgaacaac
     1021 ttcctggagg acctctatcc aagacacatc agcttcttct cagaaccact ggactcggag
     1081 cacgcagacc accagggaat ctcaaaccag caccctaaca cacagaacca cttcaactcc
     1141 ttctttctct ccaagtgtac acaatgtgac agggactgtt tctcagaaga catctccttc
     1201 aggtgaaaca gctacctcat ccctctgtag tgtcacaaac acatccatga tgacatcaga
     1261 gaagataaca gtgacaacct ccacaggctc cactcttgga aacccagggg agacatcatc
     1321 agtacctgtt actggaagtc ttatgccagt cacctcagca gccttagtaa cagttgatcc
…
     3421 ttcctcggca tccacaggtc acaccacccc tcttcatgtc accgatgctt cctcagtatc
     3481 cacaggtcac gccacccctc ttcctgtcac cagcccttcc tcagtatcca caggtcacac
     3541 cacccctctt cctgtcaccg acgcttcctc agtatccaca ggccacccca cccca
//
```

Protein Databases

- Protein sequence
  - PIR
  - MIPS
  - UniProt
  - GenePept
- Protein families
  - PFAM
  - PROSITE
  - PRINTS
- Protein structure
  - PDB
- Protein-protein interactions
  - DIP
  - KEGG
- Functional groups
  - Gene ontology
  - InterPro

Summary

- Gene sequence databases store the genomic data obtained from many species
  - Millions of sequences along with annotations on
    - Associated genes
    - Sequence features
    - Relevant literature
- These databases follow their own file formats for storing and displaying the data
- Coordination between the partner institutions maintaining the corresponding databases addresses key issues of
  - Consistency
  - Error-free data dissemination
  - Cross-referencing