BIOINFORMATICS

PRINCIPLES

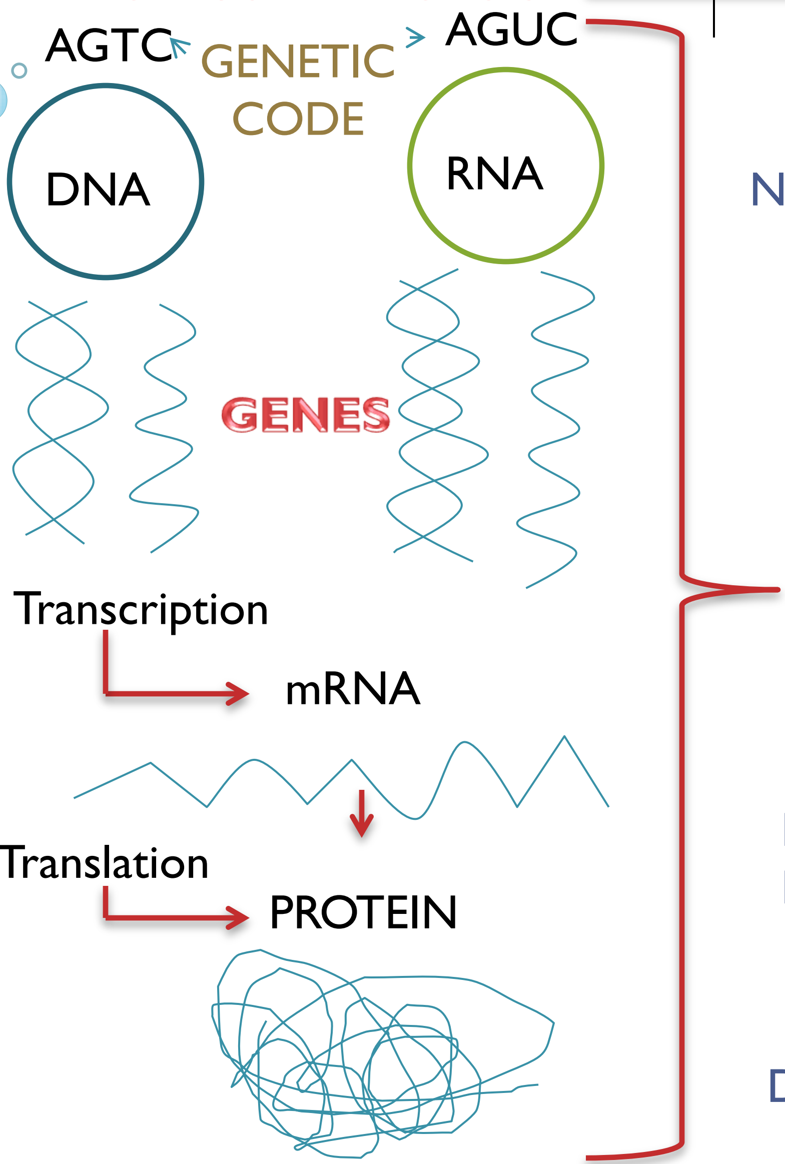
&

APPLICATIONS

AUTHOR: EUGENE MADZOKERE
Bachelor of Science Honors Degree in Biotechnology Student (2013)
Chinhoyi University of Technology (CUT)
Zimbabwe
Correspondence: madzokere@gmail.com

SUMMARY: Bioinformatics Principles & Applications

MOLECULAR BIOLOGY — INFORMATION TECHNOLOGY



AGTC

GENETIC CODE

AGUC

DNA

RNA

GENES

Transcription

mRNA

Translation

PROTEIN

HIGH-THROUGHPUT SCREENING (HTS)

↓

NEXT-GENERATION SEQUENCING (NGS)

↓

SEQUENCE ALIGNMENT, ANALYSIS & INTERPRETATION

↓

GENE FINDING

↓

GENOME MAPPING & ASSEMBLY

↓

PROTEIN STRUCTURE ALIGNMENT & PREDICTION

↓

PREDICTION OF GENE EXPRESSION & PROTEIN-PROTEIN INTERACTIONS

↓

EVOLUTION MODELLING

↓

DRUG DISCOVERY & DEVELOPMENT OF BIOPHARMACEUTICALS

Bioinformatics Defined!

- *Bioinformatics is the acquisition, application of computational tools, storage, arrangement, identification, archiving, analysis, interpretation, visualization and communication approaches for expanding the use of biological, medical, behavioral or health data in molecular biology research development.*
- The term "bioinformatics" is short for "*biological informatics*".
- *Antony Kerlavage* of Celera Genomics defined bioinformatics as "*Any application of computation to the field of biology including data management, algorithm development, and data mining*".

Bioinformatics Defined!

- Simply put, Bioinformatics is "*the application of information technology to the field of molecular biology*".
- As a research field, *Bioinformatics* entails:

1. *Creation & advancement of databases, algorithms, computational & statistical techniques* and;
2. *Creation and advancement of theory to solve formal and practical problems arising from the management and analysis of biological data*.

Bioinformatics Defined!

- Bioinformatics *strives to further our knowledge* of *biological systems* and *capacity to interpret biological processes* for *utilization in different applications*.
- This is evidenced by it's development and use of computationally intensive techniques
- That said, common activities include:

I. *Mapping & analyzing DNA, RNA, Protein, Amino Acid, & Lipid sequences.*

II. *Sequence Alignment & Analysis.*

III. *Creation and Visualization of 3-D structure models for biological molecules of significance e.g. protein.*

IV. *Genome Annotation.*

Biοinformatics Defined!

- Critical research areas include:

1. *Sequence Alignment, Annotation, Analysis & Interpretation;*
2. *Gene Finding/Identification & Synthesis;*
3. *Genome Mapping and Assembly;*
4. *Protein Structure Alignment and Prediction;*
5. *Prediction of Gene Expression & Protein-Protein Interactions;*
6. *Evolution Modeling &;*
7. *Drug Discovery and Development of Biopharmaceuticals.*

- These are dependent largely on high-throughput screening (HTS), characterization, expression and next generation sequencing (NGS) technologies.

Bioinformatics Databases Defined!

- Interest in Bioinformatics was propelled by the necessity to create databases of biological sequences.
- The first database was created shortly after the *Insulin protein sequence* was made available in 1956.
- To store and manage efficiently the large amounts of data generated in a genomic and molecular research era, *Information Technology* (IT) has been & is being used to develop & improve *Bioinformatics/Biological Databases* (*BD*).
- A *BD* *is a large, organized body of persistent data, usually associated with computerized software designed to update, query, and retrieve components of the data stored within the system.*
- The *Sequence BD*'s are amongst the most synonymous.

Bioinformatics Databases Defined!

- More specifically, a *Biological database is a structured collection of information (biological) consisting of basic units called "records"* or *"entries".*
- *Each record consists of fields holding predefined data related to the record.*
- The simplest database might be a single file containing many records, each of which includes the same set of information.
- E.g., A record associated with a *Nucleotide Sequence Database* typically contains information such as:

1. *A contact name;*
2. *The input sequence with a description of the type of molecule;*
3. *The scientific name of the source organism from which it was isolated; and, often,*
4. *Literature citations associated with the sequence.*

- *In such a case, nucleotide sequences would represent the record/entry, whilst information such as the base count and origin, sequence length, etc, constitutes the field!*
- Each *Database* has a *summary or checksum line.*

Purpose of Bioinformatics Databases!

- *Databases are designed to collect, archive, visualize and organize data to enable intelligent data description/interpretation, discovery, retrieval & invocation.*
- *Databases exist as a way of ensuring interoperability and integration between different research institutes, research databases, data mining tools, soft-wares and ordinary end-user with minimal restriction.*

Querying Bioinformatics Databases!

- A database *query is "a method to retrieve information from the database"*.
- *Organization of database records into predetermined fields, enables end-users to query on fields*.
- Database querying is made easier by algorithms.

Purpose of Bioinformatics Algorithms!

- An *Algorithm* is "*a soft-ware programme designed to improve sequence based database query (search) by increasing the speed, precision, accuracy and efficiency of identifying and making sense of similarities and/or dissimilarities from sequence alignments.*"
- *Examples of algorithms include:*

1. *Hidden Markov Models;*
2. *Smith-Waterman algorithm & the;*
3. *Needleman-Wunch algorithm.*

Advances in Bioinformatics Algorithm Design!

- However, although some algorithms are highly sensitive and increase accuracy of searches, they still take more time to execute the search.
- For this reason, algorithms are now being developed around *biotechnologists* and *bioinformaticians* which permits:

1. *Integration of diverse data and tools under a common Graphic User Interface (GUI).*
2. *Sharing information &;*
3. *Creation of powerful solutions useful in data archiving.*

Value of Sequence Alignment Process!

- Sequence Alignment informs us on the:

1. Function or activity of a new gene/protein.
2. Structure or shape of a new protein.
3. Location or preferred position of a protein.
4. Stability of a gene or protein.
5. Origin of a gene or protein.
6. Origin or phylogeny of an organelle.
7. Origin or phylogeny of an organism.

Bioinformatics Databases: Maintainer Status!

- Data submitted for storage in a database must be:

1. *Easy to access & extract to answer a specific biological question/research area*.

- Two forms of database maintainer status exist, namely:

1. *Public repositories:* have no legal restrictions & offer free public data access and retrieval, however databases are often riddled with redundancies because of limited error checking, curation, database updating and lack of strict data submission rules.
2. *Private repositories:* have higher quality data & legal restrictions attached to copyrights, patents, and often access is only available upon some form of monetary payment to data managing companies or their agents through a server network.

- *Data in Private repositories* is regularly curated and updated and end users follow strict data submission rules which limits redundancies and errors.
- *"A databases maintainer status thus directly influences the quality, access, dynamic nature, heterogeneity & type of data submitted for storage".*

Bioinformatics Database Characteristics!

- The Fundamental bioinformatics database characteristics include:

1. *Hierarchical data organization:* refers to data ranging from molecules, molecular pathways, cells, tissues, to organisms and populations.
2. *Complex data type*: describes data existing in databases as text-based sequences, blobs, images of cells and tissues, three dimensional molecular structures and complex data structural biochemical pathways.
3. *Dynamic nature*: describes the nature of data content and the resultant constant changes in the database schema.
4. *Quality:* describe the integrity of constraints within databases, the need to curate and update data constantly.
5. *Heterogeneous content*: most bioinformatics databases are heterogeneous in their content and may even have common semantic, database size, location, and syntactic differences (such as storage format or the access method).
6. *Accessibility*: describes the necessity for internet access, a search/browsing facility, flexibility to support external analysis tools and federation.

SEQUENCE DATABASES DEFINED!

- A <u>*Sequence Database*</u> (*SD*) *is a large collection of computerized (digital) nucleic acid sequences, protein sequences and/or other sequences stored on a computer.*
- *SD's* are an organized way of storing and managing copious loads of sequence information accumulating worldwide.
- Annotation in *SD's* is "*the process of adding* <u>*biological information*</u> *&* <u>*predictions*</u> *to a sequenced framework*".
- Without this annotation, *genome, nucleic acid, contig, gene &/or protein sequences* are virtually useless to bioinformatics & molecular biology research development.

SEQUENCE DATABASES DEFINED!

- Sequence Databases are classified as:

1. *Genome sequence databases.*
2. *Nucleic acid sequence databases.*
3. *Protein sequence databases.*
4. *Amino acid sequence databases.*

- *SD's* also fall into three database categories:

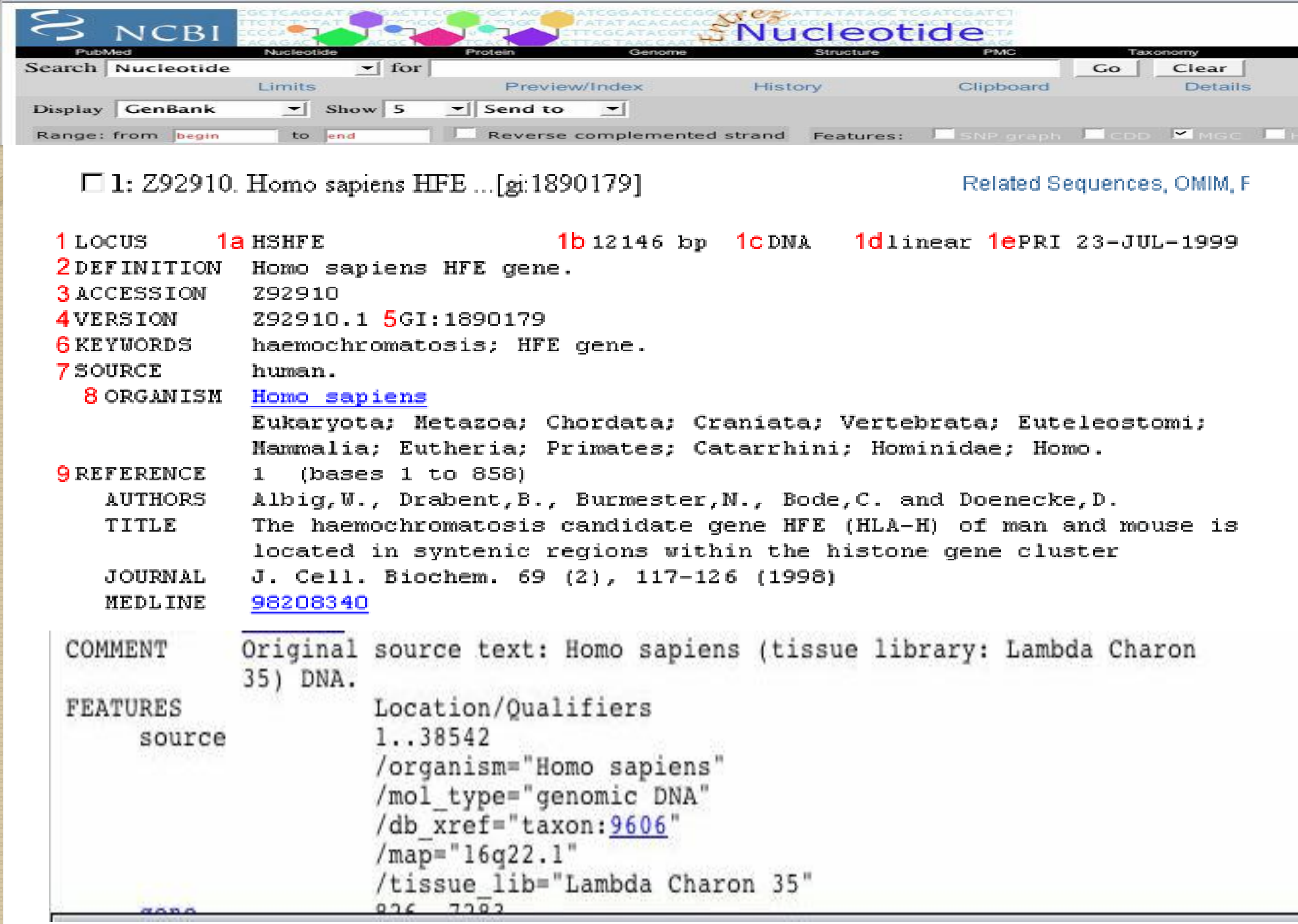
1. *Primary databases;*
2. *Secondary databases;*
3. *Composite databases.*

FUNDAMENTAL ELEMENTS OF SEQUENCE DATABASES

- All of the following elements represent the "*ideal minimal content of annotation entry in a Sequence Database*"

1. *Name :LOCUS, ENTRY, ID all unique identifiers*
2. *Definition: A brief, one-line, textual sequence description.*
3. *Accession:* A constant data identifier.
4. *Version*
5. *Gene identifier (GI)*
6. *Comments & Keywords*
7. *Source*
8. *Organism & Taxonomy Information*
9. *Literature References*
10. *Features table*
11. *Base count & Origin*
12. *And the Sequence itself!!!*

FUNDAMENTAL ELEMENTS OF SEQUENCE DATABASES



□ 1: Z92910. Homo sapiens HFE ...[gi:1890179]

```
1 LOCUS      1a HSHFE          1b 12146 bp  1c DNA   1d linear  1e PRI 23-JUL-1999
2 DEFINITION    Homo sapiens HFE gene.
3 ACCESSION     Z92910
4 VERSION       Z92910.1 5 GI:1890179
6 KEYWORDS      haemochromatosis; HFE gene.
7 SOURCE        human.
  8 ORGANISM    Homo sapiens
                Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
                Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
9 REFERENCE     1  (bases 1 to 858)
    AUTHORS     Albig,W., Drabent,B., Burmester,N., Bode,C. and Doenecke,D.
    TITLE       The haemochromatosis candidate gene HFE (HLA-H) of man and mouse is
                located in syntenic regions within the histone gene cluster
    JOURNAL     J. Cell. Biochem. 69 (2), 117-126 (1998)
    MEDLINE     98208340
COMMENT     Original source text: Homo sapiens (tissue library: Lambda Charon
            35) DNA.
FEATURES             Location/Qualifiers
     source          1..38542
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /map="16q22.1"
                     /tissue_lib="Lambda Charon 35"
     gene            926..7283
```

FUNDAMENTAL ELEMENTS OF SEQUENCE DATABASES

GenBank and GenPept format —

```
•    LOCUS       HSEF1AR               1506 bp    mRNA    linear   PRI 12-SEP-1993
•    DEFINITION  Human mRNA for elongation factor 1 alpha subunit (EF-1 alpha).
•    ACCESSION   X03558
•    VERSION     X03558.1  GI:31097
•    KEYWORDS    elongation factor; elongation factor 1.
•    SOURCE      human.
•      ORGANISM  Homo sapiens  Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
•    REFERENCE   1  (bases 1 to 1506)
•      AUTHORS   Brands,J.H., Maassen,J.A., van Hemert,F.J., Amons,R. and Moller,W.
•      TITLE     The primary structure of the alpha subunit of human elongation……
•      JOURNAL   Eur. J. Biochem. 155 (1), 167-171 (1986)
•      MEDLINE   86136120
•    FEATURES            Location/Qualifiers
       source          1..1506
                /organism="Homo sapiens"
                /db_xref="taxon:9606"
  CDS            54..1442
                /note="EF-1 alpha (aa 1-463)"
                /codon_start=1
                /protein_id="CAA27245.1"
                /db_xref="GI:31098"
                /db_xref="SWISS-PROT:P04720"
                /translation="MGKEKTHINIVVIGHVDSGKSTTTGHLIYKCGGIDKRTIEKFEK
                EAAEMGKGSFKYAWVLDKLKAERERGITIDISLWKFETSKYYVTIIDAPGHRDFIKNM
                ……VTKSAQKAQKAK"
•    BASE COUNT      412 a    337 c    387 g    370 t
•    ORIGIN
      1 acgggtttgc cgccagaaca caggtgtcgt gaaaactacc cctaaaagcc aaaatgggaa
     61 aggaaaagac tcatatcaac attgtcgtca ttggacacgt agattcgggc aagtccacca……….
   1501 aactgt
//
```

FUNDAMENTAL ELEMENTS OF SEQUENCE DATABASES

1. The **LOCUS field:** It consists of five different subfields, namely:

- ***1a Locus Name* (e.g. HSHFE) -** It is a tag for grouping similar sequences. The first two or three letters usually designate the organism. In this case ***HS*** stands for *Homo sapiens*. *The last several characters are associated with another group* designation, such as gene product. In this example, the last three digits represent the gene symbol, *HFE*. Currently, the only requirement for assigning a locus name to a record is that it is unique.
- ***1b Sequence Length* (12146 *bp*) –** It is the total number of nucleotide base pairs (or amino acid residues) in the sequence record.
- ***1c Molecule Type*** (e.g. ***DNA***) - Type of molecule that was sequenced. All sequence data in an entry must be of the same type.
- ***1d GenBank Division* (*PRI*) -** GenBank has different divisions**.** In this example, *PRI* stands for *primate sequences*. Other divisions include *ROD* (*rodent sequences*), *MAM* (*other mammal sequences*), *PLN* (*plant, fungal, and algal sequences*), & *BCT* (*bacterial sequences*).
- ***1e Modification Date* (23-July-1999) -** Date of most recent modification made to the record. The date of first public release is not available in the sequence record. This information can be obtained only by contacting NCBI at info@ncbi.nlm.nih.gov.

FUNDAMENTAL ELEMENTS OF SEQUENCE DATABASES

2. DEFINITION: – *It is a brief description of the sequence.*

- The description may include *source organism name, gene or protein name, or designation as untranscribed or untranslated sequences* (e.g., *a promoter region*).
- For sequences containing a *coding region* (*CDS*), the definition field may also contain a "*completeness*" *qualifier* such as "*complete CDS*" or "*exon 1*."

FUNDAMENTAL ELEMENTS OF SEQUENCE DATABASES

3. *ACCESSION (Z92910)*: *– It is a unique identifier assigned to a complete sequence record.*

- This number never changes, even if the record is modified.
- An "accession number" *is a combination of letters and numbers that are usually in the format of one letter followed by five digits* (e.g., *M12345*) *or two letters followed by six digits* (e.g., *AC123456*).

FUNDAMENTAL ELEMENTS OF SEQUENCE DATABASES

4. *VERSION* (*Z92910.1*) – *It is an identification number assigned to a single, specific sequence in the database.*

- This number is in the format "*accession.version.*"
- If any changes are made to the sequence data, the version part of the number will increase by one.
- E.g. *U12345.1* becomes *U12345.2*.
- A version number of *Z92910.1* for this *HFE* sequence indicates that the sequence data has not been altered thus it is an original submission.

FUNDAMENTAL ELEMENTS OF SEQUENCE DATABASES

5. Gene Identifier (GI) (*1890179*) - Also a sequence identification number.

- Whenever a sequence is changed, the version number is increased and a new *GI* is assigned.
- If a nucleotide sequence record contains a protein translation of the sequence, the translation will have its own *GI* number.

FUNDAMENTAL ELEMENTS OF SEQUENCE DATABASES

6. *KEYWORDS* (*haemochromatosis; HFE gene*) -

- A "*keyword*" can be "*any word or phrase used to describe the sequence*".
- Keywords are not taken from a controlled vocabulary. Notice that in this record the keyword, "*haemochromatosis*," employs British spelling, rather than the American "*hemochromatosis*."
- Many records have no keywords.
- A period is placed in this field for records without keywords.

FUNDAMENTAL ELEMENTS OF SEQUENCE DATABASES

7. *SOURCE* (*human*) - *Usually contains an abbreviated or common name of the source organism*.

8. *ORGANISM* (*Homo sapiens*) - *The scientific name (usually genus & species) & phylogenetic lineage*.

- Refer to the *NCBI Taxonomy Homepage* for more information about the classification scheme used to construct taxonomic lineages.

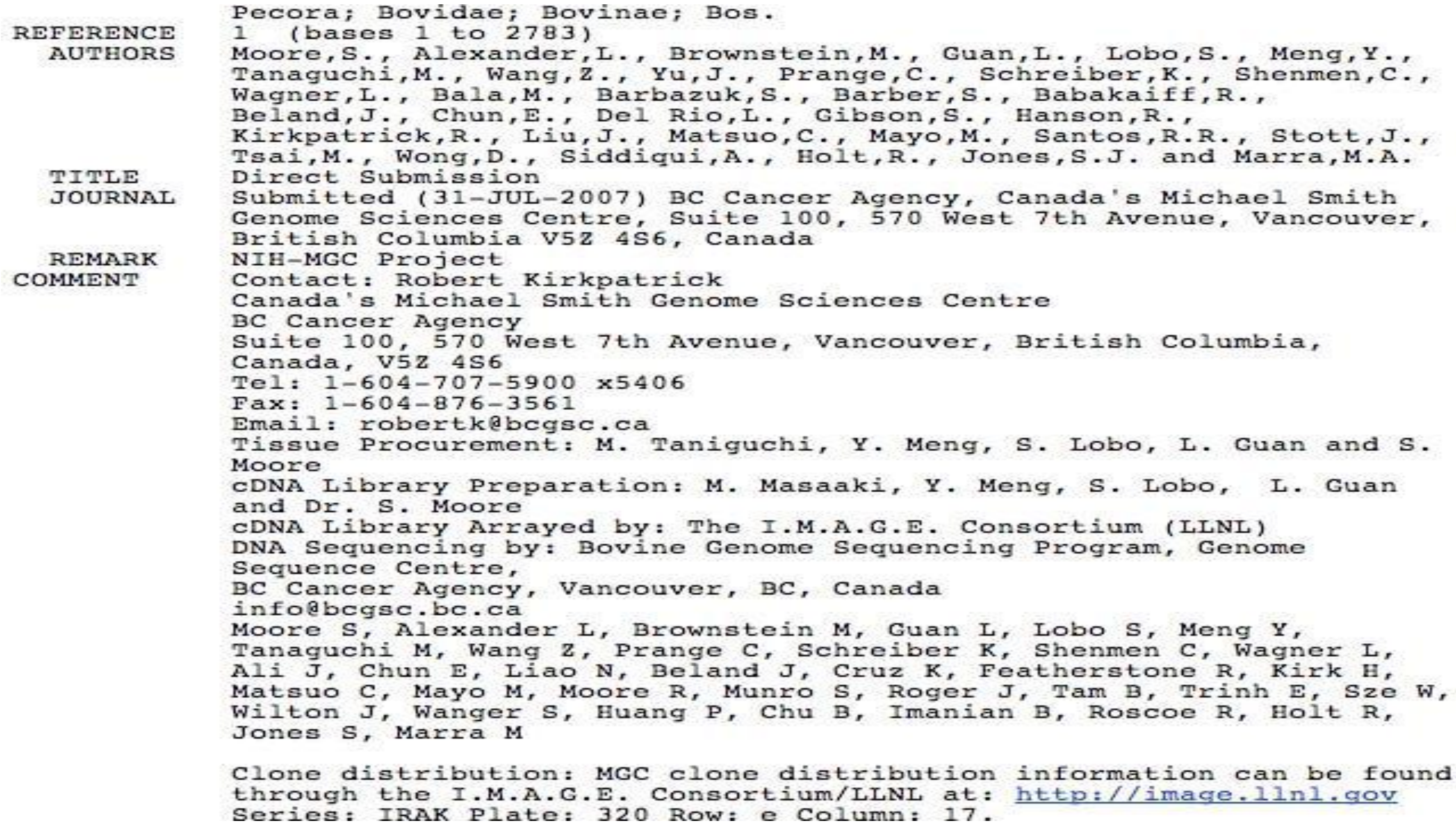
FUNDAMENTAL ELEMENTS OF SEQUENCE DATABASES

9. *REFERENCE – It is a citation of publications by sequence authors that supports information presented in the sequence record.*

- Several references may be included in one record.
- References are automatically sorted from the oldest to the newest.
- Cited publications are searchable by author, article or publication title, journal title, or *MEDLINE unique identifier* (*UID*).
- The *UID* links the sequence record to the *MEDLINE* record.

FUNDAMENTAL ELEMENTS OF SEQUENCE DATABASES

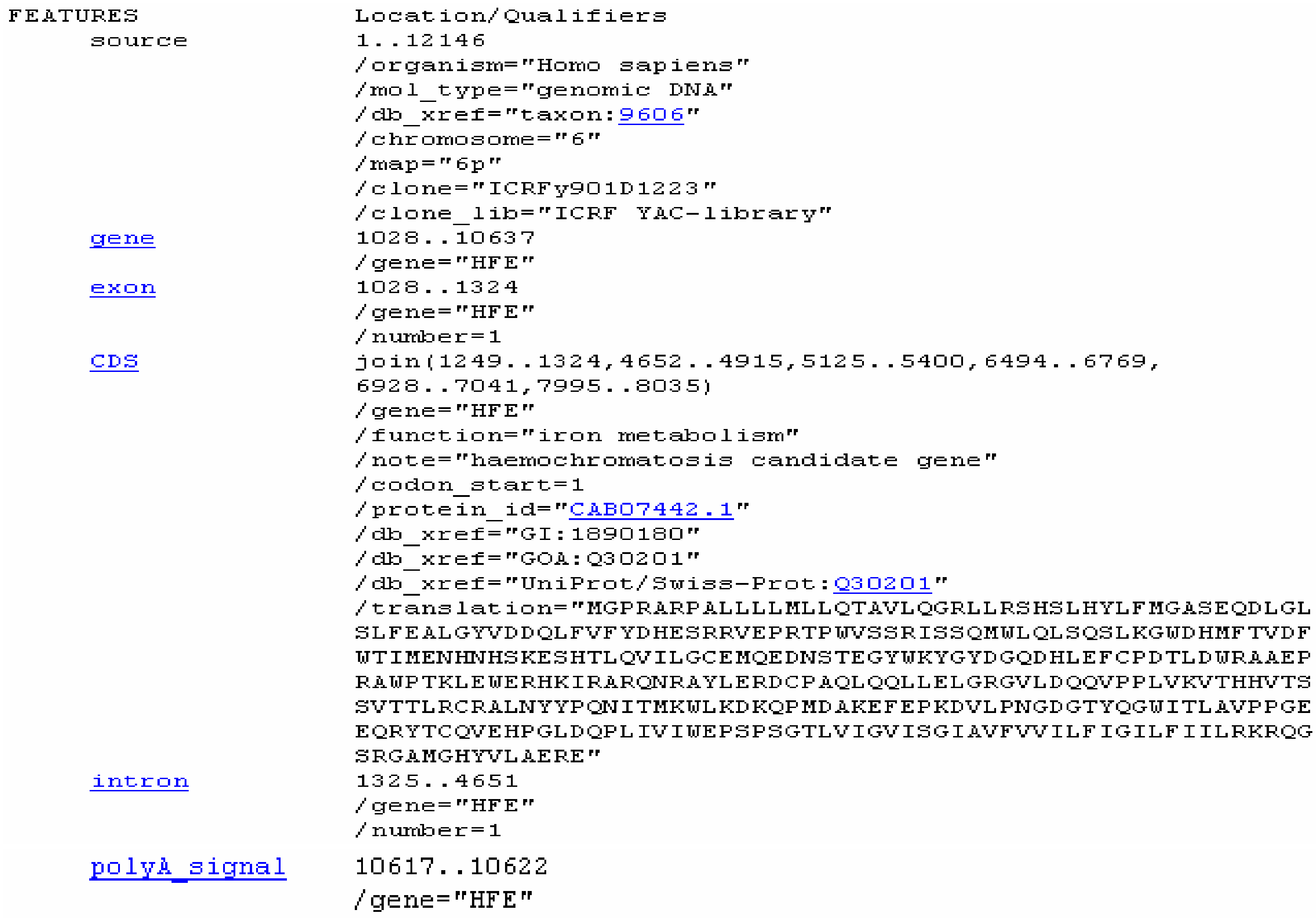
•When the *REFERENCE TITLE* contains the words "*Direct Submission*“, contact information for the submitter(s) is provided.



```
            Pecora; Bovidae; Bovinae; Bos.
REFERENCE   1  (bases 1 to 2783)
  AUTHORS   Moore,S., Alexander,L., Brownstein,M., Guan,L., Lobo,S., Meng,Y.,
            Tanaguchi,M., Wang,Z., Yu,J., Prange,C., Schreiber,K., Shenmen,C.,
            Wagner,L., Bala,M., Barbazuk,S., Barber,S., Babakaiff,R.,
            Beland,J., Chun,E., Del Rio,L., Gibson,S., Hanson,R.,
            Kirkpatrick,R., Liu,J., Matsuo,C., Mayo,M., Santos,R.R., Stott,J.,
            Tsai,M., Wong,D., Siddiqui,A., Holt,R., Jones,S.J. and Marra,M.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (31-JUL-2007) BC Cancer Agency, Canada's Michael Smith
            Genome Sciences Centre, Suite 100, 570 West 7th Avenue, Vancouver,
            British Columbia V5Z 4S6, Canada
  REMARK    NIH-MGC Project
COMMENT     Contact: Robert Kirkpatrick
            Canada's Michael Smith Genome Sciences Centre
            BC Cancer Agency
            Suite 100, 570 West 7th Avenue, Vancouver, British Columbia,
            Canada, V5Z 4S6
            Tel: 1-604-707-5900 x5406
            Fax: 1-604-876-3561
            Email: robertk@bcgsc.ca
            Tissue Procurement: M. Taniguchi, Y. Meng, S. Lobo, L. Guan and S.
            Moore
            cDNA Library Preparation: M. Masaaki, Y. Meng, S. Lobo,  L. Guan
            and Dr. S. Moore
            cDNA Library Arrayed by: The I.M.A.G.E. Consortium (LLNL)
            DNA Sequencing by: Bovine Genome Sequencing Program, Genome
            Sequence Centre,
            BC Cancer Agency, Vancouver, BC, Canada
            info@bcgsc.bc.ca
            Moore S, Alexander L, Brownstein M, Guan L, Lobo S, Meng Y,
            Tanaguchi M, Wang Z, Prange C, Schreiber K, Shenmen C, Wagner L,
            Ali J, Chun E, Liao N, Beland J, Cruz K, Featherstone R, Kirk H,
            Matsuo C, Mayo M, Moore R, Munro S, Roger J, Tam B, Trinh E, Sze W,
            Wilton J, Wanger S, Huang P, Chu B, Imanian B, Roscoe R, Holt R,
            Jones S, Marra M

            Clone distribution: MGC clone distribution information can be found
            through the I.M.A.G.E. Consortium/LLNL at: http://image.llnl.gov
            Series: IRAK Plate: 320 Row: e Column: 17.
```

10. The FEATURES Table:



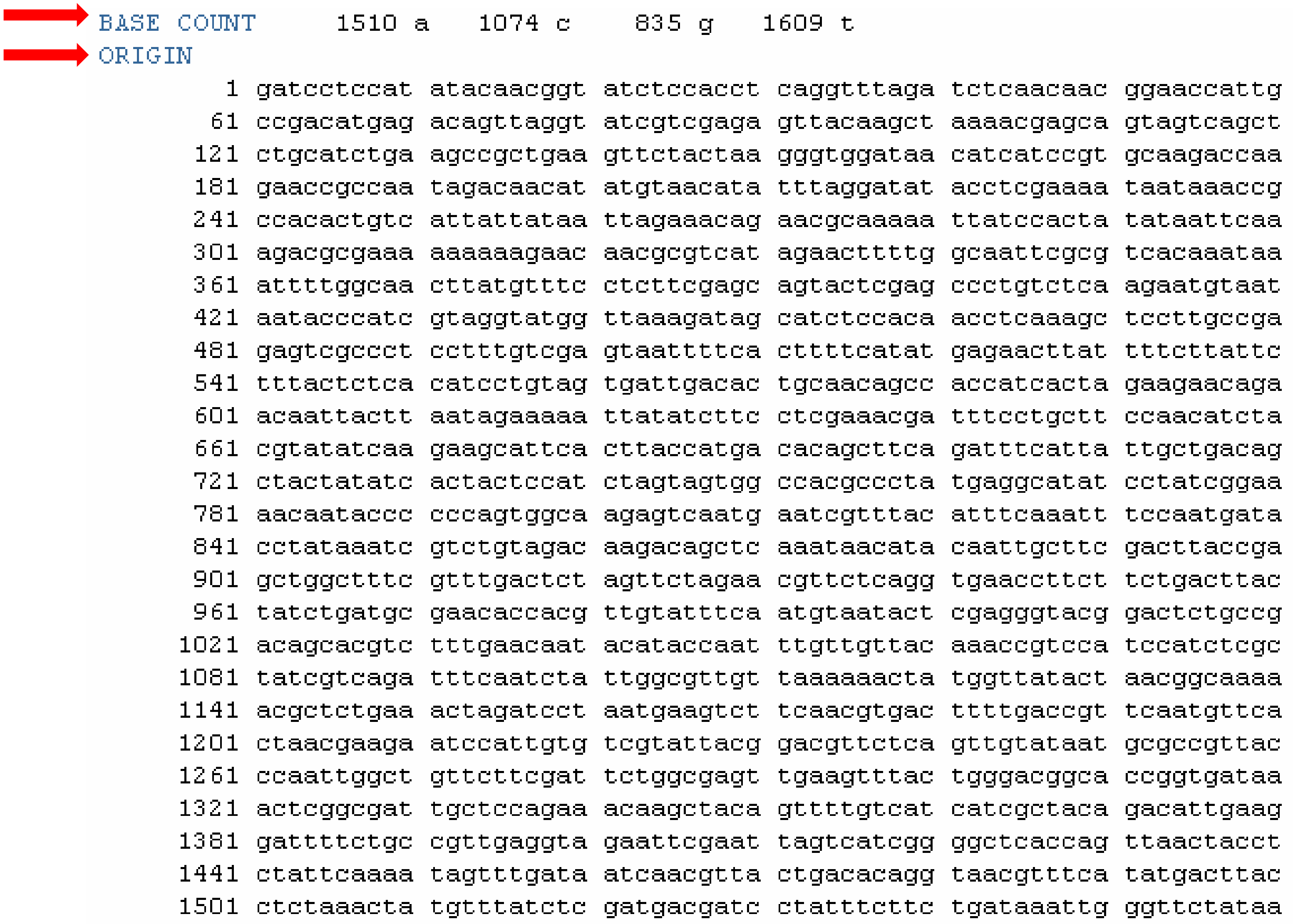
```
FEATURES             Location/Qualifiers
     source          1..12146
                     /organism="Homo sapiens"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:9606"
                     /chromosome="6"
                     /map="6p"
                     /clone="ICRFy901D1223"
                     /clone_lib="ICRF YAC-library"
     gene            1028..10637
                     /gene="HFE"
     exon            1028..1324
                     /gene="HFE"
                     /number=1
     CDS             join(1249..1324,4652..4915,5125..5400,6494..6769,
                     6928..7041,7995..8035)
                     /gene="HFE"
                     /function="iron metabolism"
                     /note="haemochromatosis candidate gene"
                     /codon_start=1
                     /protein_id="CAB07442.1"
                     /db_xref="GI:1890180"
                     /db_xref="GOA:Q30201"
                     /db_xref="UniProt/Swiss-Prot:Q30201"
                     /translation="MGPRARPALLLLMLLQTAVLQGRLLRSHSLHYLFMGASEQDLGL
                     SLFEALGYVDDQLFVFYDHESRRVEPRTPWVSSRISSQMWLQLSQSLKGWDHMFTVDF
                     WTIMENHNHSKESHTLQVILGCEMQEDNSTEGYWKYGYDGQDHLEFCPDTLDWRAAEP
                     RAWPTKLEWERHKIRARQNRAYLERDCPAQLQQLLELGRGVLDQQVPPLVKVTHHVTS
                     SVTTLRCRALNYYPQNITMKWLKDKQPMDAKEFEPKDVLPNGDGTYQGWITLAVPPGE
                     EQRYTCQVEHPGLDQPLIVIWEPSPSGTLVIGVISGIAVFVVILFIGILFIILRKRQG
                     SRGAMGHYVLAERE"
     intron          1325..4651
                     /gene="HFE"
                     /number=1
     polyA_signal    10617..10622
                     /gene="HFE"
```

FUNDAMENTAL ELEMENTS OF SEQUENCE DATABASES

11. BASE COUNT & ORIGIN:

- BASE COUNT - *Base Count gives the total number of adenine (A), cytosine (C), guanine (G), and thymine (T) bases in the sequence.*
- ORIGIN - *Origin contains the sequence data, which begins on the line immediately below the field title.*

FUNDAMENTAL ELEMENTS OF SEQUENCE DATABASES



```
BASE COUNT      1510 a    1074 c     835 g    1609 t
ORIGIN
        1 gatcctccat atacaacggt atctccacct caggtttaga tctcaacaac ggaaccattg
       61 ccgacatgag acagttaggt atcgtcgaga gttacaagct aaaacgagca gtagtcagct
      121 ctgcatctga agccgctgaa gttctactaa gggtggataa catcatccgt gcaagaccaa
      181 gaaccgccaa tagacaacat atgtaacata tttaggatat acctcgaaaa taataaaccg
      241 ccacactgtc attattataa ttagaaacag aacgcaaaaa ttatccacta tataattcaa
      301 agacgcgaaa aaaaaagaac aacgcgtcat agaacttttg gcaattcgcg tcacaaataa
      361 attttggcaa cttatgtttc ctcttcgagc agtactcgag ccctgtctca agaatgtaat
      421 aatacccatc gtaggtatgg ttaaagatag catctccaca acctcaaagc tccttgccga
      481 gagtcgccct cctttgtcga gtaattttca cttttcatat gagaacttat tttcttattc
      541 tttactctca catcctgtag tgattgacac tgcaacagcc accatcacta gaagaacaga
      601 acaattactt aatagaaaaa ttatatcttc ctcgaaacga tttcctgctt ccaacatcta
      661 cgtatatcaa gaagcattca cttaccatga cacagcttca gatttcatta ttgctgacag
      721 ctactatatc actactccat ctagtagtgg ccacgcccta tgaggcatat cctatcggaa
      781 aacaataccc cccagtggca agagtcaatg aatcgtttac atttcaaatt tccaatgata
      841 cctataaatc gtctgtagac aagacagctc aaataacata caattgcttc gacttaccga
      901 gctggctttc gtttgactct agttctagaa cgttctcagg tgaaccttct tctgacttac
      961 tatctgatgc gaacaccacg ttgtatttca atgtaatact cgagggtacg gactctgccg
     1021 acagcacgtc tttgaacaat acataccaat ttgttgttac aaaccgtcca tccatctcgc
     1081 tatcgtcaga tttcaatcta ttggcgttgt taaaaaacta tggttatact aacggcaaaa
     1141 acgctctgaa actagatcct aatgaagtct tcaacgtgac ttttgaccgt tcaatgttca
     1201 ctaacgaaga atccattgtg tcgtattacg gacgttctca gttgtataat gcgccgttac
     1261 ccaattggct gttcttcgat tctggcgagt tgaagtttac tgggacggca ccggtgataa
     1321 actcggcgat tgctccagaa acaagctaca gttttgtcat catcgctaca gacattgaag
     1381 gattttctgc cgttgaggta gaattcgaat tagtcatcgg ggctcaccag ttaactacct
     1441 ctattcaaaa tagtttgata atcaacgtta ctgacacagg taacgtttca tatgacttac
     1501 ctctaaacta tgtttatctc gatgacgatc ctatttcttc tgataaattg ggttctataa
```

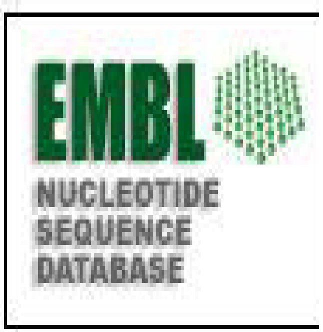
NUCLEIC ACID SEQUENCE DATABASES (NASD)

- *NASD's* are *repositories that accept nucleic acid sequence data and avail it for public use.*
- They hold heterogeneous (meaning the source of material is either genomic &/or cDNA), and either partially, completely or un-annotated nucleic acid sequence data.

TYPES OF NUCLEIC ACID DATABASES

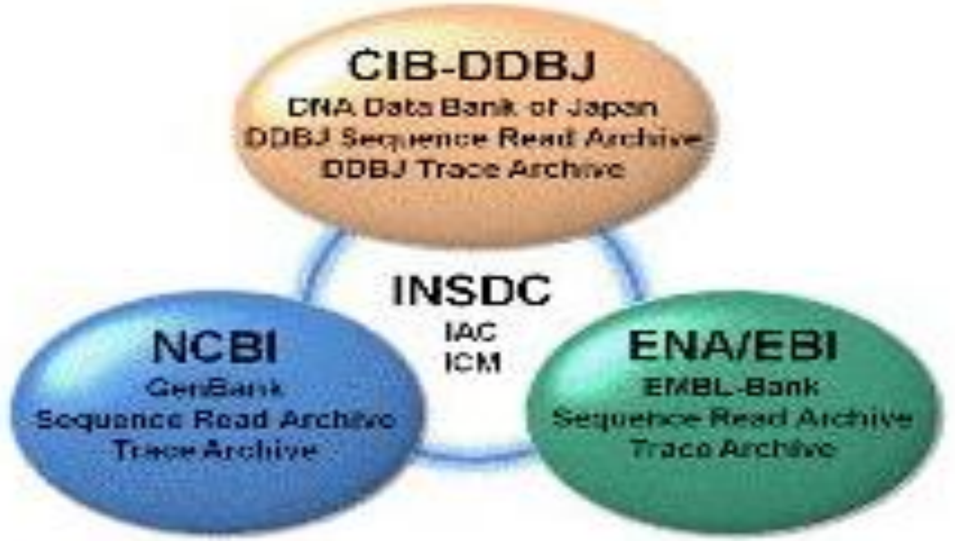
I. PRIMARY NUCLEIC ACID DATABASES

- Contain complete annotations of all the nucleic acid sequence information of organisms whose genomes have been successfully sequenced.
- Examples include *GenBank*, *DDBJ* and *EMBL*.



International Nucleotide Sequence Database Collaboration (INSDC)

- These 3 combined make-up the *International Nucleotide Sequence Database Collaboration (INSDC)*.



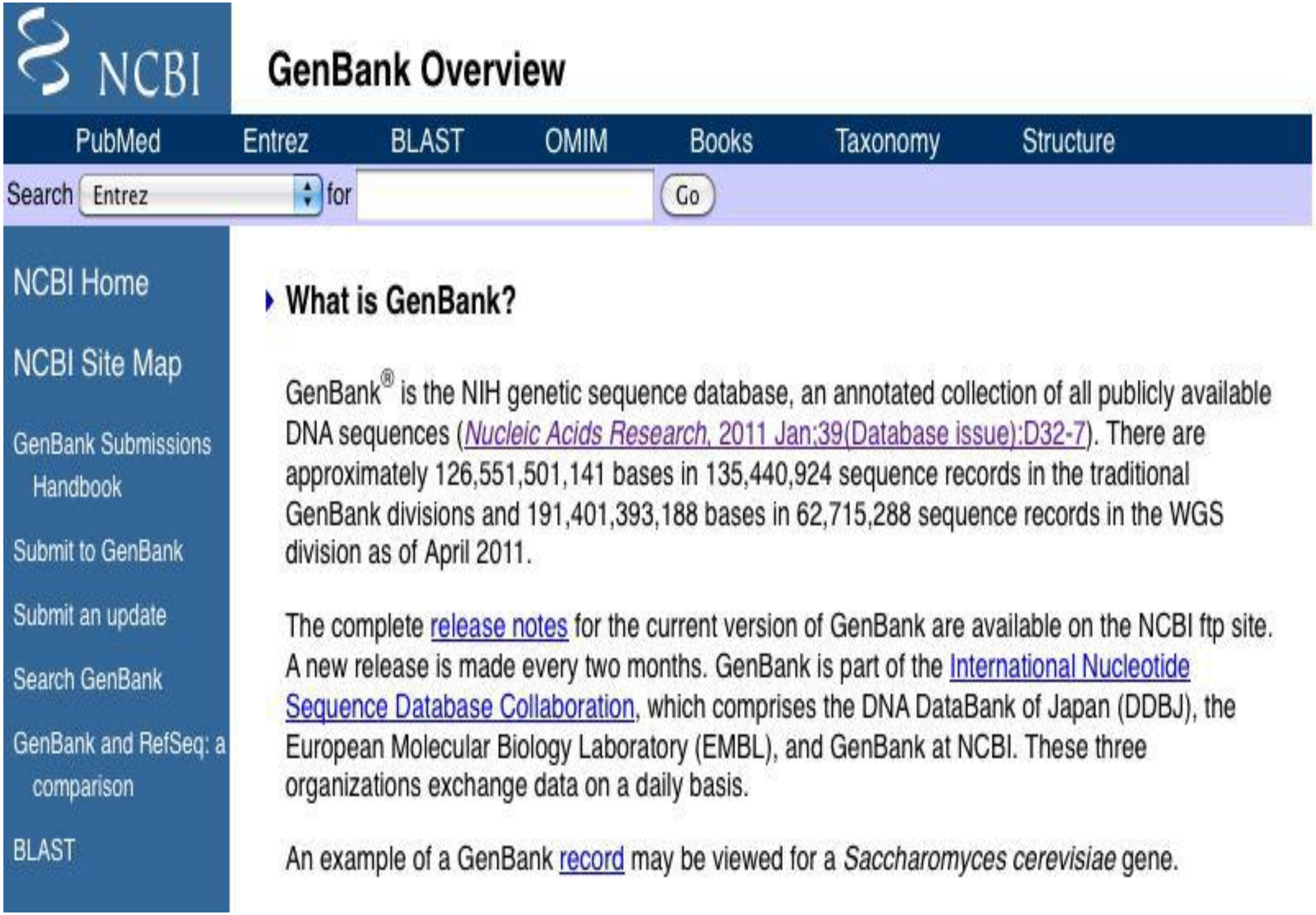
International Nucleotide Sequence Database Collaboration

International Nucleotide Sequence Database Collaboration (INSDC)

- *INSDC* is a synchronization of *GenBank*, *DDBJ* and *EMBL* databases done daily.
- Properties of *INSDC* include:

1. Consistent Accession numbers;
2. No legal restrictions. Although there are some patented sequences stored and managed.
3. Holds both sequences submitted directly by scientists and genome sequencing groups & sequences taken from literature & patents.
4. Has very limited error checking thus there is a fair amount of redundancy.
5. Access is provided via ftp & www interfaces; &
6. Sequences are listed in the 5’-3’ orientation.

[A] GenBank



GenBank Overview

What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Nucleic Acids Research, 2011 Jan;39(Database issue):D32-7). There are approximately 126,551,501,141 bases in 135,440,924 sequence records in the traditional GenBank divisions and 191,401,393,188 bases in 62,715,288 sequence records in the WGS division as of April 2011.

The complete release notes for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

An example of a GenBank record may be viewed for a *Saccharomyces cerevisiae* gene.

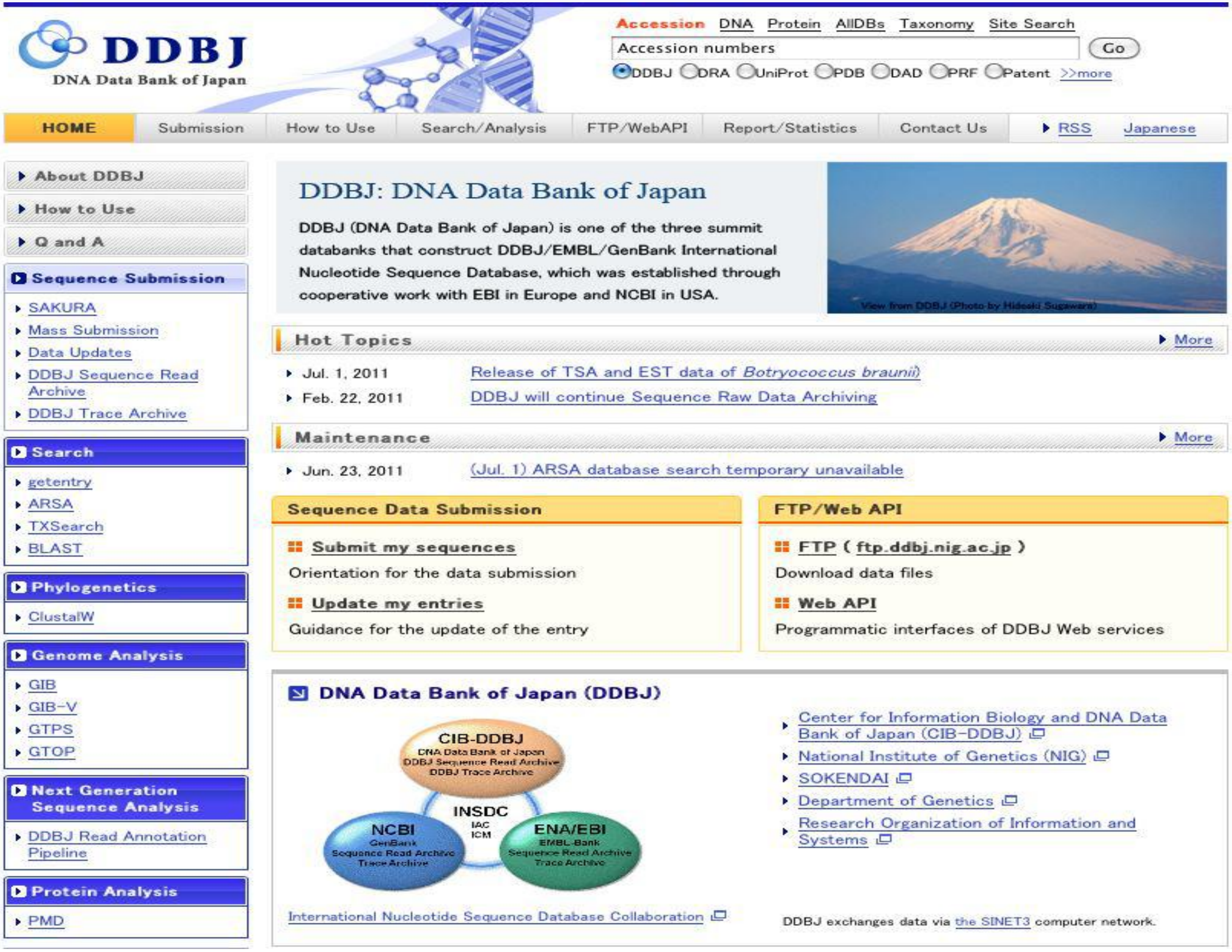
[A] GenBank

- *GenBank* (*Genetic Sequence Databank*) is one of the fastest growing repositories of known genetic sequences.
- It has a flat file structure that is an ASCII text file, readable & downloadable by both humans and computers.
- It is maintained by the National Center for Biotechnology (NCB).
- Entry data contains information on:

1. *The sequence;*
2. *Accession numbers;*
3. *The scientific and gene names;*
4. *Taxonomy/phylogenetic classification of the source organism;*
5. *A feature that identifies coding regions;*
6. *References to published literature;*
7. *Transcription units &;*
8. *Mutation sites.*

- There are approximately 286,730,369,256 sequence records in the traditional GenBank divisions as of 2011.

[B] DNA Data Bank of Japan (DDBJ)



DDBJ
DNA Data Bank of Japan

Accession DNA Protein AllDBs Taxonomy Site Search

Accession numbers Go

● DDBJ ○ DRA ○ UniProt ○ PDB ○ DAD ○ PRF ○ Patent >>more

HOME | Submission | How to Use | Search/Analysis | FTP/WebAPI | Report/Statistics | Contact Us | ▸ RSS | Japanese

- About DDBJ
- How to Use
- Q and A

Sequence Submission

- SAKURA
- Mass Submission
- Data Updates
- DDBJ Sequence Read Archive
- DDBJ Trace Archive

Search

- getentry
- ARSA
- TXSearch
- BLAST

Phylogenetics

- ClustalW

Genome Analysis

- GIB
- GIB-V
- GTPS
- GTOP

Next Generation Sequence Analysis

- DDBJ Read Annotation Pipeline

Protein Analysis

- PMD

DDBJ: DNA Data Bank of Japan

DDBJ (DNA Data Bank of Japan) is one of the three summit databanks that construct DDBJ/EMBL/GenBank International Nucleotide Sequence Database, which was established through cooperative work with EBI in Europe and NCBI in USA.

View from DDBJ (Photo by Hideaki Sugawara)

Hot Topics ▸ More

- Jul. 1, 2011 Release of TSA and EST data of *Botryococcus braunii*)
- Feb. 22, 2011 DDBJ will continue Sequence Raw Data Archiving

Maintenance ▸ More

- Jun. 23, 2011 (Jul. 1) ARSA database search temporary unavailable

Sequence Data Submission

Submit my sequences
Orientation for the data submission

Update my entries
Guidance for the update of the entry

FTP/Web API

FTP (ftp.ddbj.nig.ac.jp)
Download data files

Web API
Programmatic interfaces of DDBJ Web services

DNA Data Bank of Japan (DDBJ)

- Center for Information Biology and DNA Data Bank of Japan (CIB-DDBJ)
- National Institute of Genetics (NIG)
- SOKENDAI
- Department of Genetics
- Research Organization of Information and Systems

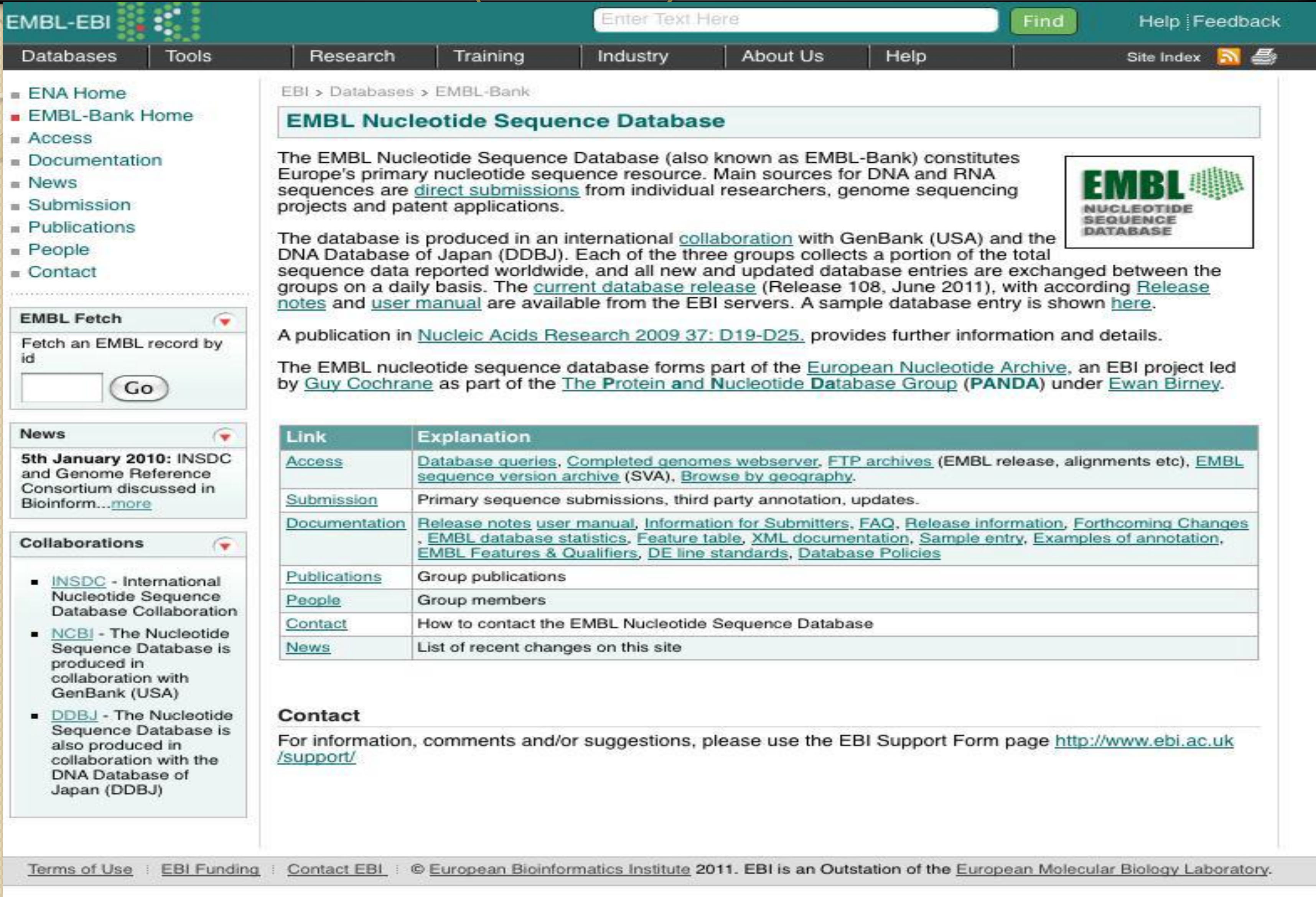
International Nucleotide Sequence Database Collaboration

DDBJ exchanges data via the SINET3 computer network.

[B] DNA Data Bank of Japan (DDBJ)

- Collects and supplies DNA data since its inception in 1986.
- Data entry as in *GenBank*.
- DDBJ exchanges data via the *SINET3* computer network.

[C] European Molecular Biology Laboratories (EMBL)



EMBL Fetch

Fetch an EMBL record by id

News

5th January 2010: INSDC and Genome Reference Consortium discussed in Bioinform...more

Collaborations

- INSDC - International Nucleotide Sequence Database Collaboration
- NCBI - The Nucleotide Sequence Database is produced in collaboration with GenBank (USA)
- DDBJ - The Nucleotide Sequence Database is also produced in collaboration with the DNA Database of Japan (DDBJ)

EMBL Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.

The database is produced in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. The current database release (Release 108, June 2011), with according Release notes and user manual are available from the EBI servers. A sample database entry is shown here.

A publication in Nucleic Acids Research 2009 37: D19-D25. provides further information and details.

The EMBL nucleotide sequence database forms part of the European Nucleotide Archive, an EBI project led by Guy Cochrane as part of the The **P**rotein **and** **N**ucleotide **D**atabase **G**roup (**PANDA**) under Ewan Birney.

Link	Explanation
Access	Database queries, Completed genomes webserver, FTP archives (EMBL release, alignments etc), EMBL sequence version archive (SVA), Browse by geography.
Submission	Primary sequence submissions, third party annotation, updates.
Documentation	Release notes user manual, Information for Submitters, FAQ, Release information, Forthcoming Changes , EMBL database statistics, Feature table, XML documentation, Sample entry, Examples of annotation, EMBL Features & Qualifiers, DE line standards, Database Policies
Publications	Group publications
People	Group members
Contact	How to contact the EMBL Nucleotide Sequence Database
News	List of recent changes on this site

Contact

For information, comments and/or suggestions, please use the EBI Support Form page http://www.ebi.ac.uk /support/

Terms of Use | EBI Funding | Contact EBI | © European Bioinformatics Institute 2011. EBI is an Outstation of the European Molecular Biology Laboratory.

[C] European Molecular Biology Laboratories (EMBL)

- It is a comprehensive database of DNA and RNA sequences collected from the scientific literature and patent applications and directly submitted from researchers and sequencing groups.
- Data collection is done in collaboration with *GenBank* (USA) and the *DNA Database of Japan* (*DDBJ*).
- It doubles in size every 18 months and as of June 1994 it contained nearly 2 million bases from 182,615 sequence entries.
- It is maintained by the *European Bioinformatics Institute* (*EBI*).
- Data entry is friendly both to computers and humans.
- Standard English used (explanations, descriptions *etc*).
- Sequences are stored in the database as they would occur in the biological state.

TYPES OF NUCLEIC ACID DATABASES

2. *SECONDARY NUCLEIC ACID DATABASES*

- They contain additional information derived from analysis of data available in primary repositories.
- They deal with particular classes of sequences.
- Examples include *UniGene*, the *HIV sequence database* and *REBASE*.

[A] *UniGene* SEQUENCE DATABASE

- *UniGene* has records with unique gene clusters.
- Each cluster contains: sequences that represent a unique gene and related information e.g tissue types in which the gene have been expressed.
- The database is populated with *Expressed Sequence Tags* (EST's).

[B] HIV SEQUENCE DATABASE

- The *HIV Sequence Database* (*HSD*) collects, curates & annotates HIV sequence data.

PROTEIN SEQUENCE DATABABES

- They consists of:

1. All the proteins that have been translated from the RNA sequences and;
2. Protein sequenced.

- Three (3) types of protein sequence databases exist, namely:

1. Primary protein databases;
2. Secondary protein databases and;
3. Composite protein databases.

PRIMARY PROTEIN SEQUENCE DATABASES

- Synonymous examples of primary protein sequence databases are:

1. *SWISS-PROT* &;
2. *PIR.*

- Both SWISS-PROT & PIR are curated.
- This means groups of designated curators (*database managers*) prepare the entries from literature and/ or contacts with external experts prior to submission into the respective databases.

[1] SWISS-PROT

- *Swiss-Prot* provides high level notations describing:

1. Functions of a protein;
2. Protein domain structure;
3. Post-translational modifications; &
4. Protein variants and other variables (Bairoch and Apweiler, 2000).

- It also provides a minimum level of redundancy & a high level of integration with other databases.
- It has legal restrictions in that entries are copyrighted, but freely accessible and usable by academic researchers.
- Commercial companies have to purchase a license from the Swiss Institute of Bioinformatics (SIB) to access the database.
- It classifies data into:

a. Core data (*Data - sequence references and taxonomic details)and;
b. The Annotation (*Annotation -sequence variants , functions, domains, secondary and quaternary structures and post translational modifications.).

[2] PROTEIN INFORMATION RESOURCE (PIR)

- It is a division of the *National Biomedical Research Foundation* (*NBRF*) in the US.
- It is a database that produces the *NRL-3D* (a database of sequences extracted from the three dimensional structures in the *Protein Databank* (*PDB*)).
- It's existence allows sequence information in *PDB* to be available for similarity searches & retrieval & provides cross reference information for use with other *PIR* Protein Sequence databases.
- It provides comprehensive, well organized, & accurate information about proteins such as sequence similarity.
- It also maintains the *PSD*, *NREF* & the *iProClass* to support researchers to understand genomic and proteomic research.

SECONDARY PROTEIN SEQUENCE DATABASES

- Major examples of Secondary protein sequence databases are: *TrEMBL, SP-TrEMBL, REM-TrEMBL and SPRT (SWALL), Prosite* and *Pfam*.

[I] TrEMBL:

- *TrEMBL* stands for Translation of EMBL nucleotide sequence database.
- It is a computer-annotated supplement of *SWISS-PROT.*
- It contains all translations of *EMBL* nucleotide sequence entries not yet integrated in *SWISS-PROT*.
- *TrEMBL* speeds new sequence information to the public.

SECONDARY PROTEIN SEQUENCE DATABASES

[2] SP-TrEMBL:

- It focuses on entries to be incorporated later into Swiss-Prot.

[3] REM-TrEMBL:

- It contains other data that will not be integrated because it may be redundant or are truncated or are not proteins or are fragments legitimately translated in-vivo.

[4] SPRT (SWALL):

- It provides data by focusing on data currency in *Swiss-Prot*, ignoring *REM-TrEMBL*, and by performing sequence comparisons against a database of all known isoforms.

SECONDARY PROTEIN SEQUENCE DATABASES

- *[5] PROSITE:*
- It is a database of protein families and domains.
- It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs.
- It is part of and is maintained much like *Swiss-Prot*.
- It is based on regular expressions describing characteristic subsequences of specific protein families or domains.
- The *Ras GTPase - activating protein signature pattern* is an example of a *PROSITE* regular expression.

SECONDARY PROTEIN SEQUENCE DATABASES

[6] Pfam:

- It is a database of protein families defined as domains (contiguous segments of entire protein sequences).
- For each domain, it contains a multiple alignment of a set of defining sequences and the other sequences in Swiss-Prot and *TrEMBL* that can be matched to that alignment.
- Alignments can be converted into *Hidden Markov Models* (*HMM*), which can be used to search for domains in a query sequence.
- It can be searched and used to identify domains in sequence.
- It is licensed under the *GNU General Public License* making it available to anyone,
- However, *Pfam* imposes restrictions that derivative works (new databases, &/or modifications) must be made available in source form.

COMPOSITE DATABASES

- They compile and filter sequence data from different primary databases to produce combined non-redundant sets that are more complete than the individual databases.
- An example of a composite database is *OWL*.
- *OWL* combines 4 publicly available primary sources: *SWISS-PROT*, *PIR*, *GenBank* and *NRL-3D.*

SOFTWARE TOOLS FOR DATA MINING IN BIOINFORMATICS

- These range from simple command-line tools to more complex graphical programs and standalone web-services available from various bioinformatics companies and public institutions.
- The computational biology tool best-known among biologists is probably *BLAST*.
- *BLAST* is an algorithm for determining the similarity of arbitrary sequences against other sequences, possibly from curated databases of protein or DNA sequences.
- The NCBI provides a popular web-based implementation that searches their databases.
- BLAST is one of a number of generally available programs for doing sequence alignment.

WEB SERVICES IN BIOINFORMATICS

- A web service is *"a program/software that can be executed on a remote machine owning to the industry efforts to standardize web service description, discovery and invocation"*.
- These efforts have led to standards such as WSDL (Christenson *et al*, 2001), UDDI (UDDI2002).
- The European Bioinformatics Institute (EBI) has classified basic bioinformatics web services into three categories:

1. SSS (*Sequence Search Services*);
2. MSA (*Multiple Sequence Alignment*) and;
3. BSA (*Biological Sequence Analysis*).

WEB SERVICES IN BIOINFORMATICS

- Availability of these service-oriented bioinformatics resources demonstrates the applicability of web based bioinformatics solutions.
- The web services range from a collection of standalone tools with a common data format under a single, standalone or web-based interface, to integrative, distributed and extensible bioinformatics workflow management systems.

WEB SERVICES IN BIOINFORMATICS

- *SOAP* and *REST*-based interfaces have been developed for a wide variety of bioinformatics applications.
- *This allows an application running on one computer in one part of the world to use algorithms, data and computing resources on servers in other parts of the world.*
- *Also, end users do not have to deal with software and database maintenance overheads.*

Transition: Web-based tools to Web Services

- Before Web-services came into force, web-based tools were widely used to manage data.
- However, web based tools alone, faced several hindrances including:

1. *Applications were not language and platform independent;*
2. *Lack of machine friendly web interface.*
3. *Non-standard input and output data format of the web interfaces, application interface and message exchange protocol.*
4. *Transport protocol for the remote messaging were often not fire-wall friendly.*
5. *Lack of automated service description, discovery and integration.*

Advantages of WEB SERVICES

1. Eliminate the need to develop and rely on ad-hoc screen scrapping mechanism.
2. Offer a single uniform method for the application integration of through the internet.
3. Provide a model for web applications in which their public interfaces and bindings are defined and described using an XML standard format.
4. Use of XML-based messaging render the web services infrastructure platform-and language-independent and changes to the interface can immediately be detected by client software.

WEB SERVICE Architecture and Service Model IN BIOINFORMATICS

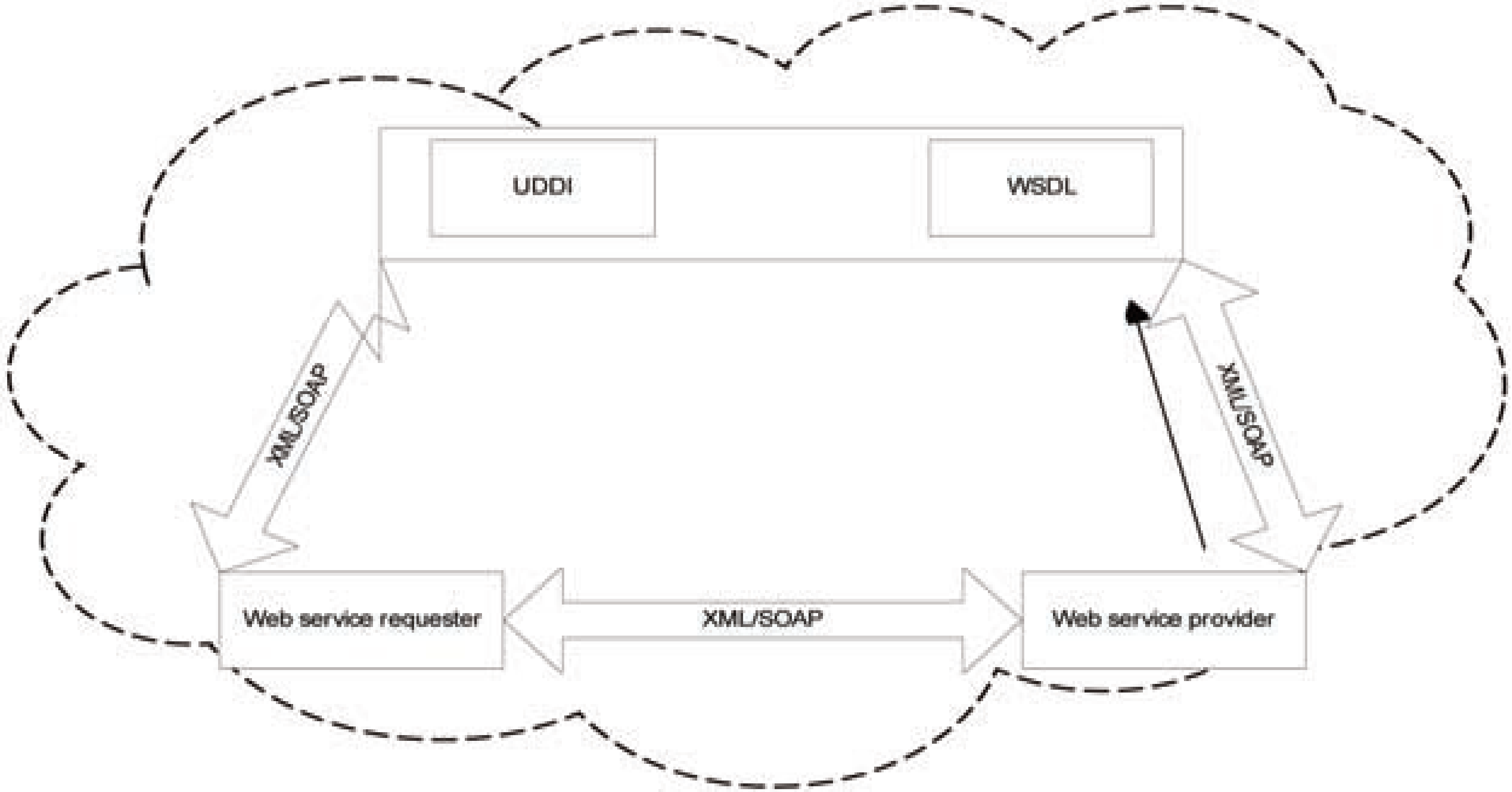


Figure 1.0 The basic profile of the interoperability model of Web Services (WS-I)

WEB SERVICES IN BIOINFORMATICS

1. The Web Service Description Language (WSDL) (http://www.w3.org/TR/wsdl) uses the XML standard format that describes a web service interface and the exchange of messages between the provider and requester in an abstract manner.

- *Service providers are generally specialized genome" centers such as National Center for Biotechnology Information (NCBI), European Bioinformatics Institute (EBI).*
- *Service consumers mostly are working in smaller laboratories and research groups with smaller, non-specialist resources.*

WEB SERVICES IN BIOINFORMATICS

2. *Simple Object Access Protocol (SOAP) is an XML-based protocol for the stateless message exchange which, in general, has been developed on the top of HTTP.*

- *This makes WS -I firewall friendly as opposed to the protocols used by (CORBA).*

WEB SERVICES IN BIOINFORMATICS

3. Universal Description, Discovery and Integration (UDDI) are a standard protocol designed to publish details about an organization and the web services.

- It provides a description and definition of web services in a central repository, which functions as yellow pages for web services.
- WSDL and SOAP are the W3C standards, while UDDI is an Organization for the Advancement of Structured Information Standards (OASIS) standard.
- For a client to use a web service it only needs WSDL with SOAP that is commonly being used as the default protocol.

Limitations of WEB SERVICES

- *The desired automation of the discovery, composition and invocation of web services and workflows is inhibited because:*

1. *UDDI search capabilities in its current form are limited to the keyword-based matching. It does not capture semantic relationships between entries in its directories.*

Limitations of WEB SERVICES

2. *UDDI supports search based on only the high-level information specified about businesses and services, i.e., the final state specification.*

- *The transitory and intermediate capabilities of the web service are not specified.*
- *However, UDDI service registrations may include references to the WSDL descriptions, which may facilitate the limited automation of the discovery and invocation.*
- *But, the absence of any explicit semantic information limits the automated comprehension of the WSDL description to simple ontologies in domains without contextual and conceptual differences.*

Limitations of WEB SERVICES

3. With the parameterized input invocation for filtering and delimiting the search domain is not available.

4. The search facilities in UDDI are restricted to exact matches because the search is syntax based, and thus discourages service composition and workflows.

5. Owning to the limitation of range imposed by non-semantic descriptions, not all WSDL documents describe the non-functional attributes such as authenticity, currency, efficiency, performance, scalability, etc. Even in the way WSDL+OWL-S, the mapping OWL-S into WSDL may lose much semantic information because WSDL can not express the abundance semantics of OWL.

Limitations of WEB SERVICES

6. Both the service providers and service consumers want to remain back-ward compatible to the legacy formats.

- The service consumers want their data in legacy formats so that the existing tools can operate over it.
- The service providers are wary of changing requirements of myriad of the existing data formats.
- Although this is not a serious problem for the simple data types, it has serious implications for most of the biological data which is highly complex and internally structured.

Limitations of WEB SERVICES

7. Scripts which are used to compose work flows are monolithic and complex and hence lack reusability.

Applications of BIOINFORMATICS

1. *GENOMICS:*

- *Estimating the number of genes in an organism basing on the number of nucleotide base pairs was not reliable, due to the presence of high numbers of redundant copies of many genes.*
- *Genomics has corrected this situation. Useful genes can be selected from a gene library thus constructed and inserted into other organisms for improvement or harmful genes can be silenced.*
- *In the areas of Structural genomics, Functional genomics and Nutritional genomics, bioinformatics plays a vital role.*

Applications of BIOINFORMATICS

a) Structural Genomics:- Focuses on large scale genome structure determination, gene identification and characterization.

b) Functional Genomics:- Focuses on predicting and identifying and characterization of genes and genomes based on function.

c) Nutritional Genomics:- Focuses on characterizing and inferring nutritional relevance to identified genes.

Applications of BIOINFORMATICS

2. PROTEOMICS:

- *Involves the sequencing of amino acids in a protein, determining its three dimensional structure and relating it to the function of the protein.*
- *Extensive data, generated through crystallography and NMR, are required for proteomic studies.*
- *With such data on known proteins, the structure and its relationship to function of newly discovered proteins can be understood in a very short time.*
- *In such areas, bioinformatics has an enormous analytical and predictive potential.*
- *It can help develop better understanding of how proteins fold and interact with one another and with other biological molecules which in turn will give scientists and doctors better insight into diseases and ways to combat them.*

Applications of BIOINFORMATICS

3. CHEMINFORMARTICS & DRUG DESIGN:

- *Cheminformatics involves organization of chemical data in a logical form to facilitate the process of understanding chemical properties, their relationship to structures and making inferences.*
- *In silico approaches now enable researchers:-*

1.To identify and structurally modify a natural product;

2.To design a drug with the desired properties and;

3.To assess its therapeutic effects, theoretically.

- *NB: [in silico (in the computer, based on silicon chip technology)].*
- *The risk involved in the earlier (in-vitro & in-vivo)random processes of drug discovery methods is largely removed by bioinformatics.*

Applications of BIOINFORMATICS

4. MOLECULAR PHYLOGENIES:

- Phylogeny is the origin and evolution of organisms.
- Biological systems of classifications for the known organisms, plants & animal included have been constructed.
- Amino acid sequences and characteristics of proteins are also used in systematics.
- Similarly, new work is arising in areas to do with modeling the phylogenetic evolution of microbes, viruses and sub-viral organisms as a way of understanding, amongst other things:

1. The effect of different patterns of severity on control of infectiousness;
2. Relationship between sub-type variation, infectivity and disease progression and control.
3. Factors triggering severity of infectiousness.

Applications of BIOINFORMATICS

5. DRUG MODIFICATION:

- Several synthetic products are quite useful but cannot be used by one and all for certain side effects in some people.
- E.g., Aspartame (marketed under different trade names) is a dipeptide of aspartic acid and phenylalanine, and is 300 times sweeter than cane sugar.
- Aspartame is widely used as an alternate sweetener by diabetics and others who cannot take sweeteners loaded with calories.
- Unfortunately, pregnant women and people suffering from phenylketonuria, a disorder due to an impaired metabolism of phenylalanine, should not use aspartame.
- It would be useful if phenylalanine were substituted by some other amino acid without affecting its sweetness, to remove the restriction on its use.

Applications of BIOINFORMATICS

- *The list of applications continues to grow daily and can be seen in new "OMICS" areas such as:*

1. *Metagenomics - Analysis and manipulation of microbial genomes without culturing.*
2. *Transcriptomics - Expression profiles of mRNA.*
3. *Metabolomics – Analysis, Modeling & Interpretation of Signaling & Metabolic Pathways.*
4. *Cellinomics - Analysis, Modeling & Interpretation of Cell-cell interactions.*

CONCLUSION

THE END