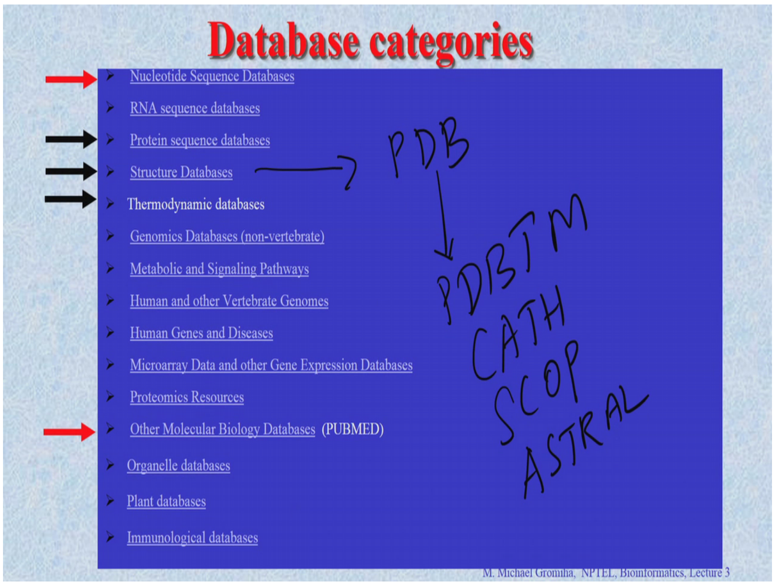
Bioinformatics
Prof. M. Michael Gromiha
Department of Biotechnology
Indian Institute of Technology, Madras

Lecture - 3b
Databases Categories

(Refer Slide Time: 00:17)



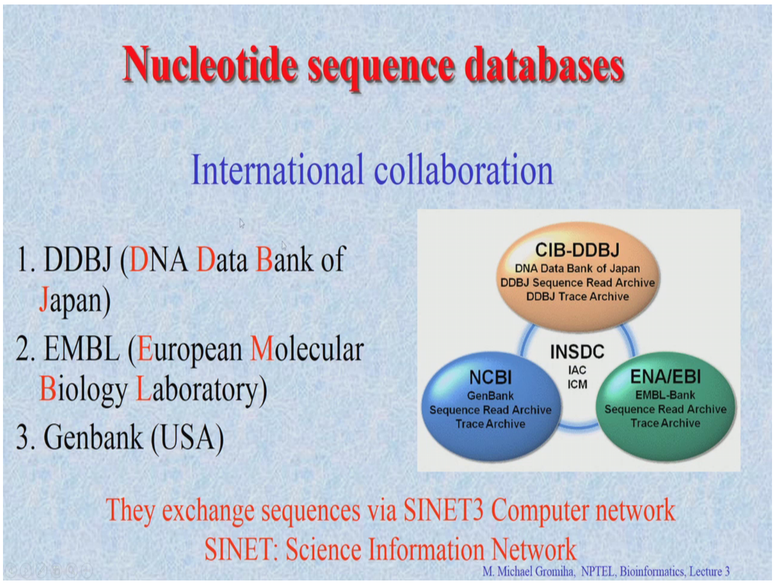
Database categories

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Thermodynamic databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases (PUBMED)
- Organelle databases
- Plant databases
- Immunological databases

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, this lecture while cover few databases mainly the DNA database and protein databases right. So, in the last class we discussed mainly about the DNA. So, I will move on to little bit about the DNA databases. So, it is very important to get the sequences of a nucleic acids right.

(Refer Slide Time: 00:41)



Nucleotide sequence databases

International collaboration

1. DDBJ (DNA Data Bank of Japan)
2. EMBL (European Molecular Biology Laboratory)
3. Genbank (USA)

They exchange sequences via SINET3 Computer network
SINET: Science Information Network

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

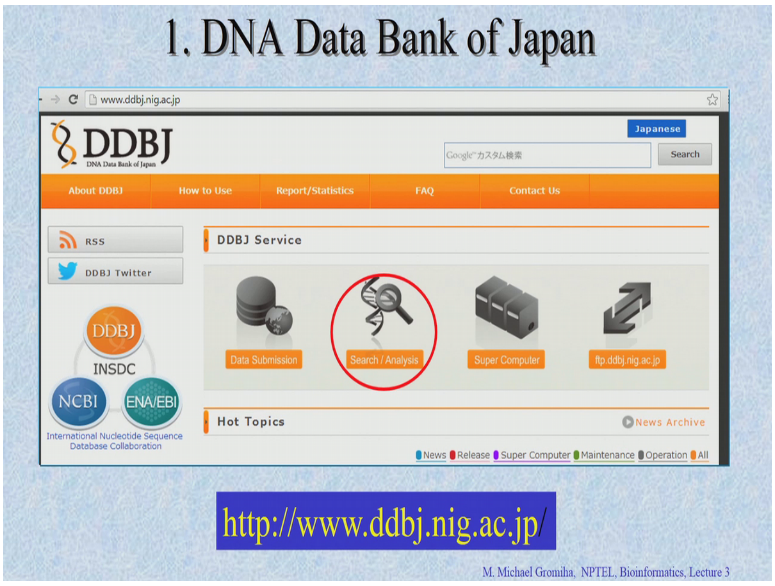
So, at the same time there are different types of databases have been developed for example. So, one is a DDBJ we say DNA data bank Japan right. So, this is in Japan and the second is EMBL, EMBL is the European molecular biology laboratory. So, they also developed this collecting nucleated data and GenBank USA they also sort out to collect.

Is the similar time because it is several people they think in a similar way right. We cannot say this is very unique that only we will do. Because science is open and research is open. So, it is very competitive right. If you have some information immidiately several people think in a same way, this is the reason we need to be very fast and we very accurate and we have to provide reliable data. So, they started collecting the data and then they also developed some tools to analyze the data right then what happened in this case? Each databases they have different types of data right. So, in this case there is some discrepancies. So, users have to access all the databases, that biggest sometimes you get the data from DDBJ may not be available EMBL.

Sometimes you get the data from EMBL may not be available in GenBank. So, for getting a data set, they have to check all the databases. Second aspect is the search options and the display options and the format may not be same if it is different then again we have to rework right. So, then (Refer Time: 02:09). So, they join together using these science information network. So, here is the DDBJ right. So, NCBI and we have the EMBL. So, joint together. So, they share the data. So, now, the data what we get from

DDBJ you get from EMBL or from the GenBank. So, wherever you use or access the databank database you will get all information right.

(Refer Slide Time: 02:34)



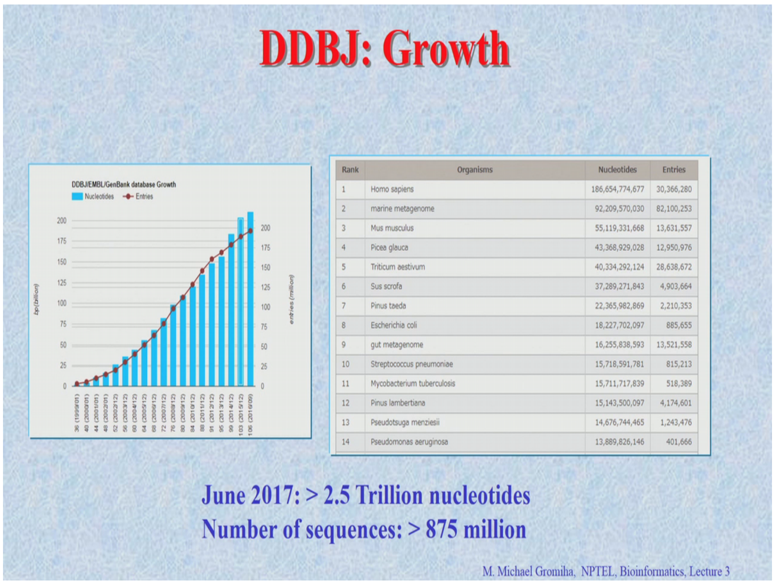
1. DNA Data Bank of Japan

http://www.ddbj.nig.ac.jp/

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, now will just demonstrative bit about these databases, DDBJ this is in Japan in (Refer Time: 02:42) Japan a beautiful city in Japan you can have this a (Refer Time: 02:45) and all fine. So, here they have a different options, you can submit the data and you can search and analyze the data and all they have the facilities and also you can FTP other information right. So, this is the website DDBJ dot nig that national institute of genetics this is the institute there maintaining the database and ac dot JP. So, you can access this website and you can get the data of the DNA sequences fine.

(Refer Slide Time: 03:14)



DDBJ: Growth

Rank	Organisms	Nucleotides	Entries
1	Homo sapiens	186,654,774,677	30,366,280
2	marine metagenome	92,209,570,030	82,100,253
3	Mus musculus	55,119,331,668	13,631,557
4	Picea glauca	43,368,929,028	12,950,976
5	Triticum aestivum	40,334,292,124	28,638,672
6	Sus scrofa	37,289,271,843	4,903,664
7	Pinus taeda	22,365,982,869	2,210,353
8	Escherichia coli	18,227,702,097	885,655
9	gut metagenome	16,255,838,593	13,521,558
10	Streptococcus pneumoniae	15,718,591,781	815,213
11	Mycobacterium tuberculosis	15,711,717,839	518,389
12	Pinus lambertiana	15,143,500,097	4,174,601
13	Pseudotsuga menziesii	14,676,744,465	1,243,476
14	Pseudomonas aeruginosa	13,889,826,146	401,666

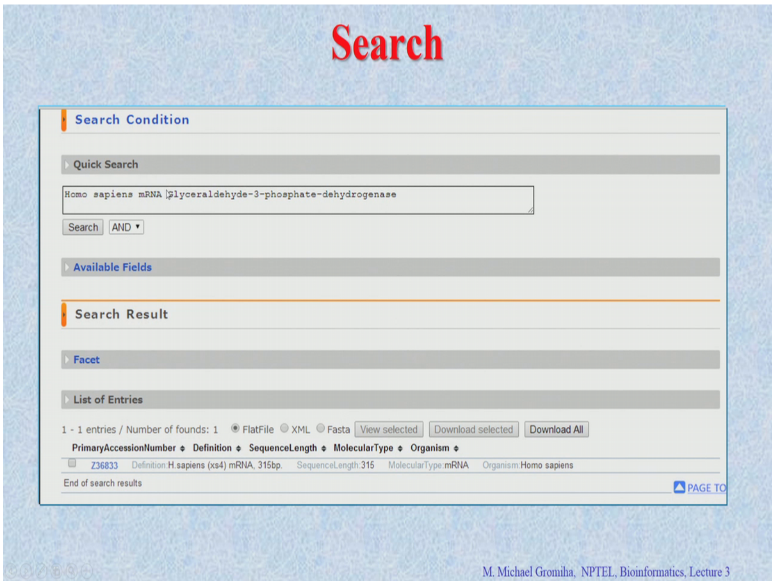
June 2017: > 2.5 Trillion nucleotides
Number of sequences: > 875 million

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, this is a growth get this is the started long term ago. So, it is in 1999. So, these the in January this is the release number 36, before that we started to its initiated database and it is a current status. So, this is 2016 106 release. So, this they we have the 2.5 trillion nucleotides as of June 2017 this let us update and about it; so it 875 million sequences. So, if a plenty of sequences for the analysis and we look into the organisms as we expected. So, we have the top most human and followed the mus musculus and then you can see call a other organisms. So, you can this is the top most organisms it the DDBJ in DDBJ fine. This is nucleotides this is the number of entries available for each organism right.

So now how to get the data; so there are various way is to get the data from DDBJ, the first two condition is the very simple search there is a quick search.

(Refer Slide Time: 04:15)



Search

Search Condition

Quick Search

```
Homo sapiens mRNA glyceraldehyde-3-phosphate-dehydrogenase
```

Search Result

1 - 1 entries / Number of founds: 1 ● FlatFile ○ XML ○ Fasta

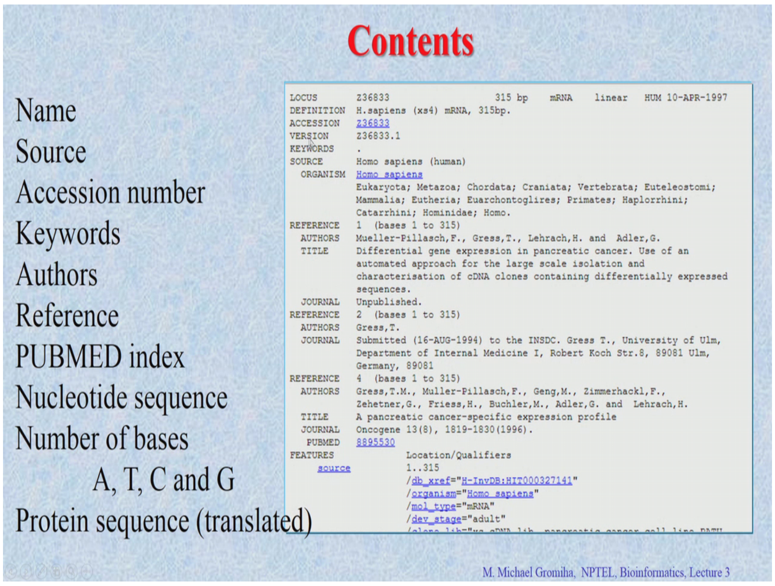
	PrimaryAccessionNumber ⬍	Definition ⬍	SequenceLength ⬍	MolecularType ⬍	Organism ⬍
□	Z36833	Definition:H.sapiens (xs4) mRNA, 315bp.	SequenceLength:315	MolecularType:mRNA	Organism:Homo sapiens

End of search results

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

If you want to get the data for any of these for your DNA or RNA; so here you put the homo sapiens mrna for this protein right. So, if you search right then you will get the link right. When you click on these accession number right, these are the 36833.

(Refer Slide Time: 04:33)



Contents

Name
Source
Accession number
Keywords
Authors
Reference
PUBMED index
Nucleotide sequence
Number of bases
A, T, C and G
Protein sequence (translated)

```
LOCUS       Z36833                   315 bp    mRNA    linear   HUM 10-APR-1997
DEFINITION  H.sapiens (xs4) mRNA, 315bp.
ACCESSION   Z36833
VERSION     Z36833.1
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 315)
  AUTHORS   Mueller-Pillasch,F., Gress,T., Lehrach,H. and  Adler,G.
  TITLE     Differential gene expression in pancreatic cancer. Use of an
            automated approach for the large scale isolation and
            characterisation of cDNA clones containing differentially expressed
            sequences.
  JOURNAL   Unpublished.
REFERENCE   2  (bases 1 to 315)
  AUTHORS   Gress,T.
  JOURNAL   Submitted (16-AUG-1994) to the INSDC. Gress T., University of Ulm,
            Department of Internal Medicine I, Robert Koch Str.8, 89081 Ulm,
            Germany, 89081
REFERENCE   4  (bases 1 to 315)
  AUTHORS   Gress,T.M., Muller-Pillasch,F., Geng,M., Zimmerhackl,F.,
            Zehetner,G., Friess,H., Buchler,M., Adler,G. and  Lehrach,H.
  TITLE     A pancreatic cancer-specific expression profile
  JOURNAL   Oncogene 13(8), 1819-1830(1996).
   PUBMED   8895530
FEATURES             Location/Qualifiers
     source          1..315
                     /db_xref="H-InvDB:HIT000327141"
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /dev_stage="adult"
```

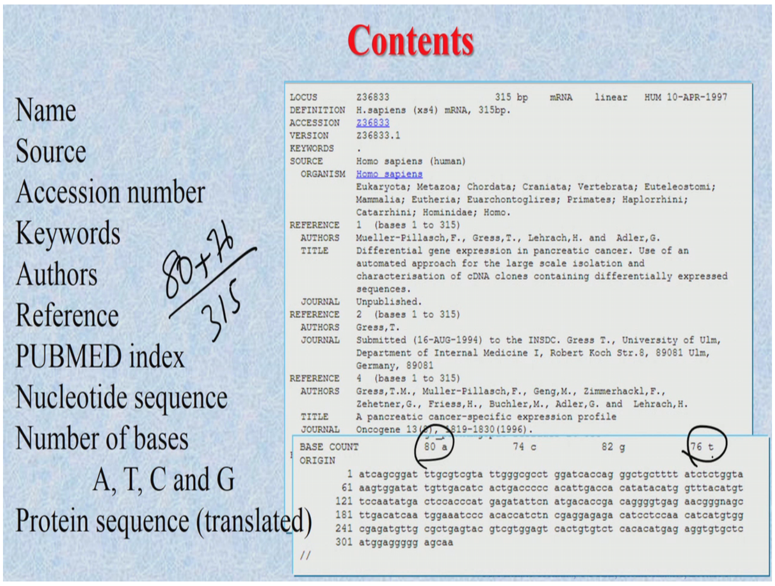
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

Then you get the complete data. So, now, you see; what are the contents, which are available in DDBJ. So, first you give the name right. So, I have gave some of the data. So, I can give the homo sapiens mrna three and 15 base pairs right and the source, source is homo sapiens and here they have given accession number specifically for this a DDBJ.

So, these accession number and (Refer Time: 04:56) keywords right. So, in this case you can search the data base on any specific keywords.

Right and the others who are though and who sequences data and the reference; you give the complete reference and you give the PUBMED index, they give the link to the PUBMED; PUBMED this literature database ok.

(Refer Slide Time: 05:18)



Contents

- Name
- Source
- Accession number
- Keywords
- Authors
- Reference
- PUBMED index
- Nucleotide sequence
- Number of bases
- A, T, C and G
- Protein sequence (translated)

$\frac{80+76}{315}$

```
LOCUS       Z36833                  315 bp    mRNA    linear   HUM 10-APR-1997
DEFINITION  H.sapiens (xs4) mRNA, 315bp.
ACCESSION   Z36833
VERSION     Z36833.1
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 315)
  AUTHORS   Mueller-Pillasch,F., Gress,T., Lehrach,H. and  Adler,G.
  TITLE     Differential gene expression in pancreatic cancer. Use of an
            automated approach for the large scale isolation and
            characterisation of cDNA clones containing differentially expressed
            sequences.
  JOURNAL   Unpublished.
REFERENCE   2  (bases 1 to 315)
  AUTHORS   Gress,T.
  JOURNAL   Submitted (16-AUG-1994) to the INSDC. Gress T., University of Ulm,
            Department of Internal Medicine I, Robert Koch Str.8, 89081 Ulm,
            Germany, 89081
REFERENCE   4  (bases 1 to 315)
  AUTHORS   Gress,T.M., Muller-Pillasch,F., Geng,M., Zimmerhackl,F.,
            Zehetner,G., Friess,H., Buchler,M., Adler,G. and  Lehrach,H.
  TITLE     A pancreatic cancer-specific expression profile
  JOURNAL   Oncogene 13(8), 1819-1830(1996).
BASE COUNT          80 a          74 c          82 g          76 t
ORIGIN
        1 atcagcggat ttgcgtcgta ttgggcgcct ggatcaccag ggctgctttt atctctggta
       61 aagtggatat tgttgacatc actgaccccc acattgacca catatacatg gtttacatgt
      121 tccaatatga ctccacccat gagatattcn atgacaccga caggggtgag aacgggnagc
      181 ttgacatcaa tggaaatccc acaccatctn cgaggagaga catcctccaa catcatgtgg
      241 cgagatgttg cgctgagtac gtcgtggagt cactgtgtct cacacatgag aggtgtgctc
      301 atggaggggg agcaa
//
```

That I will explain little bit later then and they give the data of number of as T C and G. So, here is the number of A, number of C, number of G number of T and this is the complete sequence right 315 10 15 base points right. So, in these sequence right. So, what is the AG; AG at contents, what is that AT contents?

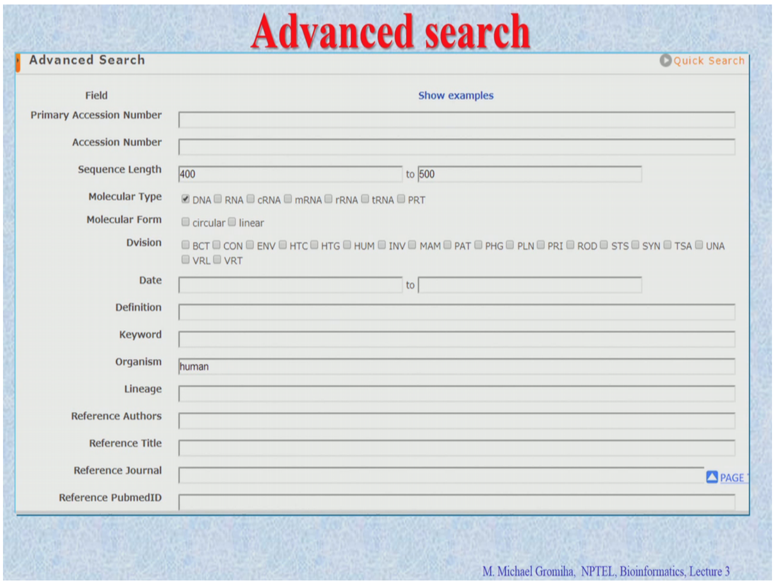
Student: (Refer Time: 05:38).

At plus 76 divided by 300 and.

Student: 15.

15 right (Refer Time: 05:51) calculate, this is number of as this is number of ts and you can calculate the a t conduct right. Same thing they give the translation protein sequence also right the fine.

(Refer Slide Time: 05:59)



Advanced search

Advanced Search

Field | Show examples

Primary Accession Number:

Accession Number:

Sequence Length: 400 to 500

Molecular Type: ☑ DNA □ RNA □ cRNA □ mRNA □ rRNA □ tRNA □ PRT

Molecular Form: □ circular □ linear

Dvision: □ BCT □ CON □ ENV □ HTC □ HTG □ HUM □ INV □ MAM □ PAT □ PHG □ PLN □ PRI □ ROD □ STS □ SYN □ TSA □ UNA □ VRL □ VRT

Date: to

Definition:

Keyword:

Organism: human

Lineage:

Reference Authors:

Reference Title:

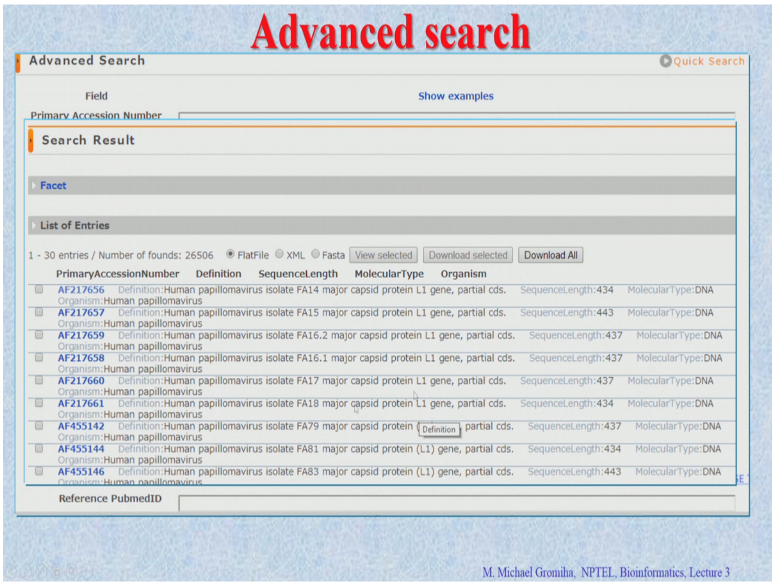
Reference Journal:

Reference PubmedID:

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, that one search; so they give lot of options right some of them are familiar, some of them are art familiar. So, in this case they give the ontology of each terms. Now if you click on this one you will get the details of right the specific term. So, here I put the sequence length 400 to 500 and orgs organism human. You can use any of these search options to get your sequence.

(Refer Slide Time: 06:25)



Advanced search

Advanced Search

Field | Show examples

Primary Accession Number

Search Result

1 - 30 entries / Number of founds: 26506 ● FlatFile ○ XML ○ Fasta [View selected] [Download selected] [Download All]

PrimaryAccessionNumber Definition SequenceLength MolecularType Organism

□ AF217656 Definition:Human papillomavirus isolate FA14 major capsid protein L1 gene, partial cds. SequenceLength:434 MolecularType:DNA Organism:Human papillomavirus

□ AF217657 Definition:Human papillomavirus isolate FA15 major capsid protein L1 gene, partial cds. SequenceLength:443 MolecularType:DNA Organism:Human papillomavirus

□ AF217659 Definition:Human papillomavirus isolate FA16.2 major capsid protein L1 gene, partial cds. SequenceLength:437 MolecularType:DNA Organism:Human papillomavirus

□ AF217658 Definition:Human papillomavirus isolate FA16.1 major capsid protein L1 gene, partial cds. SequenceLength:437 MolecularType:DNA Organism:Human papillomavirus

□ AF217660 Definition:Human papillomavirus isolate FA17 major capsid protein L1 gene, partial cds. SequenceLength:437 MolecularType:DNA Organism:Human papillomavirus

□ AF217661 Definition:Human papillomavirus isolate FA18 major capsid protein L1 gene, partial cds. SequenceLength:434 MolecularType:DNA Organism:Human papillomavirus

□ AF455142 Definition:Human papillomavirus isolate FA79 major capsid protein (Definition partial cds. SequenceLength:437 MolecularType:DNA Organism:Human papillomavirus

□ AF455144 Definition:Human papillomavirus isolate FA81 major capsid protein (L1) gene, partial cds. SequenceLength:434 MolecularType:DNA Organism:Human papillomavirus

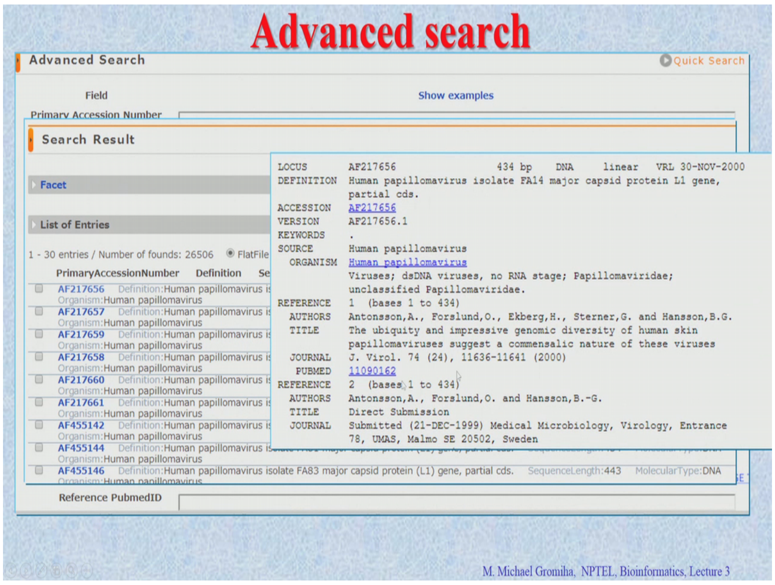
□ AF455146 Definition:Human papillomavirus isolate FA83 major capsid protein (L1) gene, partial cds. SequenceLength:443 MolecularType:DNA Organism:Human papillomavirus

Reference PubmedID

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

Now, if you give these are a list of sequences, there are so many error data for extremely that two twenty 26506 entries between 400 and 500 from human right.

(Refer Slide Time: 06:38)



Advanced search

1 - 30 entries / Number of founds: 26506

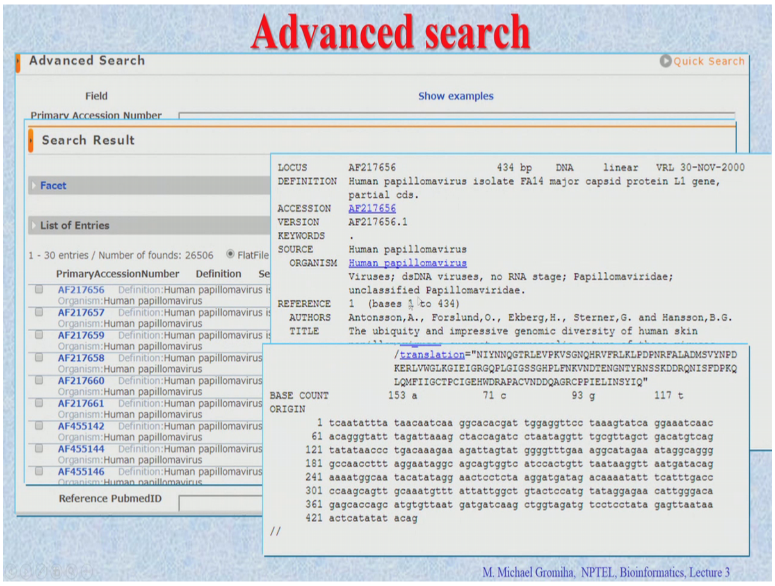
PrimaryAccessionNumber	Definition
AF217656	Definition:Human papillomavirus is Organism:Human papillomavirus
AF217657	Definition:Human papillomavirus is Organism:Human papillomavirus
AF217659	Definition:Human papillomavirus is Organism:Human papillomavirus
AF217658	Definition:Human papillomavirus is Organism:Human papillomavirus
AF217660	Definition:Human papillomavirus is Organism:Human papillomavirus
AF217661	Definition:Human papillomavirus is Organism:Human papillomavirus
AF455142	Definition:Human papillomavirus is Organism:Human papillomavirus
AF455144	Definition:Human papillomavirus is Organism:Human papillomavirus
AF455146	Definition:Human papillomavirus isolate FA83 major capsid protein (L1) gene, partial cds. SequenceLength:443 MolecularType:DNA Organism:Human papillomavirus

```
LOCUS       AF217656                 434 bp    DNA     linear   VRL 30-NOV-2000
DEFINITION  Human papillomavirus isolate FA14 major capsid protein L1 gene,
            partial cds.
ACCESSION   AF217656
VERSION     AF217656.1
KEYWORDS    .
SOURCE      Human papillomavirus
  ORGANISM  Human papillomavirus
            Viruses; dsDNA viruses, no RNA stage; Papillomaviridae;
            unclassified Papillomaviridae.
REFERENCE   1  (bases 1 to 434)
  AUTHORS   Antonsson,A., Forslund,O., Ekberg,H., Sterner,G. and Hansson,B.G.
  TITLE     The ubiquity and impressive genomic diversity of human skin
            papillomaviruses suggest a commensalic nature of these viruses
  JOURNAL   J. Virol. 74 (24), 11636-11641 (2000)
   PUBMED   11090162
REFERENCE   2  (bases 1 to 434)
  AUTHORS   Antonsson,A., Forslund,O. and Hansson,B.-G.
  TITLE     Direct Submission
  JOURNAL   Submitted (21-DEC-1999) Medical Microbiology, Virology, Entrance
            78, UMAS, Malmo SE 20502, Sweden
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, then if we click any of this access number we will get the full data right. We discussed earlier. So, they give the definition, they give the keywords and the general name right.

(Refer Slide Time: 06:47)



Advanced search

1 - 30 entries / Number of founds: 26506

PrimaryAccessionNumber	Definition
AF217656	Definition:Human papillomavirus is Organism:Human papillomavirus
AF217657	Definition:Human papillomavirus is Organism:Human papillomavirus
AF217659	Definition:Human papillomavirus is Organism:Human papillomavirus
AF217658	Definition:Human papillomavirus Organism:Human papillomavirus
AF217660	Definition:Human papillomavirus Organism:Human papillomavirus
AF217661	Definition:Human papillomavirus Organism:Human papillomavirus
AF455142	Definition:Human papillomavirus Organism:Human papillomavirus
AF455144	Definition:Human papillomavirus Organism:Human papillomavirus
AF455146	Definition:Human papillomavirus Organism:Human papillomavirus

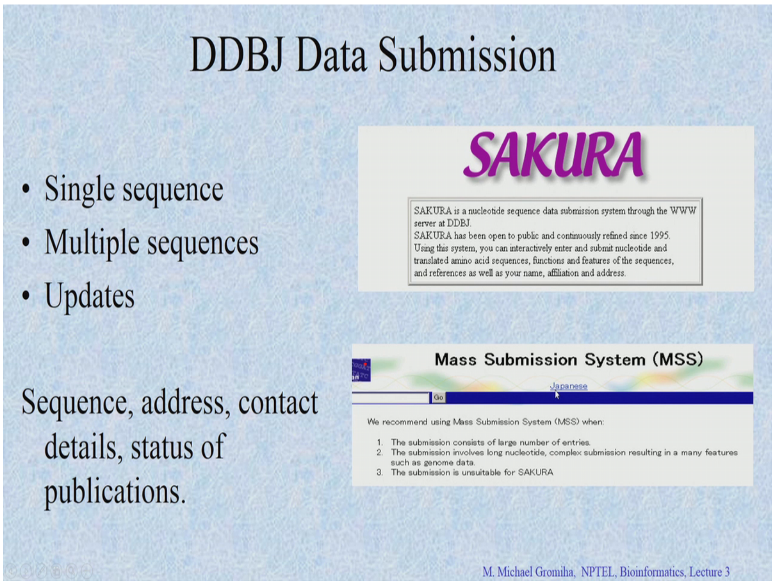
```
LOCUS       AF217656                 434 bp    DNA     linear   VRL 30-NOV-2000
DEFINITION  Human papillomavirus isolate FA14 major capsid protein L1 gene,
            partial cds.
ACCESSION   AF217656
VERSION     AF217656.1
KEYWORDS    .
SOURCE      Human papillomavirus
  ORGANISM  Human papillomavirus
            Viruses; dsDNA viruses, no RNA stage; Papillomaviridae;
            unclassified Papillomaviridae.
REFERENCE   1  (bases 1 to 434)
  AUTHORS   Antonsson,A., Forslund,O., Ekberg,H., Sterner,G. and Hansson,B.G.
  TITLE     The ubiquity and impressive genomic diversity of human skin
```

```
                     /translation="NIYNNQGTRLEVPKVSGNQHRVFRLKLPDPNRFALADMSVYNPD
                     KERLVWGLKGIEIGRGQPLGIGSSGHPLFNKVNDTENGNTYRNSSKDDRQNISFDPKQ
                     LQMFIIGCTPCIGEHWDRAPACVNDDQAGRCPPIELINSYIQ"
BASE COUNT      153 a        71 c        93 g       117 t
ORIGIN
        1 tcaatattta taacaatcaa ggcacacgat tggaggttcc taaagtatca ggaaatcaac
       61 acagggtatt tagattaaag ctaccagatc ctaataggtt tgcgttagct gacatgtcag
      121 tatataaccc tgacaaagaa agattagtat ggggtttgaa aggcatagaa ataggcaggg
      181 gccaaccttt aggaataggc agcagtggtc atccactgtt taataaggtt aatgatacag
      241 aaaatggcaa tacatatagg aactcctcta aggatgatag acaaaatatt tcatttgacc
      301 ccaagcagtt gcaaatgttt attattggct gtactccatg tataggagaa cattgggaca
      361 gagcaccagc atgtgttaat gatgatcaag ctggtagatg tcctcctata gagttaataa
      421 actcatatat acag
//
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, this is the translator protein sequence right, this is the niy where with the last class we discussed about the coding right is the one codes for a protein, I mean amino acid right. So, this is a protein sequence this is DNA sequence, I will get a protein sequence fine.

(Refer Slide Time: 07:02)



DDBJ Data Submission

- Single sequence
- Multiple sequences
- Updates

Sequence, address, contact details, status of publications.

SAKURA

SAKURA is a nucleotide sequence data submission system through the WWW server at DDBJ.
SAKURA has been open to public and continuously refined since 1995.
Using this system, you can interactively enter and submit nucleotide and translated amino acid sequences, functions and features of the sequences, and references as well as your name, affiliation and address.

Mass Submission System (MSS)

We recommend using Mass Submission System (MSS) when:

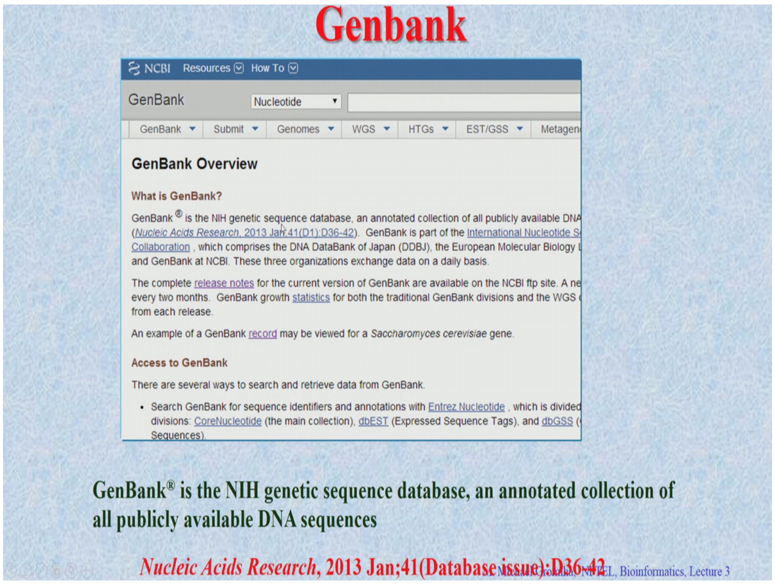
1. The submission consists of large number of entries.
2. The submission involves long nucleotide, complex submission resulting in a many features such as genome data.
3. The submission is unsuitable for SAKURA

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

Now, the DDBJ they have the option to upload your data if we have any sequence right you can also upload the data. So, either you give the single sequence or the multiple sequences.

So, they receive the data and they first validate, and if are all aspects if the data is clean then they will a include your data in the database right that is fine similar way. Now, we discussed about DDBJ, similar data are maintain in the GenBank and EMBL right. So, what is GenBank, how many of you use GenBank? Fine right. So, what is a GenBank right?

(Refer Slide Time: 07:47)



Genbank

GenBank Overview

What is GenBank?

GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide S Collaboration , which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology L and GenBank at NCBI. These three organizations exchange data on a daily basis.

The complete release notes for the current version of GenBank are available on the NCBI ftp site. A ne every two months. GenBank growth statistics for both the traditional GenBank divisions and the WGS from each release.

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

Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with Entrez Nucleotide , which is divided divisions: CoreNucleotide (the main collection), dbEST (Expressed Sequence Tags), and dbGSS (Sequences).

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences

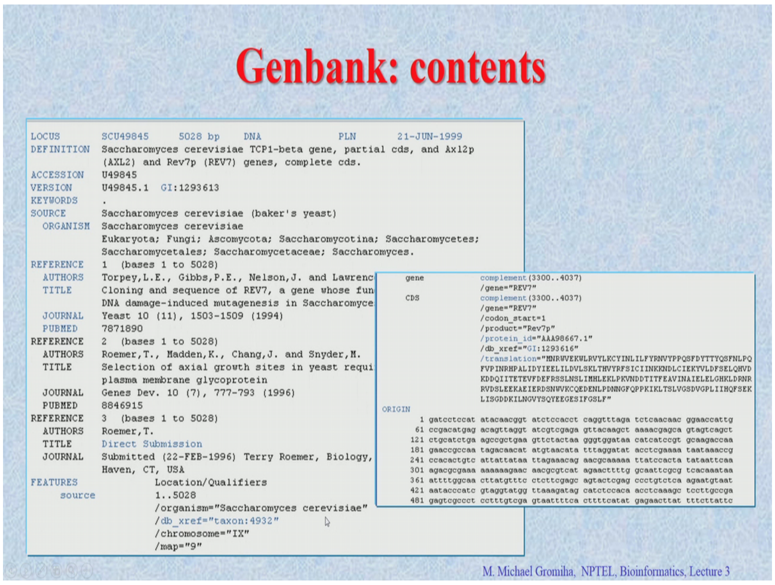
Nucleic Acids Research, 2013 Jan;41(Database issue):D36-42

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

It is a NIH genetic sequence database right, they provide the collection of the information regarding the DNA sequences right. So, they give all the information this is the website. So, consider GenBank will get the information regarding the sequence available in GenBank. So, it is the developed by NIH.

So, in CBI a there maintaining this database right on the all the d-th they publish in your database issues anywhere because anywhere a nucleic acids research is the general as they discussed earlier. So, they publish the details about the databases and you can get all the information regarding that fine.

(Refer Slide Time: 08:20)



Genbank: contents

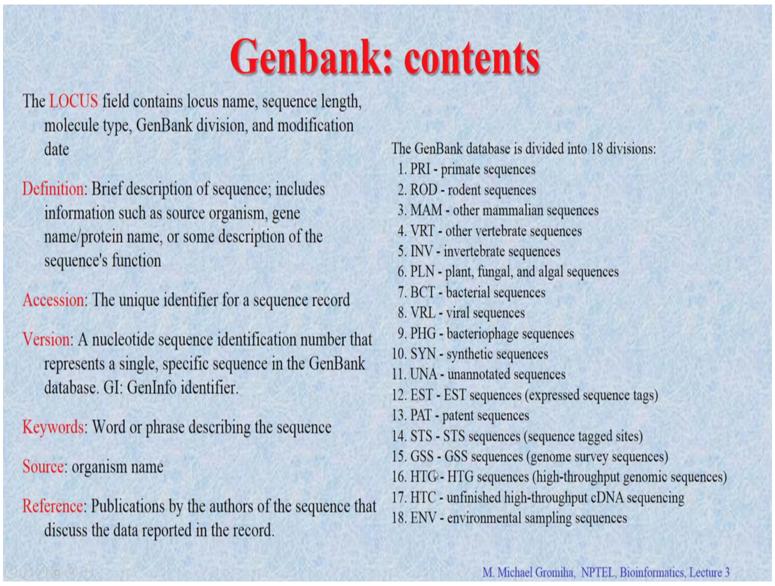
```
LOCUS       SCU49845     5028 bp    DNA             PLN       21-JUN-1999
DEFINITION  Saccharomyces cerevisiae TCP1-beta gene, partial cds, and Axl2p
            (AXL2) and Rev7p (REV7) genes, complete cds.
ACCESSION   U49845
VERSION     U49845.1  GI:1293613
KEYWORDS    .
SOURCE      Saccharomyces cerevisiae (baker's yeast)
  ORGANISM  Saccharomyces cerevisiae
            Eukaryota; Fungi; Ascomycota; Saccharomycotina; Saccharomycetes;
            Saccharomycetales; Saccharomycetaceae; Saccharomyces.
REFERENCE   1  (bases 1 to 5028)
  AUTHORS   Torpey,L.E., Gibbs,P.E., Nelson,J. and Lawrenc
  TITLE     Cloning and sequence of REV7, a gene whose fun
            DNA damage-induced mutagenesis in Saccharomyce
  JOURNAL   Yeast 10 (11), 1503-1509 (1994)
  PUBMED    7871890
REFERENCE   2  (bases 1 to 5028)
  AUTHORS   Roemer,T., Madden,K., Chang,J. and Snyder,M.
  TITLE     Selection of axial growth sites in yeast requi
            plasma membrane glycoprotein
  JOURNAL   Genes Dev. 10 (7), 777-793 (1996)
  PUBMED    8846915
REFERENCE   3  (bases 1 to 5028)
  AUTHORS   Roemer,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (22-FEB-1996) Terry Roemer, Biology,
            Haven, CT, USA
FEATURES             Location/Qualifiers
     source          1..5028
                     /organism="Saccharomyces cerevisiae"
                     /db_xref="taxon:4932"
                     /chromosome="IX"
                     /map="9"
```

```
     gene            complement(3300..4037)
                     /gene="REV7"
     CDS             complement(3300..4037)
                     /gene="REV7"
                     /codon_start=1
                     /product="Rev7p"
                     /protein_id="AAA98667.1"
                     /db_xref="GI:1293616"
                     /translation="MNRWVEKWLRVYLKCYINLILFYRNVYPPQSFDYTTYQSFNLPQ
                     FVPINRHPALIDYIEELILDVLSKLTHVYRFSICIINKKNDLCIEKYVLDFSELQHVD
                     KDDQIITETEVFDEFRSSLNSLIMHLEKLPKVNDDTITFEAVINAIELELGHKLDRNR
                     RVDSLEEKAEIERDSNWVKCQEDENLPDNNGFQPPKIKLTSLVGSDVGPLIIHQFSEK
                     LISGDDKILNGVYSQYEEGESIFGSLF"
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
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, this is the contents of GenBank. So, like we discussed in DDBJ. So, GenBank also a similar type of contents right; maybe we can say look into this website and get more details just I will go through quickly. So, they has the definition because this is. So, this is saccharomyces cerevisiae right. So, this is accession number right here you give the organism and if the others names right the reference and the features.

(Refer Slide Time: 08:45)



Genbank: contents

The LOCUS field contains locus name, sequence length, molecule type, GenBank division, and modification date

Definition: Brief description of sequence; includes information such as source organism, gene name/protein name, or some description of the sequence's function

Accession: The unique identifier for a sequence record

Version: A nucleotide sequence identification number that represents a single, specific sequence in the GenBank database. GI: GenInfo identifier.

Keywords: Word or phrase describing the sequence

Source: organism name

Reference: Publications by the authors of the sequence that discuss the data reported in the record.

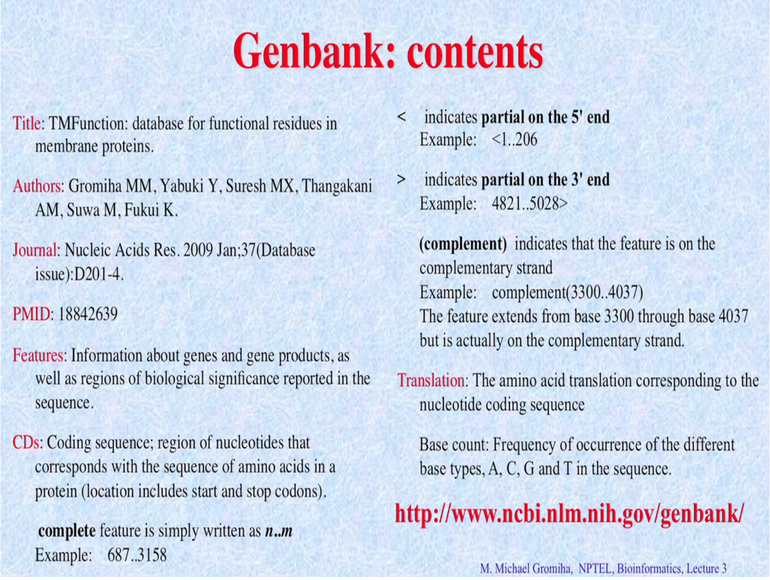
The GenBank database is divided into 18 divisions:

1. PRI - primate sequences
2. ROD - rodent sequences
3. MAM - other mammalian sequences
4. VRT - other vertebrate sequences
5. INV - invertebrate sequences
6. PLN - plant, fungal, and algal sequences
7. BCT - bacterial sequences
8. VRL - viral sequences
9. PHG - bacteriophage sequences
10. SYN - synthetic sequences
11. UNA - unannotated sequences
12. EST - EST sequences (expressed sequence tags)
13. PAT - patent sequences
14. STS - STS sequences (sequence tagged sites)
15. GSS - GSS sequences (genome survey sequences)
16. HTG - HTG sequences (high-throughput genomic sequences)
17. HTC - unfinished high-throughput cDNA sequencing
18. ENV - environmental sampling sequences

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, here you give the various divisions, they give the primary sequences and rodent sequences, mammalian sequences, they have the various conditions, they various classifications they given the separately. So, if you are interested any given specific sequences for example, viral sequences right they give the database for the data for the viral sequences, they have various classifications right.

(Refer Slide Time: 09:07)



Genbank: contents

Title: TMFunction: database for functional residues in membrane proteins.

Authors: Gromiha MM, Yabuki Y, Suresh MX, Thangakani AM, Suwa M, Fukui K.

Journal: Nucleic Acids Res. 2009 Jan;37(Database issue):D201-4.

PMID: 18842639

Features: Information about genes and gene products, as well as regions of biological significance reported in the sequence.

CDs: Coding sequence; region of nucleotides that corresponds with the sequence of amino acids in a protein (location includes start and stop codons).

complete feature is simply written as ***n..m***
Example: 687..3158

< indicates **partial on the 5' end**
Example: <1..206

> indicates **partial on the 3' end**
Example: 4821..5028>

(complement) indicates that the feature is on the complementary strand
Example: complement(3300..4037)
The feature extends from base 3300 through base 4037 but is actually on the complementary strand.

Translation: The amino acid translation corresponding to the nucleotide coding sequence

Base count: Frequency of occurrence of the different base types, A, C, G and T in the sequence.

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

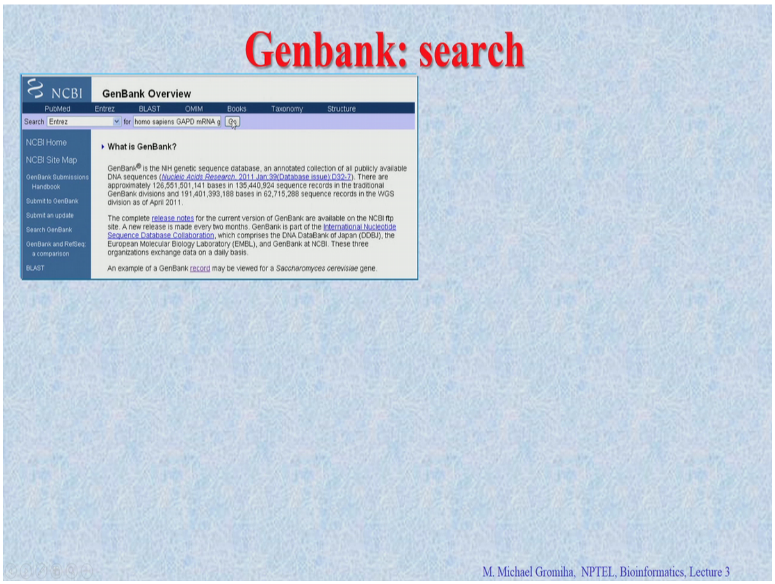
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, then again they give the data about the literature and what the different features and the coding sequence. So, what is the coding sequence?

Student: (Refer Time: 09:17) translated protein.

Translated into protein so that nuclei types that corresponding the sequence of amino acid in protein neither they how they translate. So, this give from which two which. So, we will put n dot dot means from n to m for example, 687 to 3158 see this is the coding region. Then if you partial one they put the less than symbol right we have partial at the 5 dash n then if it is greater than symbol they put partial at the 3 (Refer Time: 09:44). So, they give the information right we also give the complementary stand information right and the translation.

(Refer Slide Time: 09:51)



Genbank: search

NCBI GenBank Overview

Search: Entrez for: homo sapiens GAPD mRNA g

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.

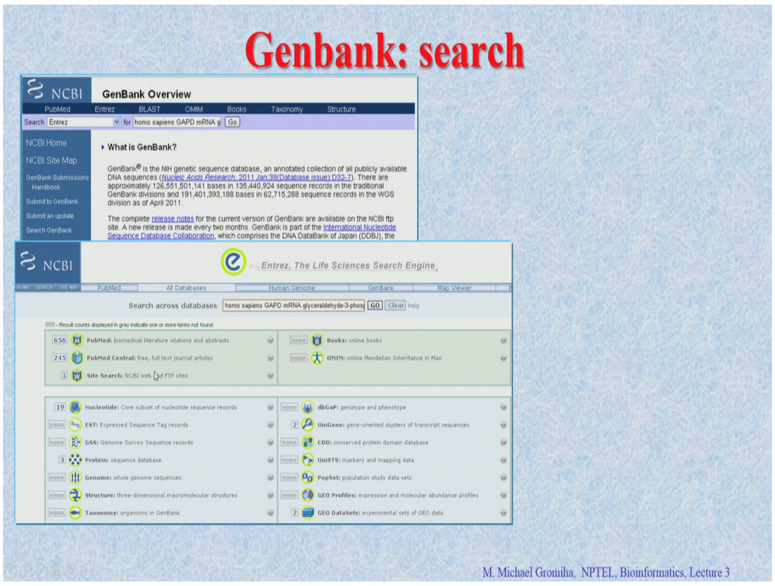
The complete release notes for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (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.

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, here this is the search option available. So, you put the keyword here right if you click on this go, and then you will get the list.

(Refer Slide Time: 10:00)



Genbank: search

NCBI GenBank Overview

Search: Entrez for: homo sapiens GAPD mRNA g

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

NCBI Entrez, The Life Sciences Search Engine

Search across databases: homo sapiens GAPD mRNA glyceraldehyde-3-phos

- Result counts displayed in gray indicate one or more terms not found

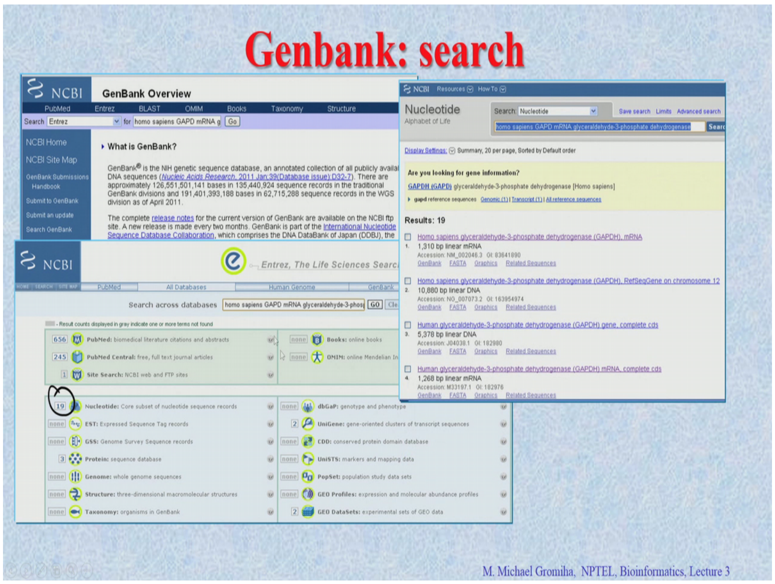
656	PubMed: biomedical literature citations and abstracts	none	Books: online books
245	PubMed Central: free, full text journal articles	none	OMIM: online Mendelian Inheritance in Man
1	Site Search: NCBI web and FTP sites		

19	Nucleotide: Core subset of nucleotide sequence records	none	dbGaP: genotype and phenotype
none	EST: Expressed Sequence Tag records	2	UniGene: gene-oriented clusters of transcript sequences
none	GSS: Genome Survey Sequence records	none	CDD: conserved protein domain database
3	Protein: sequence database	none	UniSTS: markers and mapping data
none	Genome: whole genome sequences	none	PopSet: population study data sets
none	Structure: three-dimensional macromolecular structures	none	GEO Profiles: expression and molecular abundance profiles
none	Taxonomy: organisms in GenBank	2	GEO DataSets: experimental sets of GEO data

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, they give you different types of information, not only the nucleic acid sequences they give the data about the how many literature in the buffered, right if order the how many nucleic rate sequences, I mean GST sequences right to protein sequence and so on. So, if you click on the nuclei type. So, there are 19.

(Refer Slide Time: 10:19)



Genbank: search

NCBI GenBank Overview

PubMed Entrez BLAST OMIM Books Taxonomy Structure

Search Entrez for homo sapiens GAPD mRNA g Go

NCBI Home
NCBI Site Map
GenBank Submissions Handbook
Submit to GenBank
Submit an update
Search GenBank

What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly availal 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

NCBI Entrez, The Life Sciences Searc

HOME SEARCH SITE MAP PubMed All Databases Human Genome GenBank

Search across databases homo sapiens GAPD mRNA glyceraldehyde-3-phos GO Cle

- Result counts displayed in gray indicate one or more terms not found

656 PubMed: biomedical literature citations and abstracts
245 PubMed Central: free, full text journal articles
1 Site Search: NCBI web and FTP sites
none Books: online books
none OMIM: online Mendelian In
19 Nucleotide: Core subset of nucleotide sequence records
none EST: Expressed Sequence Tag records
none GSS: Genome Survey Sequence records
3 Protein: sequence database
none Genome: whole genome sequences
none Structure: three-dimensional macromolecular structures
none Taxonomy: organisms in GenBank
none dbGaP: genotype and phenotype
2 UniGene: gene-oriented clusters of transcript sequences
none CDD: conserved protein domain database
none UniSTS: markers and mapping data
none PopSet: population study data sets
none GEO Profiles: expression and molecular abundance profiles
2 GEO DataSets: experimental sets of GEO data

NCBI Resources How To

Nucleotide
Alphabet of Life

Search: Nucleotide Save search Limits Advanced search
homo sapiens GAPD mRNA glyceraldehyde-3-phosphate dehydrogenase Searc

Display Settings: Summary, 20 per page, Sorted by Default order

Are you looking for gene information?
GAPDH (GAPD) glyceraldehyde-3-phosphate dehydrogenase [Homo sapiens]
gapd reference sequences Genomic (1) | Transcript (1) | All reference sequences

Results: 19

1. Homo sapiens glyceraldehyde-3-phosphate dehydrogenase (GAPDH), mRNA
1,310 bp linear mRNA
Accession: NM_002046.3 GI: 83641890
GenBank FASTA Graphics Related Sequences

2. Homo sapiens glyceraldehyde-3-phosphate dehydrogenase (GAPDH), RefSeqGene on chromosome 12
10,880 bp linear DNA
Accession: NG_007073.2 GI: 163954974
GenBank FASTA Graphics Related Sequences

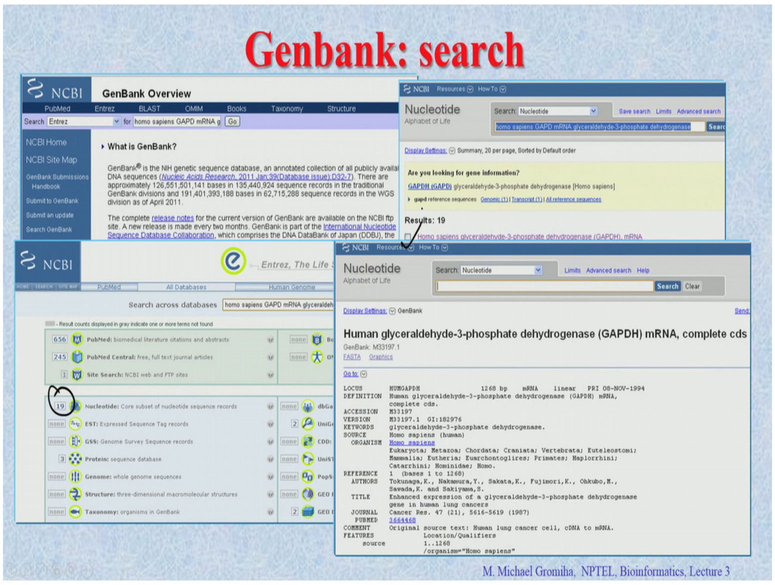
3. Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene, complete cds
5,378 bp linear DNA
Accession: J04038.1 GI: 182980
GenBank FASTA Graphics Related Sequences

4. Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA, complete cds
1,268 bp linear mRNA
Accession: M33197.1 GI: 182976
GenBank FASTA Graphics Related Sequences

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, now if you get all the 19 correct these are the 19 sequences you get from GenBank.

(Refer Slide Time: 10:28)



Genbank: search

NCBI GenBank Overview

What is GenBank?

Search across databases homo sapiens GAPD mRNA glyceraldehyde

Display Settings: Summary, 20 per page, Sorted by Default order

Are you looking for gene information?
GAPDH (GAPD) glyceraldehyde-3-phosphate dehydrogenase [Homo sapiens]

Results: 19

Homo sapiens glyceraldehyde-3-phosphate dehydrogenase (GAPDH), mRNA

NCBI Resources How To

Nucleotide
Alphabet of Life

Search: Nucleotide Limits Advanced search Help
Search Clear

Display Settings: GenBank Send:

Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA, complete cds
GenBank: M33197.1
FASTA Graphics

Go to:

```
LOCUS       HUMGAPDH                1268 bp    mRNA    linear   PRI 08-NOV-1994
DEFINITION  Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA,
            complete cds.
ACCESSION   M33197
VERSION     M33197.1  GI:182976
KEYWORDS    glyceraldehyde-3-phosphate dehydrogenase.
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 1268)
  AUTHORS   Tokunaga,K., Nakamura,Y., Sakata,K., Fujimori,K., Ohkubo,M.,
            Sawada,K. and Sakiyama,S.
  TITLE     Enhanced expression of a glyceraldehyde-3-phosphate dehydrogenase
            gene in human lung cancers
  JOURNAL   Cancer Res. 47 (21), 5616-5619 (1987)
   PUBMED   3664468
COMMENT     Original source text: Human lung cancer cell, cDNA to mRNA.
FEATURES             Location/Qualifiers
     source          1..1268
                     /organism="Homo sapiens"
```

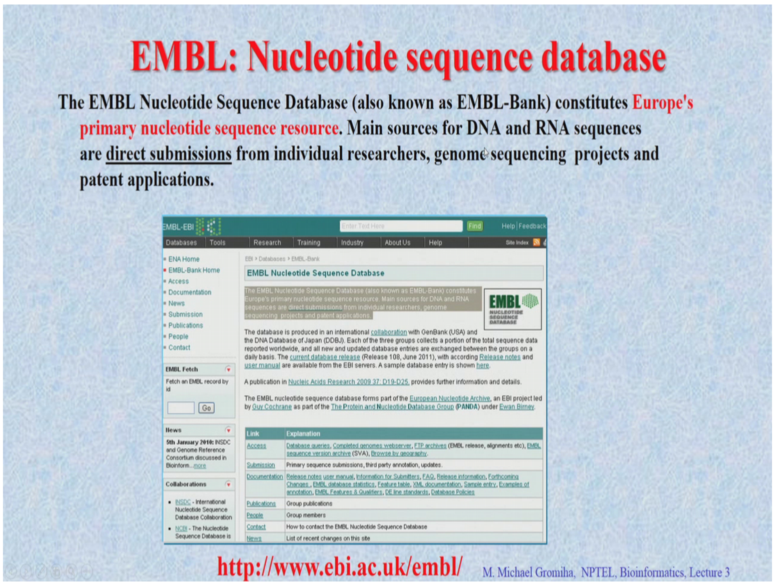
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

Then if you click on this first one right then it will show the details right fine. So, now, get the data from the GenBank right we discussed about the DDBJ, I am discussed about the GenBank. So, another one is EMBL.

EMBL is maintain from European micro biology laboratory right. So, one is in US one is in Europe and another one is Japan right because they have they very good first commuting facility under facility long time ago this is way first the developed there right. Using most of the databases called PDB Uniprot and all these databases right they have the (Refer Time: 11:01) between all these developed countries, because internet came there first and they are very first computing facilities and all. So, they started there because familiar trying to use there right because India we got very late right because when I visited Japan in 1997, they ask with you to do with the all everything with the email and all.

But we can only facts at the time right, but currently it is synchronized right whatever we get in the developed countries, we immediately in India also right. This is the reason why almost of the early developed databases right ever thing is from US or in the Europe or in Japan fine; so EMBL. So, they also start we have the database for the nuclei acid sequence database right.

(Refer Slide Time: 11:43)



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.

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 Protein and Nucleotide Database Group (PANDA) under Ewan Birney.

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

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

http://www.ebi.ac.uk/embl/ M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

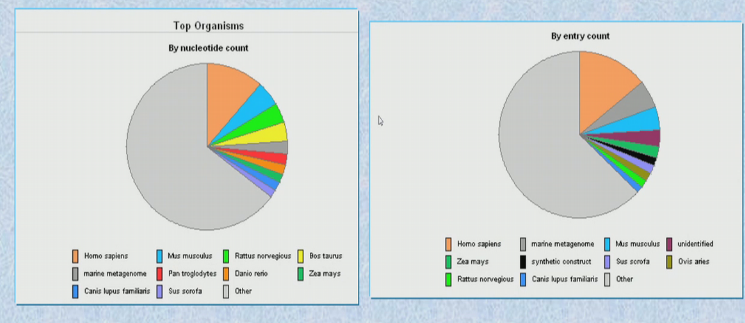
Here the main resources sources are the DNA and RNA sequences right. So, they also accept direct submissions, they are also (Refer Time: 11:50) data right. So, mainly from the genome projects as well as the patent applications, this is the website for the EMBL right you can the EBI dot ac dot UK and you got the EMBL. So, this is the website.

(Refer Slide Time: 12:06)

EMBL: Statistics

Release on 16-Jan-2014.

The release contains 798.2 million sequence entries comprising 2004.9 billion nucleotides.

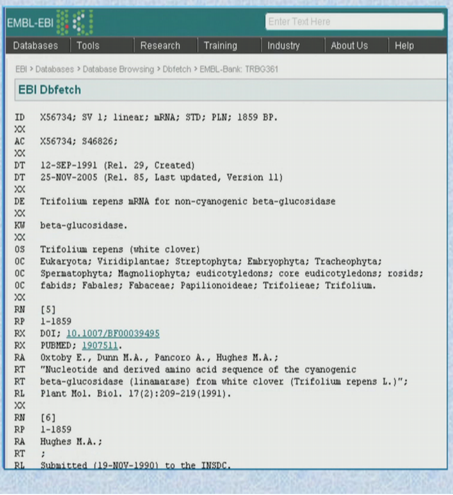


M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, now if you got to the EMBL this is the statistics, but similar to DDBJ, now correctly see the mainly this is the organisms right Homo sapiens is the most right and the forward by the other organisms, if are also this is entry count and when nuclei count. So, you can see the similar level of the statistics fine ok.

(Refer Slide Time: 12:25)

Database entry

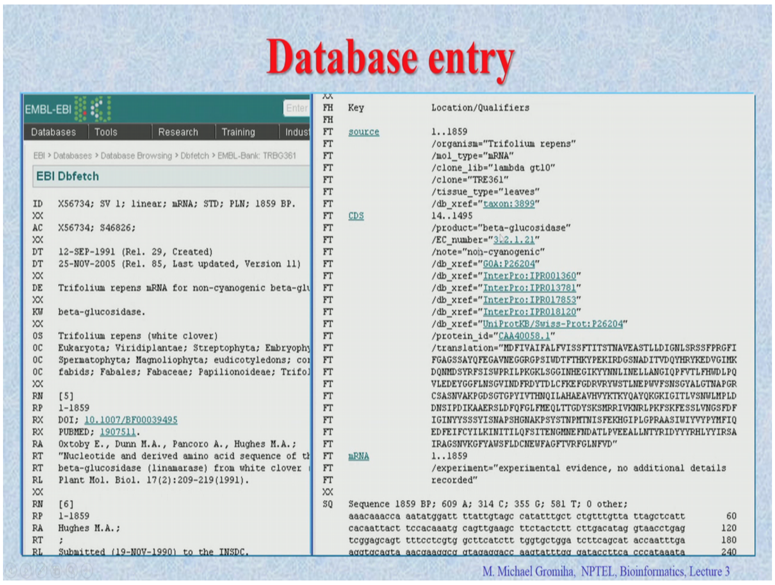


```
ID   X56734; SV 1; linear; mRNA; STD; PLN; 1859 BP.
XX
AC   X56734; S46826;
XX
DT   12-SEP-1991 (Rel. 29, Created)
DT   25-NOV-2005 (Rel. 85, Last updated, Version 11)
XX
DE   Trifolium repens mRNA for non-cyanogenic beta-glucosidase
XX
KW   beta-glucosidase.
XX
OS   Trifolium repens (white clover)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids;
OC   fabids; Fabales; Fabaceae; Papilionoideae; Trifolieae; Trifolium.
XX
RN   [5]
RP   1-1859
RX   DOI; 10.1007/BF00039495
RX   PUBMED; 1907511.
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of the cyanogenic
RT   beta-glucosidase (linamarase) from white clover (Trifolium repens L.)";
RL   Plant Mol. Biol. 17(2):209-219(1991).
XX
RN   [6]
RP   1-1859
RA   Hughes M.A.;
RT   ;
RL   Submitted (19-NOV-1990) to the INSDC.
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

Now, go to the database entry, here also if you see the similar to the GenBank or familiar DDBJ, we see the names and the PUBMED entries and we have the coding regions this is the proteins sequence and we have the DNA sequence.

(Refer Slide Time: 12:37)



Database entry

EMBL-EBI
Enter
Databases
Tools
Research
Training
Indust
EBI > Databases > Database Browsing > Dbfetch > EMBL-Bank: TRBG361

EBI Dbfetch

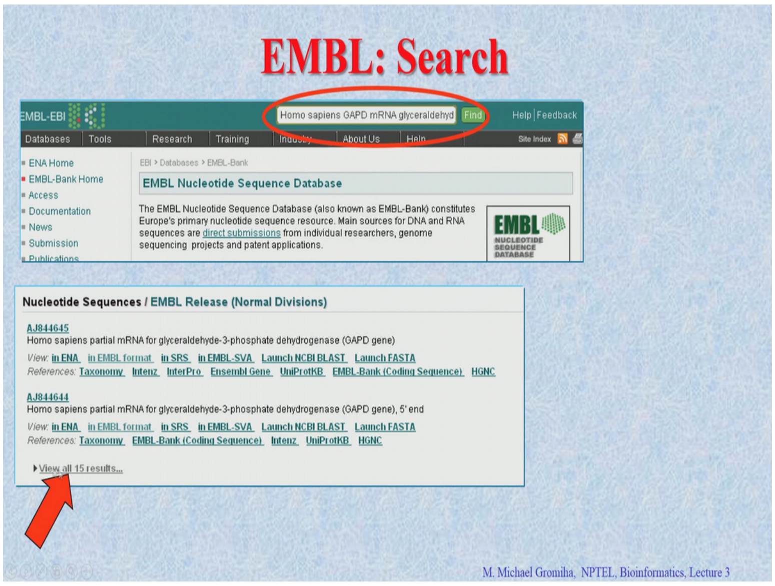
```
ID   X56734; SV 1; linear; mRNA; STD; PLN; 1859 BP.
XX
AC   X56734; S46826;
XX
DT   12-SEP-1991 (Rel. 29, Created)
DT   25-NOV-2005 (Rel. 85, Last updated, Version 11)
XX
DE   Trifolium repens mRNA for non-cyanogenic beta-glu
XX
KW   beta-glucosidase.
XX
OS   Trifolium repens (white clover)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophy
OC   Spermatophyta; Magnoliophyta; eudicotyledons; cor
OC   fabids; Fabales; Fabaceae; Papilionoideae; Trifol
XX
RN   [5]
RP   1-1859
RX   DOI; 10.1007/BF00039495
RX   PUBMED; 1907511.
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of th
RT   beta-glucosidase (linamarase) from white clover (
RL   Plant Mol. Biol. 17(2):209-219(1991).
XX
RN   [6]
RP   1-1859
RA   Hughes M.A.;
RT   ;
RL   Submitted (19-NOV-1990) to the INSDC.
```

```
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1859
FT                   /organism="Trifolium repens"
FT                   /mol_type="mRNA"
FT                   /clone_lib="lambda gt10"
FT                   /clone="TRE361"
FT                   /tissue_type="leaves"
FT                   /db_xref="taxon:3899"
FT   CDS             14..1495
FT                   /product="beta-glucosidase"
FT                   /EC_number="3.2.1.21"
FT                   /note="non-cyanogenic"
FT                   /db_xref="GOA:P26204"
FT                   /db_xref="InterPro:IPR001360"
FT                   /db_xref="InterPro:IPR013781"
FT                   /db_xref="InterPro:IPR017853"
FT                   /db_xref="InterPro:IPR018120"
FT                   /db_xref="UniProtKB/Swiss-Prot:P26204"
FT                   /protein_id="CAA40058.1"
FT                   /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT                   FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT                   DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT                   VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT                   CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT                   DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
FT                   IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT                   EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT                   IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
FT   mRNA            1..1859
FT                   /experiment="experimental evidence, no additional details
FT                   recorded"
XX
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
     aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt        60
     cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag       120
     tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga       180
     aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata       240
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

This has 1859 base pairs and you can see the condense of ACG and T right fine. So, EMBL also have a search options.

(Refer Slide Time: 12:51)



EMBL: Search

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.

Nucleotide Sequences / EMBL Release (Normal Divisions)

AJ844645
Homo sapiens partial mRNA for glyceraldehyde-3-phosphate dehydrogenase (GAPD gene)
View: in ENA in EMBL format in SRS in EMBL-SVA Launch NCBI BLAST Launch FASTA
References: Taxonomy Intenz InterPro Ensembl Gene UniProtKB EMBL-Bank (Coding Sequence) HGNC

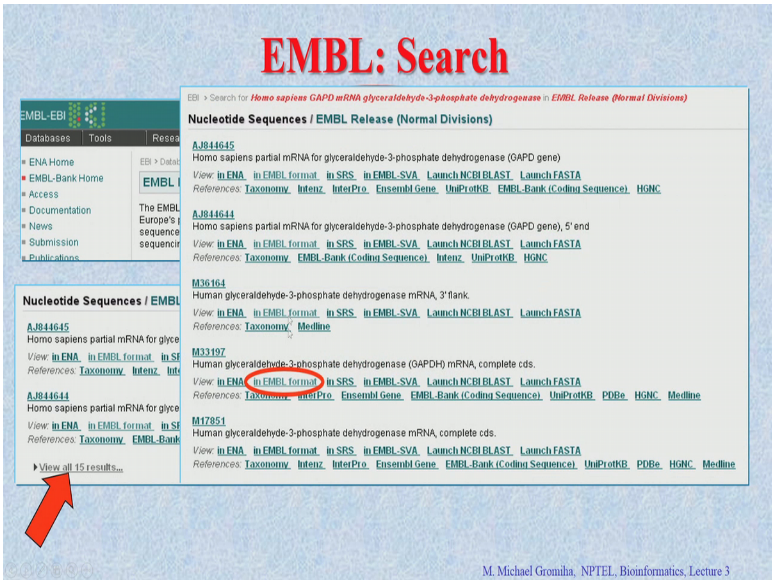
AJ844644
Homo sapiens partial mRNA for glyceraldehyde-3-phosphate dehydrogenase (GAPD gene), 5' end
View: in ENA in EMBL format in SRS in EMBL-SVA Launch NCBI BLAST Launch FASTA
References: Taxonomy EMBL-Bank (Coding Sequence) Intenz UniProtKB HGNC

View all 15 results...

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

You can search to the homo sapiens gapdm RNA right glyceraldehyd site. Now, go there. So, they have a various options right.

(Refer Slide Time: 13:06)



EMBL: Search

EBI > Search for *Homo sapiens GAPD mRNA glyceraldehyde-3-phosphate dehydrogenase* in *EMBL Release (Normal Divisions)*

Nucleotide Sequences / EMBL Release (Normal Divisions)

AJ844645
Homo sapiens partial mRNA for glyceraldehyde-3-phosphate dehydrogenase (GAPD gene)
View: in ENA in EMBL format in SRS in EMBL-SVA Launch NCBI BLAST Launch FASTA
References: Taxonomy Intenz InterPro Ensembl Gene UniProtKB EMBL-Bank (Coding Sequence) HGNC

AJ844644
Homo sapiens partial mRNA for glyceraldehyde-3-phosphate dehydrogenase (GAPD gene), 5' end
View: in ENA in EMBL format in SRS in EMBL-SVA Launch NCBI BLAST Launch FASTA
References: Taxonomy EMBL-Bank (Coding Sequence) Intenz UniProtKB HGNC

M36164
Human glyceraldehyde-3-phosphate dehydrogenase mRNA, 3' flank.
View: in ENA in EMBL format in SRS in EMBL-SVA Launch NCBI BLAST Launch FASTA
References: Taxonomy Medline

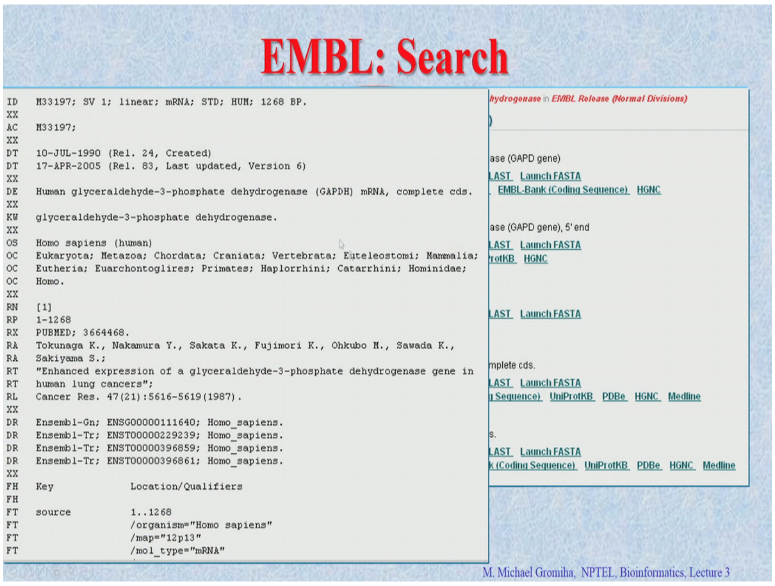
M33197
Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA, complete cds.
View: in ENA in EMBL format in SRS in EMBL-SVA Launch NCBI BLAST Launch FASTA
References: Taxonomy InterPro Ensembl Gene EMBL-Bank (Coding Sequence) UniProtKB PDBe HGNC Medline

M17851
Human glyceraldehyde-3-phosphate dehydrogenase mRNA, complete cds.
View: in ENA in EMBL format in SRS in EMBL-SVA Launch NCBI BLAST Launch FASTA
References: Taxonomy Intenz InterPro Ensembl Gene EMBL-Bank (Coding Sequence) UniProtKB PDBe HGNC Medline

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

If you give all the results these are the various results for the EMBL. So, now, we have a very format you check, whether the EMBL format if you click in the EMBL format right.

(Refer Slide Time: 13:15)



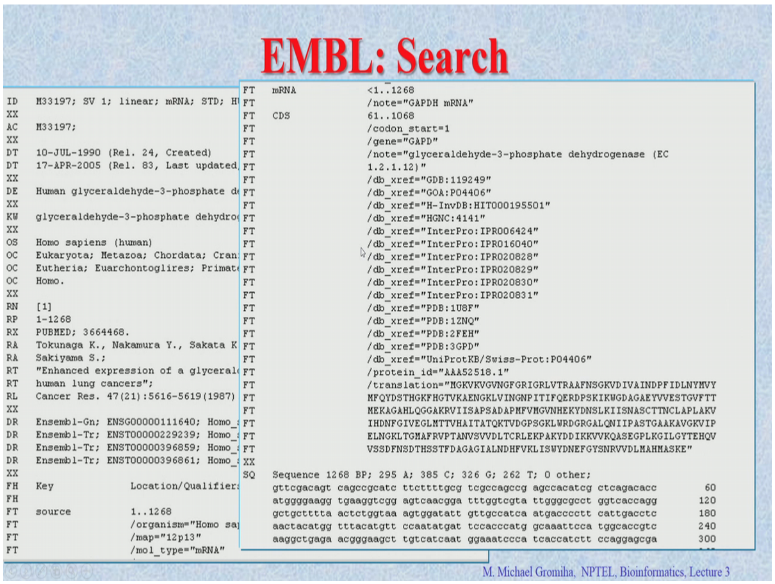
EMBL: Search

```
ID   M33197; SV 1; linear; mRNA; STD; HUM; 1268 BP.
XX
AC   M33197;
XX
DT   10-JUL-1990 (Rel. 24, Created)
DT   17-APR-2005 (Rel. 83, Last updated, Version 6)
XX
DE   Human glyceraldehyde-3-phosphate dehydrogenase (GAPDH) mRNA, complete cds.
XX
KW   glyceraldehyde-3-phosphate dehydrogenase.
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   1-1268
RX   PUBMED; 3664468.
RA   Tokunaga K., Nakamura Y., Sakata K., Fujimori K., Ohkubo M., Sawada K.,
RA   Sakiyama S.;
RT   "Enhanced expression of a glyceraldehyde-3-phosphate dehydrogenase gene in
RT   human lung cancers";
RL   Cancer Res. 47(21):5616-5619(1987).
XX
DR   Ensembl-Gn; ENSG00000111640; Homo_sapiens.
DR   Ensembl-Tr; ENST00000229239; Homo_sapiens.
DR   Ensembl-Tr; ENST00000396859; Homo_sapiens.
DR   Ensembl-Tr; ENST00000396861; Homo_sapiens.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1268
FT                   /organism="Homo sapiens"
FT                   /map="12p13"
FT                   /mol_type="mRNA"
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, it gives you the EMBL format, this is the contents of this particular database.

(Refer Slide Time: 13:20)



EMBL: Search

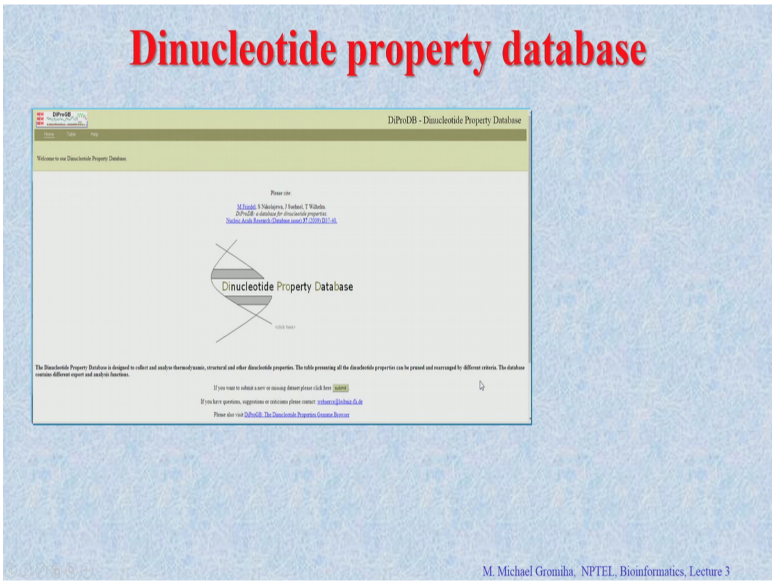
```
ID   M33197; SV 1; linear; mRNA; STD; H
XX
AC   M33197;
XX
DT   10-JUL-1990 (Rel. 24, Created)
DT   17-APR-2005 (Rel. 83, Last updated,
XX
DE   Human glyceraldehyde-3-phosphate d
XX
KW   glyceraldehyde-3-phosphate dehydro
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Cran
OC   Eutheria; Euarchontoglires; Primat
OC   Homo.
XX
RN   [1]
RP   1-1268
RX   PUBMED; 3664468.
RA   Tokunaga K., Nakamura Y., Sakata K
RA   Sakiyama S.;
RT   "Enhanced expression of a glyceral
RT   human lung cancers";
RL   Cancer Res. 47(21):5616-5619(1987)
XX
DR   Ensembl-Gn; ENSG00000111640; Homo_
DR   Ensembl-Tr; ENST00000229239; Homo_
DR   Ensembl-Tr; ENST00000396859; Homo_
DR   Ensembl-Tr; ENST00000396861; Homo_
XX
FH   Key             Location/Qualifier
FH
FT   source          1..1268
FT                   /organism="Homo sa
FT                   /map="12p13"
FT                   /mol_type="mRNA"
```

```
FT   mRNA            <1..1268
FT                   /note="GAPDH mRNA"
FT   CDS             61..1068
FT                   /codon_start=1
FT                   /gene="GAPD"
FT                   /note="glyceraldehyde-3-phosphate dehydrogenase (EC
FT                   1.2.1.12)"
FT                   /db_xref="GDB:119249"
FT                   /db_xref="GOA:P04406"
FT                   /db_xref="H-InvDB:HIT000195501"
FT                   /db_xref="HGNC:4141"
FT                   /db_xref="InterPro:IPR006424"
FT                   /db_xref="InterPro:IPR016040"
FT                   /db_xref="InterPro:IPR020828"
FT                   /db_xref="InterPro:IPR020829"
FT                   /db_xref="InterPro:IPR020830"
FT                   /db_xref="InterPro:IPR020831"
FT                   /db_xref="PDB:1U8F"
FT                   /db_xref="PDB:1ZNQ"
FT                   /db_xref="PDB:2FEH"
FT                   /db_xref="PDB:3GPD"
FT                   /db_xref="UniProtKB/Swiss-Prot:P04406"
FT                   /protein_id="AAA52518.1"
FT                   /translation="MGKVKVGVNGFGRIGRLVTRAAFNSGKVDIVAINDPFIDLNYMVY
FT                   MFQYDSTHGKFHGTVKAENGKLVINGNPITIFQERDPSKIKWGDAGAEYVVESTGVFTT
FT                   MEKAGAHLQGGAKRVIISAPSADAPMFVMGVNHEKYDNSLKIISNASCTTNCLAPLAKV
FT                   IHDNFGIVEGLMTTVHAITATQKTVDGPSGKLWRDGRGALQNIIPASTGAAKAVGKVIP
FT                   ELNGKLTGMAFRVPTANVSVVDLTCRLEKPAKYDDIKKVVKQASEGPLKGILGYTEHQV
FT                   VSSDFNSDTHSSTFDAGAGIALNDHFVKLISWYDNEFGYSNRVVDLMAHMASKE"
XX
SQ   Sequence 1268 BP; 295 A; 385 C; 326 G; 262 T; 0 other;
     gttcgacagt cagccgcatc ttcttttgcg tcgccagccg agccacatcg ctcagacacc        60
     atggggaagg tgaaggtcgg agtcaacgga tttggtcgta ttgggcgcct ggtcaccagg       120
     gctgctttta actctggtaa agtggatatt gttgccatca atgacccctt cattgacctc       180
     aactacatgg tttacatgtt ccaatatgat tccacccatg gcaaattcca tggcaccgtc       240
     aaggctgaga acgggaagct tgtcatcaat ggaaatccca tcaccatctt ccaggagcga       300
```

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

Then we have the protein sequence and the DNA sequence fine.

(Refer Slide Time: 13:24)



Dinucleotide property database

DiProDB - Dinucleotide Property Database

Welcome to our Dinucleotide Property Database.

Please cite:

M Friedel, S Nikolajewa, J Suehnel, T Wilhelm.
DiProDB: a database for dinucleotide properties.
Nucleic Acids Research (Database issue) **37** (2009) D37-40.

The Dinucleotide Property Database is designed to collect and analyse thermodynamic, structural and other dinucleotide properties. The table presenting all the dinucleotide properties can be pruned and rearranged by different criteria. The database contains different export and analysis functions.

If you want to submit a new or missing dataset please click here submit

If you have questions, suggestions or criticisms please contact: webserve@leibniz-fli.de

Please also visit DiProGB: The Dinucleotide Properties Genome Browser

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, if you use the DDBJ right r is the GenBank or is the EMBL now you can get the data for any of the all nucleated sequences. Then you can use the sequences for further analysis. So, you suggested to look into the databases and use the options available in this databases, and then try to get the sequences from different organisms and see whether anything is different or the same once again calculate different properties right.

Last class we discussed about various properties from the nucleotides right what various properties we discussed in the last class?

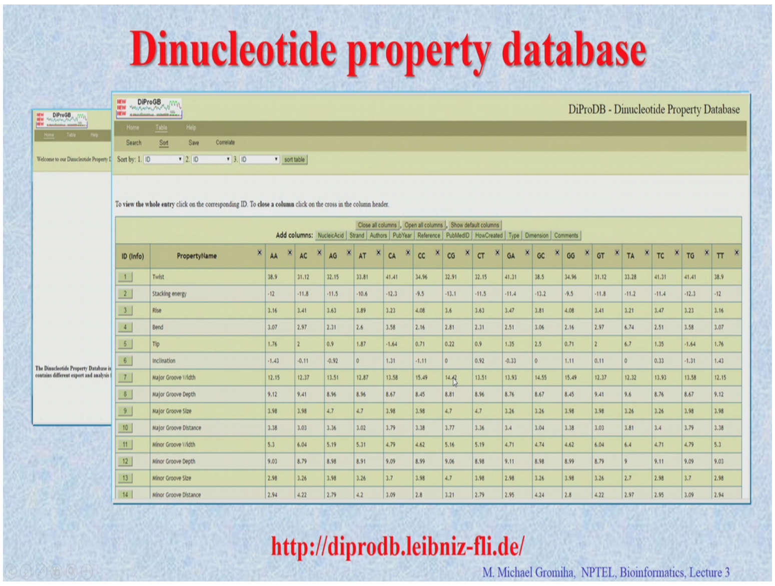
Student: Flexibility.

Flexibility.

Student: Stacking energy.

Stacking energy; so it depending upon the dinucleotide or trinucleotides right. So, we classified the nucleotide sequence in the overlapping segments, either dinucleotides or trinucleotides depending upon the availability of data, then we can calculate the average values. This will tell you how for this particular sequence is table compared with the other sequence or lot. So, there is database called dinucleotide proper database. Last class we discussed only few properties right mainly the flexibility or rigidity or the base stacking energy or hydrated bond. So, this database serious more than 140 a features right this available is this sets diprodb.

(Refer Slide Time: 14:43)



Dinucleotide property database

DiProDB - Dinucleotide Property Database

Home
Table
Help
Search
Sort
Save
Correlate
Sort by: 1. ID 2. ID 3. ID
sort table
Close all columns
Open all columns
Show default columns
Add columns: NucleicAcid Strand Authors PubYear Reference PubMedID HowCreated Type Dimension Comments

To view the whole entry click on the corresponding ID. To close a column click on the cross in the column header.

ID (Info)	PropertyName	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	Twist	38.9	31.12	32.15	33.81	41.41	34.96	32.91	32.15	41.31	38.5	34.96	31.12	33.28	41.31	41.41	38.9
2	Stacking energy	-12	-11.8	-11.5	-10.6	-12.3	-9.5	-13.1	-11.5	-11.4	-13.2	-9.5	-11.8	-11.2	-11.4	-12.3	-12
3	Rise	3.16	3.41	3.63	3.89	3.23	4.08	3.6	3.63	3.47	3.81	4.08	3.41	3.21	3.47	3.23	3.16
4	Bend	3.07	2.97	2.31	2.6	3.58	2.16	2.81	2.31	2.51	3.06	2.16	2.97	6.74	2.51	3.58	3.07
5	Tip	1.76	2	0.9	1.87	-1.64	0.71	0.22	0.9	1.35	2.5	0.71	2	6.7	1.35	-1.64	1.76
6	Inclination	-1.43	-0.11	-0.92	0	1.31	-1.11	0	0.92	-0.33	0	1.11	0.11	0	0.33	-1.31	1.43
7	Major Groove Width	12.15	12.37	13.51	12.87	13.58	15.49	14.42	13.51	13.93	14.55	15.49	12.37	12.32	13.93	13.58	12.15
8	Major Groove Depth	9.12	9.41	8.96	8.96	8.67	8.45	8.81	8.96	8.76	8.67	8.45	9.41	9.6	8.76	8.67	9.12
9	Major Groove Size	3.98	3.98	4.7	4.7	3.98	3.98	4.7	4.7	3.26	3.26	3.98	3.98	3.26	3.26	3.98	3.98
10	Major Groove Distance	3.38	3.03	3.36	3.02	3.79	3.38	3.77	3.36	3.4	3.04	3.38	3.03	3.81	3.4	3.79	3.38
11	Minor Groove Width	5.3	6.04	5.19	5.31	4.79	4.62	5.16	5.19	4.71	4.74	4.62	6.04	6.4	4.71	4.79	5.3
12	Minor Groove Depth	9.03	8.79	8.98	8.91	9.09	8.99	9.06	8.98	9.11	8.98	8.99	8.79	9	9.11	9.09	9.03
13	Minor Groove Size	2.98	3.26	3.98	3.26	3.7	3.98	4.7	3.98	2.98	3.26	3.98	3.26	2.7	2.98	3.7	2.98
14	Minor Groove Distance	2.94	4.22	2.79	4.2	3.09	2.8	3.21	2.79	2.95	4.24	2.8	4.22	2.97	2.95	3.09	2.94

http://diprodb.leibniz-fli.de/

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, these are the various information available this database here they have a twists stacking energy, rice, bent and so on right.

Here these are the various dinucleotides, we have the property names. So, you can analyze various features; that means we do not know which feature is important for the flexibility, which features important for the binding affinity right. So, you can use any of the features and try to understand why this is important what properties are important for the binding affinity or if the any mutation is casting the diseases, you have the mutation data here right then you can relates with this decision information so on fine, so this database available in the website. So, these are download able. So, you can download the data right for all the features and you can use this features to understand the different DNA sequences.

Fine; so till away discuss about the database; different types of databases, the collection of databases right and only for the DNA or the where databases we develop for the DNA.

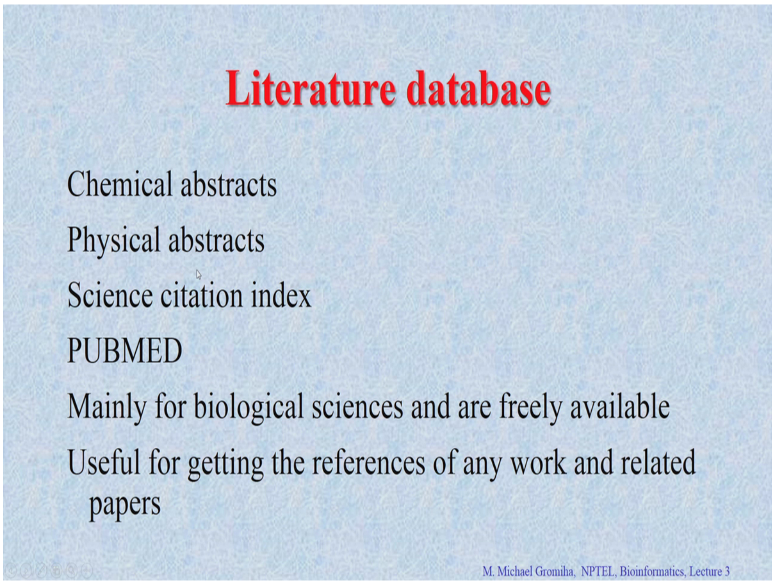
Student: DDBJ

DDBJ, EMBL, GenBank, and this, then nuclei proper database. Now, we will discuss about literature database right what is literature database?

Student: (Refer Time: 15:56).

And discuss data about the published articles, that the earlier days right.

(Refer Slide Time: 16:06)



Literature database

Chemical abstracts

Physical abstracts

Science citation index

PUBMED

Mainly for biological sciences and are freely available

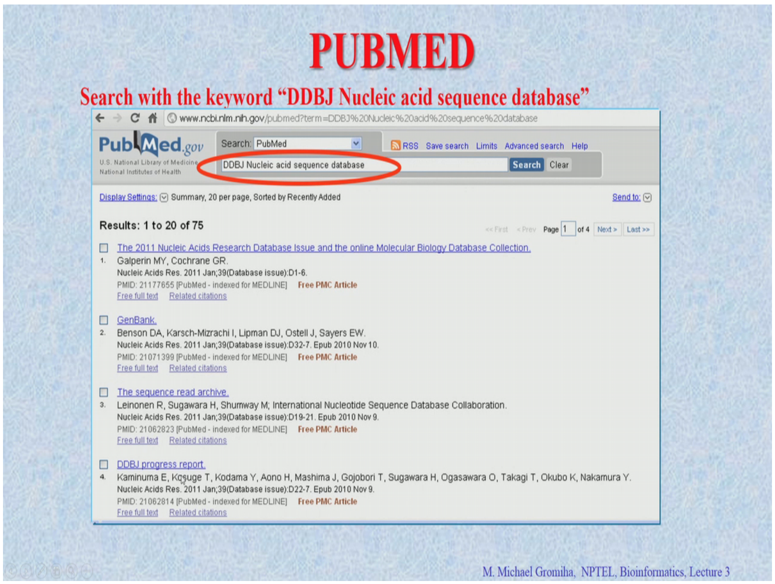
Useful for getting the references of any work and related papers

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

If you (Refer Time: 00:00) because they started to have this some abstracts like chemical abstracts the physical abstracts right. So, in this case this is very big volume very small letters, very difficult worried that. Even I have not sure the publish a number base in the computer readable form. Publish in the chemical abstract and the physical abstracts they are the very famous abstracts. Then we started the science citation index right they started to rank the different journals as the different papers, and see how many citations each article or each general published articles in generals are cited by other others.

Then PUBMED is the widely used database for the live senses right they include mainly a live sense papers, not the physics or chemistry right from several lading generals right with the reference to the all the papers published in the literature ok.

(Refer Slide Time: 16:50)



PUBMED

Search with the keyword "DDBJ Nucleic acid sequence database"

www.ncbi.nlm.nih.gov/pubmed?term=DDBJ%20Nucleic%20acid%20sequence%20database
PubMed.gov
U.S. National Library of Medicine
National Institutes of Health
RSS
Save search
Limits
Advanced search
Help
Search
Clear
Send to:
<< First
< Prev
Next >
Last >>

Search: PubMed

DDBJ Nucleic acid sequence database

Display Settings: Summary, 20 per page, Sorted by Recently Added

Results: 1 to 20 of 75

Page 1 of 4

□ The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection.
1. Galperin MY, Cochrane GR.
Nucleic Acids Res. 2011 Jan;39(Database issue):D1-6.
PMID: 21177655 [PubMed - indexed for MEDLINE] Free PMC Article
Free full text Related citations

□ GenBank.
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW.
Nucleic Acids Res. 2011 Jan;39(Database issue):D32-7. Epub 2010 Nov 10.
PMID: 21071399 [PubMed - indexed for MEDLINE] Free PMC Article
Free full text Related citations

□ The sequence read archive.
3. Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration.
Nucleic Acids Res. 2011 Jan;39(Database issue):D19-21. Epub 2010 Nov 9.
PMID: 21062823 [PubMed - indexed for MEDLINE] Free PMC Article
Free full text Related citations

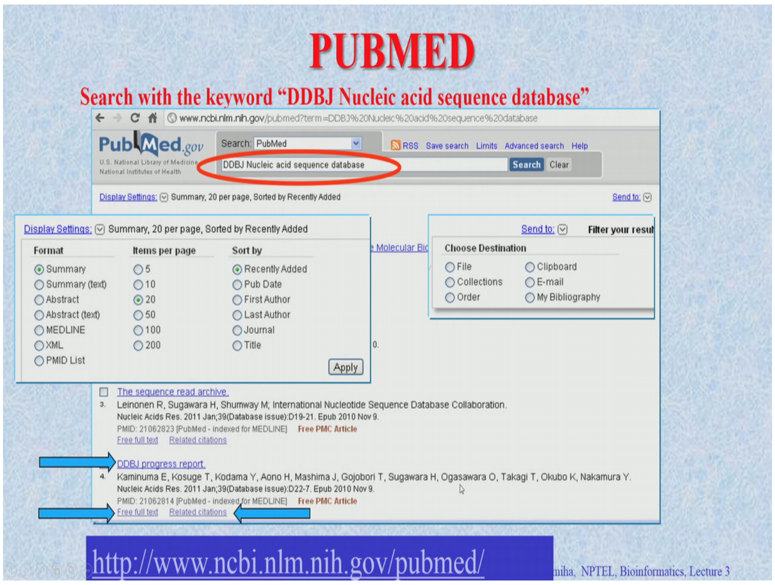
□ DDBJ progress report.
4. Kaminuma E, Kosuge T, Kodama Y, Aono H, Mashima J, Gojobori T, Sugawara H, Ogasawara O, Takagi T, Okubo K, Nakamura Y.
Nucleic Acids Res. 2011 Jan;39(Database issue):D22-7. Epub 2010 Nov 9.
PMID: 21062814 [PubMed - indexed for MEDLINE] Free PMC Article
Free full text Related citations

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, I will explain about the proper because its widely used database and we get almost all the information, there going to the biology and as well as for the medicine. So, if you want to get the information regarding DDBJ nucleic acid sequence database. So, you want to get the article published about this a database. If you search with the keyword right, you will show you the all the data right. From this you can get this is the collection, because this is also includes the DDBJ this is listed here and you can say somewhere DDBJ progress report and various other information right.

They listed other articles also, because they also linked with this DDBJ, this is the real why they get the other articles right.

(Refer Slide Time: 17:28)



PUBMED

Search with the keyword “DDBJ Nucleic acid sequence database”

www.ncbi.nlm.nih.gov/pubmed?term=DDBJ%20Nucleic%20acid%20sequence%20database

Search: PubMed

DDBJ Nucleic acid sequence database

Display Settings: Summary, 20 per page, Sorted by Recently Added

Format	Items per page	Sort by
● Summary	○ 5	● Recently Added
○ Summary (text)	○ 10	○ Pub Date
○ Abstract	● 20	○ First Author
○ Abstract (text)	○ 50	○ Last Author
○ MEDLINE	○ 100	○ Journal
○ XML	○ 200	○ Title
○ PMID List		

Apply

Send to:

Choose Destination

○ File	○ Clipboard
○ Collections	○ E-mail
○ Order	○ My Bibliography

□ The sequence read archive.

3. Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration.
Nucleic Acids Res. 2011 Jan;39(Database issue):D19-21. Epub 2010 Nov 9.
PMID: 21062823 [PubMed - indexed for MEDLINE] Free PMC Article
Free full text Related citations

DDBJ progress report.

4. Kaminuma E, Kosuge T, Kodama Y, Aono H, Mashima J, Gojobori T, Sugawara H, Ogasawara O, Takagi T, Okubo K, Nakamura Y.
Nucleic Acids Res. 2011 Jan;39(Database issue):D22-7. Epub 2010 Nov 9.
PMID: 21062814 [PubMed - indexed for MEDLINE] Free PMC Article
Free full text Related citations

http://www.ncbi.nlm.nih.gov/pubmed/

PubMed.gov
U.S. National Library of Medicine
National Institutes of Health
RSS Save search Limits Advanced search Help
Search Clear
Filter your resul
e Molecular Bio
miha, NPTEL, Bioinformatics, Lecture 3

So, we go DDBJ progress report right. So, then we can get the data. So, when you display the data there are various options; whether they need the format this summary or the abstract or any other formats and how many items are page? 5 or 10 or 20 and whether you can sort whether you need the first other or the general name or the recently added and so on right. So, when you apply this right, then we can get the information; this is what we need. So, they aspect DDBJ nucleic acid sequence database, here we have DDBJ, we will get the data. For if you want to get the full text (Refer Time: 18:03) full text then you get the complete paper, because this is the title.

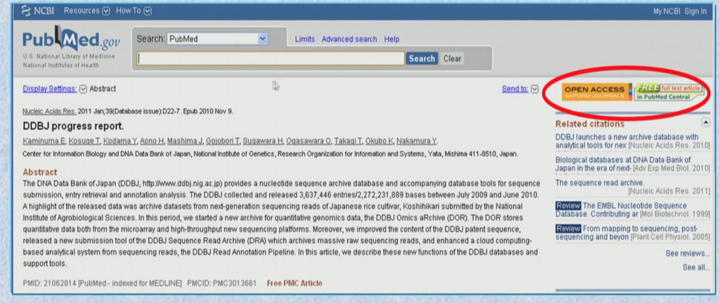
If you click here you will get the abstract and if you click here you will get the full text right, then we are the related citations what is the meaning of related citations?

Student: (Refer Time: 18:17).

The papers which are similar to DDBJ. So, we can get a GenBank, you can get the EMBL also other databases right. So, can get the different entries publish articles, which related to this DDBJ right. So, this is a website for the PUBMED and you can access this website to get the literature database right.

(Refer Slide Time: 18:36)

Abstract


NCBI
Resources
How To
My NCBI
Sign In
PubMed.gov
U.S. National Library of Medicine
National Institutes of Health
Search: PubMed
Limits
Advanced search
Help
Search
Clear
Display Settings: Abstract
Send to:
OPEN ACCESS
OXFORD JOURNALS
FREE full text article in PubMed Central

Nucleic Acids Res. 2011 Jan;39(Database issue):D22-7. Epub 2010 Nov 9.

DDBJ progress report.

Kaminuma E, Kosuge T, Kodama Y, Aono H, Mashima J, Gojobori T, Sugawara H, Ogasawara O, Takagi T, Okubo K, Nakamura Y.

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization for Information and Systems, Yata, Mishima 411-8510, Japan.

Abstract

The DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp) provides a nucleotide sequence archive database and accompanying database tools for sequence submission, entry retrieval and annotation analysis. The DDBJ collected and released 3,637,446 entries/2,272,231,889 bases between July 2009 and June 2010. A highlight of the released data was archive datasets from next-generation sequencing reads of Japanese rice cultivar, Koshihikari submitted by the National Institute of Agrobiological Sciences. In this period, we started a new archive for quantitative genomics data, the DDBJ Omics aRchive (DOR). The DOR stores quantitative data both from the microarray and high-throughput new sequencing platforms. Moreover, we improved the content of the DDBJ patent sequence, released a new submission tool of the DDBJ Sequence Read Archive (DRA) which archives massive raw sequencing reads, and enhanced a cloud computing-based analytical system from sequencing reads, the DDBJ Read Annotation Pipeline. In this article, we describe these new functions of the DDBJ databases and support tools.

PMID: 21062814 [PubMed - indexed for MEDLINE] PMCID: PMC3013661 Free PMC Article

Related citations

DDBJ launches a new archive database with analytical tools for nex [Nucleic Acids Res. 2010]

Biological databases at DNA Data Bank of Japan in the era of next- [Adv Exp Med Biol. 2010]

The sequence read archive. [Nucleic Acids Res. 2011]

Review The EMBL Nucleotide Sequence Database. Contributing ar [Mol Biotechnol. 1999]

Review From mapping to sequencing, post-sequencing and beyon [Plant Cell Physiol. 2005]

See reviews...

See all...

What is Open Access?

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, now you get the click on the DDBJ progress report, you get the abstract and here you can see get this full paper. Either you get this Oxford University process this is the publish in your site or you can get the proper central right what is open access? So, they put is open access right.

What is the meaning of open access?

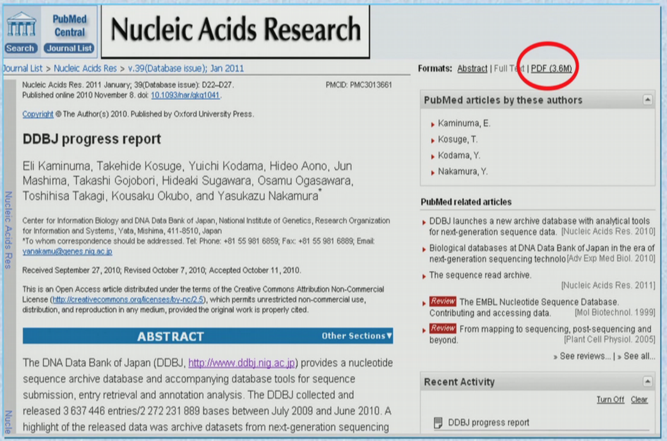
Student: You can use the free earlier.

You can use the free earlier right because to publish and article publish a general issue. So, it will was start of cast, who will bear the cast two options; one option is the subscribers they have to bear the cast, in this case (Refer Time: 19:13) have to pay anything. We have to publish the article when this accepted with the review or in the other procedures publishers to publish, how to earn the money because they we have to subscribe the generalist. When you subscribe you have to pay, though those who want to read this articles they have to pay right. These go they maintain till few years ago. In the open access means the others will play right in this case the read us do not go to pay because others will pay the charges and the tell the publishers to make it offer. So, anybody can access this article.

So, this is call though open access right there are two options, some journals they can only open access, some journals they can only sub by subscription some journals they have the option given to the others. Others can choose whether you want to make it open or not. Where is open you have to pay buts available to everyone, where is not open its not available to everyone. So, only restricted users can read the article right. So, depends on the others they can you make it as open access or not fine, this is open access means you can use it right.

(Refer Slide Time: 20:18)

Full text



Nucleic Acids Research

Journal List > Nucleic Acids Res > v.39(Database issue); Jan 2011

Formats: Abstract | Full Text | PDF (3.6M)

Nucleic Acids Res. 2011 January; 39(Database issue): D22–D27.
Published online 2010 November 8. doi: 10.1093/nar/gkq1041.

PMCID: PMC3013661

Copyright © The Author(s) 2010. Published by Oxford University Press.

DDBJ progress report

Eli Kaminuma, Takehide Kosuge, Yuichi Kodama, Hideo Aono, Jun Mashima, Takashi Gojobori, Hideaki Sugawara, Osamu Ogasawara, Toshihisa Takagi, Kousaku Okubo, and Yasukazu Nakamura*

Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Research Organization for Information and Systems, Yata, Mishima, 411-8510, Japan
*To whom correspondence should be addressed. Tel: Phone: +81 55 981 6859; Fax: +81 55 981 6889; Email: yanakamu@genes.nig.ac.jp

Received September 27, 2010; Revised October 7, 2010; Accepted October 11, 2010.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT

Other Sections▼

The DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp) provides a nucleotide sequence archive database and accompanying database tools for sequence submission, entry retrieval and annotation analysis. The DDBJ collected and released 3 637 446 entries/2 272 231 889 bases between July 2009 and June 2010. A highlight of the released data was archive datasets from next-generation sequencing

PubMed articles by these authors

- Kaminuma, E.
- Kosuge, T.
- Kodama, Y.
- Nakamura, Y.

PubMed related articles

- DDBJ launches a new archive database with analytical tools for next-generation sequence data. [Nucleic Acids Res. 2010]
- Biological databases at DNA Data Bank of Japan in the era of next-generation sequencing technolo [Adv Exp Med Biol. 2010]
- The sequence read archive. [Nucleic Acids Res. 2011]
- Review The EMBL Nucleotide Sequence Database. Contributing and accessing data. [Mol Biotechnol. 1999]
- Review From mapping to sequencing, post-sequencing and beyond. [Plant Cell Physiol. 2005]

» See reviews... | » See all...

Recent Activity

Turn Off Clear

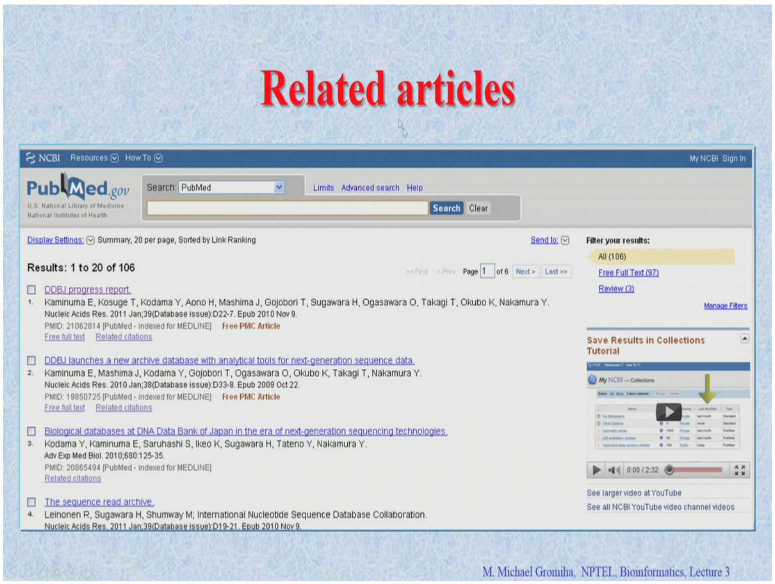
DDBJ progress report

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, you can go the full text right now this the PDF format right. So, what is PDF format? Portable dockman format right in this case you can get the full paper you can read in the format, then you can like a print right.

Here we can read it online whether if figure if PDF format, this just look at of print like (Refer Time: 00:00) in general article right you can do that.

(Refer Slide Time: 20:45)



Related articles

Display Settings: Summary, 20 per page, Sorted by Link Ranking

Send to:

Results: 1 to 20 of 106

<< First < Prev Page 1 of 6 Next > Last >>

□ DDBJ progress report.
1. Kaminuma E, Kosuge T, Kodama Y, Aono H, Mashima J, Gojobori T, Sugawara H, Ogasawara O, Takagi T, Okubo K, Nakamura Y.
Nucleic Acids Res. 2011 Jan;39(Database issue):D22-7. Epub 2010 Nov 9.
PMID: 21062814 [PubMed - indexed for MEDLINE] Free PMC Article
Free full text Related citations

□ DDBJ launches a new archive database with analytical tools for next-generation sequence data.
2. Kaminuma E, Mashima J, Kodama Y, Gojobori T, Ogasawara O, Okubo K, Takagi T, Nakamura Y.
Nucleic Acids Res. 2010 Jan;38(Database issue):D33-8. Epub 2009 Oct 22.
PMID: 19850725 [PubMed - indexed for MEDLINE] Free PMC Article
Free full text Related citations

□ Biological databases at DNA Data Bank of Japan in the era of next-generation sequencing technologies.
3. Kodama Y, Kaminuma E, Saruhashi S, Ikeo K, Sugawara H, Tateno Y, Nakamura Y.
Adv Exp Med Biol. 2010;680:125-35.
PMID: 20865494 [PubMed - indexed for MEDLINE]
Related citations

□ The sequence read archive.
4. Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration.
Nucleic Acids Res. 2011 Jan;39(Database issue):D19-21. Epub 2010 Nov 9.

Filter your results:
All (106)
Free Full Text (97)
Review (3)
Manage Filters

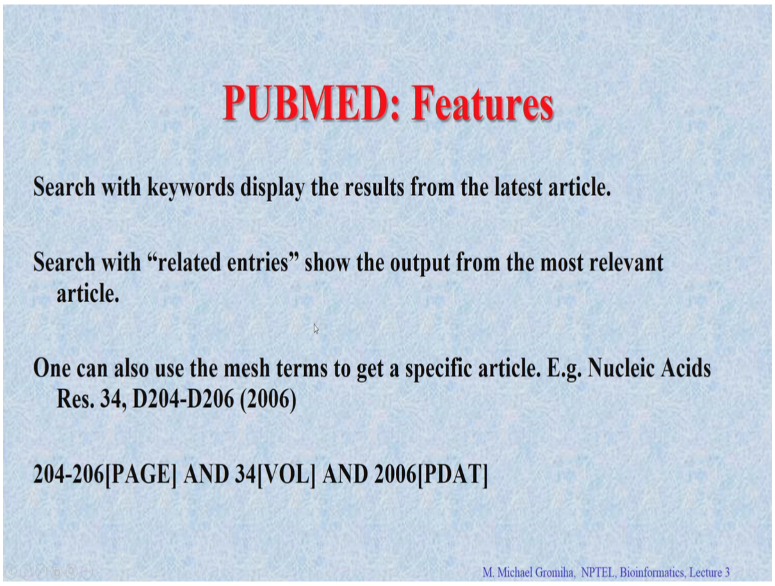
Save Results in Collections Tutorial

See larger video at YouTube
See all NCBI YouTube video channel videos

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, now it is related article. So, if you look into these related click on related articles. So, it show the all the articles, which are relevant to your search. So, for example, this is your search right the man article, if you go to related citations these are the other papers which publish in the literature which related to the article which our query article right you get all the related articles you can get that fine.

(Refer Slide Time: 21:09)



PUBMED: Features

Search with keywords display the results from the latest article.

Search with "related entries" show the output from the most relevant article.

One can also use the mesh terms to get a specific article. E.g. Nucleic Acids Res. 34, D204-D206 (2006)

204-206[PAGE] AND 34[VOL] AND 2006[PDAT]

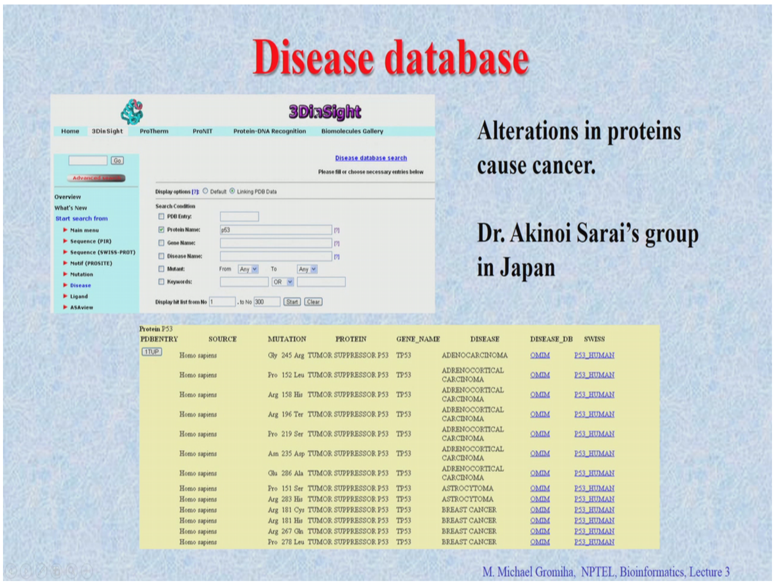
M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, what is the various features of the PUBMED. So, you can search with the any keywords, when get data for the very latest article right and search with the related entries to give the articles which are relevant to the related articles, and also you can some mesh terms to get any of these articles. So, can you they get with the page numbers, you can with the with the general name, you can search with others and so on; so main difference between currently available PUBMED and previous years well.

Previous days, unless we get the article we cannot site, but currently if you type any keyword you get all the articles. So, main problem is currently if you write a man script just you type the keyword, get all the papers, they do not data full paper, some thermal people they do not get the abstract also right. They do not know the others, just they take for topmost 5 15 and site everything.

The earlier days the original others cut proper credit, because they do not get all the papers only the important papers we have to send a request and they will send the refract, then we read, then we understand, then only be site number of citation if this is very less, because all of papers almost state by the others, but currently they said 100s, but they do not read in other papers. That is a difference between the advantage of this PUBMED right very easily they others use the get the information fine.

(Refer Slide Time: 22:42)



Disease database

Disease database search

Please fill or choose necessary entries below

Display options [?]: ○ Default ● Linking PDB Data

Search Condition

□ PDB Entry:

☑ Protein Name: p53 [?]

□ Gene Name: [?]

□ Disease Name: [?]

□ Mutant: From Any To Any

□ Keywords: OR

Display hit list from No 1 . to No 300

Protein P53

PDBENTRY	SOURCE	MUTATION	PROTEIN	GENE_NAME	DISEASE	DISEASE_DB	SWISS
1TUP	Homo sapiens	Gly 245 Arg	TUMOR SUPPRESSOR P53	TP53	ADENOCARCINOMA	OMIM	P53_HUMAN
	Homo sapiens	Pro 152 Leu	TUMOR SUPPRESSOR P53	TP53	ADRENOCORTICAL CARCINOMA	OMIM	P53_HUMAN
	Homo sapiens	Arg 158 His	TUMOR SUPPRESSOR P53	TP53	ADRENOCORTICAL CARCINOMA	OMIM	P53_HUMAN
	Homo sapiens	Arg 196 Ter	TUMOR SUPPRESSOR P53	TP53	ADRENOCORTICAL CARCINOMA	OMIM	P53_HUMAN
	Homo sapiens	Pro 219 Ser	TUMOR SUPPRESSOR P53	TP53	ADRENOCORTICAL CARCINOMA	OMIM	P53_HUMAN
	Homo sapiens	Asn 235 Asp	TUMOR SUPPRESSOR P53	TP53	ADRENOCORTICAL CARCINOMA	OMIM	P53_HUMAN
	Homo sapiens	Glu 286 Ala	TUMOR SUPPRESSOR P53	TP53	ADRENOCORTICAL CARCINOMA	OMIM	P53_HUMAN
	Homo sapiens	Pro 151 Ser	TUMOR SUPPRESSOR P53	TP53	ASTROCYTOMA	OMIM	P53_HUMAN
	Homo sapiens	Arg 283 His	TUMOR SUPPRESSOR P53	TP53	ASTROCYTOMA	OMIM	P53_HUMAN
	Homo sapiens	Arg 181 Cys	TUMOR SUPPRESSOR P53	TP53	BREAST CANCER	OMIM	P53_HUMAN
	Homo sapiens	Arg 181 His	TUMOR SUPPRESSOR P53	TP53	BREAST CANCER	OMIM	P53_HUMAN
	Homo sapiens	Arg 267 Gln	TUMOR SUPPRESSOR P53	TP53	BREAST CANCER	OMIM	P53_HUMAN
	Homo sapiens	Pro 278 Leu	TUMOR SUPPRESSOR P53	TP53	BREAST CANCER	OMIM	P53_HUMAN

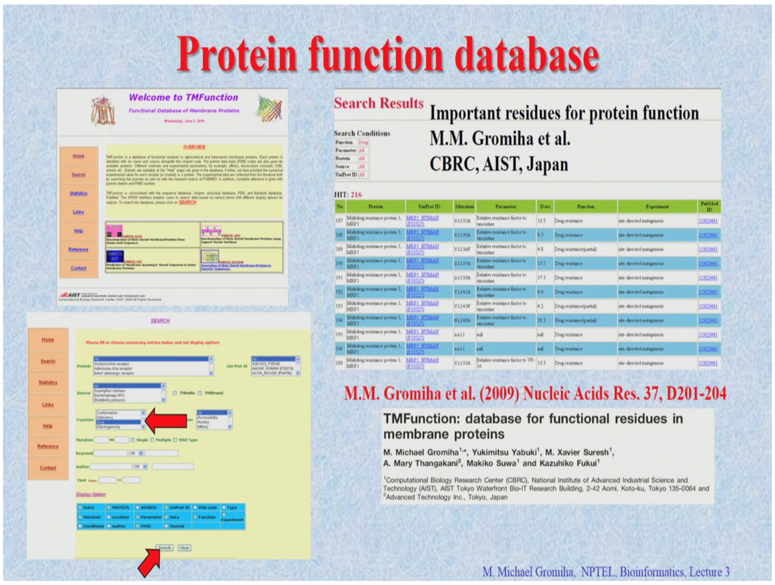
Alterations in proteins cause cancer.

Dr. Akinoi Sarai's group in Japan

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

So, this is another database called disease database right this is the 3D inside this database which contains the various information regarding the proteins sequence and they there are thermodynamics and (Refer Time: 22:49) and so on they are such added some of the information regarding the diseases. So, how are the mutations, they change the of the mutations right because any of these diseases. They obtain the data from the other databases and then develop the database for the diseases.

(Refer Slide Time: 23:06)



Protein function database

Welcome to TMFunction

Functional Database of Membrane Proteins

Wednesday, June 2, 2010

OVERVIEW

TMFunction is a database of functional residues in alpha-helical and beta-barrel membrane proteins. Each protein is identified with its name and source alongwith the Uniprot code. The protein data bank (PDB) codes are also given for available proteins. Different methods and experimental parameters, for example, affinity, dissociation constant, IC50, activity etc. (Details are available at the "Help" page) are given in the database. Futher, we have provided the numerical experimental value for each residue (or mutant) in a protein. The experimental data are collected from the literature both by searching the journals as well as with the keyword search at PUBMED. In addition, complete reference is given with journal citation and PMID number.

TMFunction is cross-linked with the sequence database, Uniprot, structural database, PDB, and literature database, PubMed. The WWW interface enables users to search data based on various terms with different display options for outputs. To search the database, please click on SEARCH

TMBETA-DISC Discrimination of Beta-Barrel Membrane Proteins from Amino Acid Sequence.

TMBETA-SVM Discrimination of Beta-Barrel Membrane Proteins using Support Vector Machines

TMBETA-NET Prediction of Membrane Spanning b-Strand Segments in Outer Membrane Proteins

TMBETA-GENOME Annotation of Beta-Barrel Membrane Proteins in Genomic Sequences.

Computational Biology Research Center, AIST, 2008 All Rights Reserved.

SEARCH

Please fill or choose necessary entries below and set display options

Protein: All / Acetylcholine receptor / Adenosine A2a receptor / beta1-adrenergic receptor

Uni Prot ID: All / A3HJG3_PSEAE / AA2AR_HUMAN (P29274) / ACHA_MOUSE (P04756)

Source: All / Aspergillus nidulans / Bacteriophage M13 / Bordetella pertussis   □ TMHelix □ TMStrand

Function: Conformation / Deficiency / Drug / Electrogenicity

Parameter: All / Accessibility / Activity / Affinity

Mutation ___ to ___ □ Single □ Multiple □ Wild Type

Keyword ___ OR ___

Author ___ OR ___

Year From ___ to ___

Display Option

☑ Entry	☑ PROTEIN	□ SOURCE	☑ UniProt ID	□ PDB code	□ Type
☑ Mutation	□ Location	☑ Parameter	☑ Data	☑ Function	☑ Experiment
□ Conditions	□ Author	☑ PMID	□ Journal		

Search Results

Important residues for protein function
M.M. Gromiha et al.
CBRC, AIST, Japan

Search Conditions

Function	Drug
Parameter	All
Protein	All
Source	All
UniProt ID	All

HIT: 216

No.	Protein	UniProt ID	Mutation	Parameter	Data	Function	Experiment	PubMed ID
187	Multidrug resistance protein 1; MRP1	MRP1_HUMAN (P33527)	S1233A	Relative resistance factor to vincristine	12.5	Drug resistance	site-directed mutagenesis	11925441
188	Multidrug resistance protein 1; MRP1	MRP1_HUMAN (P33527)	S1235A	Relative resistance factor to vincristine	9.3	Drug resistance	site-directed mutagenesis	11925441
189	Multidrug resistance protein 1; MRP1	MRP1_HUMAN (P33527)	Y1236F	Relative resistance factor to vincristine	4.8	Drug resistance(partial)	site-directed mutagenesis	11925441
190	Multidrug resistance protein 1; MRP1	MRP1_HUMAN (P33527)	S1237A	Relative resistance factor to vincristine	13.7	Drug resistance	site-directed mutagenesis	11925441
191	Multidrug resistance protein 1; MRP1	MRP1_HUMAN (P33527)	Q1239A	Relative resistance factor to vincristine	17.3	Drug resistance	site-directed mutagenesis	11925441
192	Multidrug resistance protein 1; MRP1	MRP1_HUMAN (P33527)	T1241A	Relative resistance factor to vincristine	4.4	Drug resistance	site-directed mutagenesis	11925441
193	Multidrug resistance protein 1; MRP1	MRP1_HUMAN (P33527)	Y1243F	Relative resistance factor to vincristine	4.2	Drug resistance(partial)	site-directed mutagenesis	11925441
194	Multidrug resistance protein 1; MRP1	MRP1_HUMAN (P33527)	N1245A	Relative resistance factor to vincristine	29.3	Drug resistance(partial)	site-directed mutagenesis	11925441
195	Multidrug resistance protein 1; MRP1	MRP1_HUMAN (P33527)	null	null	null	Drug resistance	site-directed mutagenesis	11925441
196	Multidrug resistance protein 1; MRP1	MRP1_HUMAN (P33527)	null	null	null	Drug resistance	site-directed mutagenesis	11925441
198	Multidrug resistance protein 1; MRP1	MRP1_HUMAN (P33527)	S1233A	Relative resistance factor to VP-16	13.3	Drug resistance	site-directed mutagenesis	11925441

M.M. Gromiha et al. (2009) Nucleic Acids Res. 37, D201-204

TMFunction: database for functional residues in membrane proteins

M. Michael Gromiha[1,*], Yukimitsu Yabuki[1], M. Xavier Suresh[1], A. Mary Thangakani[2], Makiko Suwa[1] and Kazuhiko Fukui[1]

[1]Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064 and [2]Advanced Technology Inc., Tokyo, Japan

M. Michael Gromiha, NPTEL, Bioinformatics, Lecture 3

This is another database for the protein function database, we developed few years ago. So, here it will tell you these important residues which are perform in different functions in membrane proteins.

So, you have a search options and we have the display options and we have go they have get the results. For any mutations you can say whether (Refer Time: 23:23) and. So, what is the function of the protein as well as how many specific mutation, which alters this specific function or not; likewise if you do little literature the very databases based on the protein sequence, protein structure and the thermodynamics, diseases and the literature and so on. So, look into this nucleic acid research website, we will get the information regarding the all the information so that you can use it.

Thanks for your attention.