Dept. of Computer Science and Engineering
University of Rajshahi
www.ru.ac.bd

Dr. Shamim Ahmad

Databases

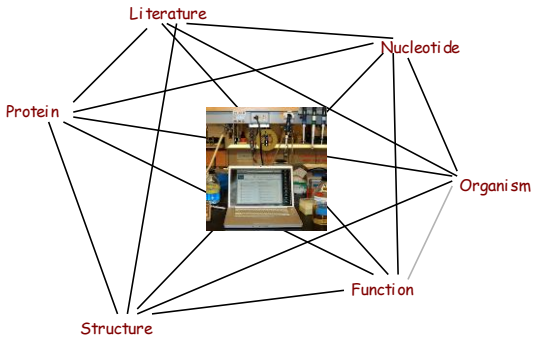
Organizing information in the post-genomic era

The rise of bioinformatics

What is a database?

Which databases are important for molecular cell biology research?

How is information processed in databases?



Biological databases use different organizing principles

Hyperlinks connect records in different databases

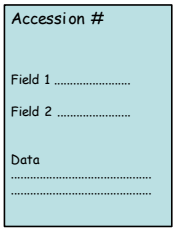
Databases are organized collections of information

Information is stored in records

Databases assign each record a unique accession number using their own numbering system

Fields are used to cross-reference the data. Records can be searched by fields.

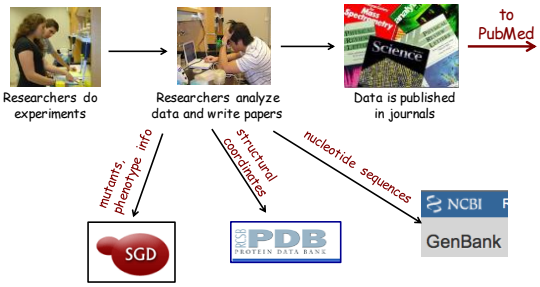
Data is entered in the record using a defined format



Bioinformaticians work with computer scientists to set up the database structure

Curators review and link records within and between databases

The information in databases ultimately derives from experimental data



Curators will process the submissions and link entries in different databases

Database Examples in Bioinformatics

- Primary (archival)
  - GenBank/EMBL/DDBJ (seqs)
  - PDB (protein structures)
  - Medline (literature)
  - IMEx databases (protein interactions)
- Secondary (curated)
  - RefSeq (seqs)
  - UniProt - SwissProt (seqs)
  - Taxon (taxonomy)
  - PROSITE (binding sites)
  - OMIM (genetics literature/reviews)
  - IMEx databases (protein interactions)

Sequence Databases

- DNA
  - NCBI: GenBank -> RefSeq
  - EBI: EMBL
- Protein
  - NCBI: GenPept
  - EBI: UniProt: TrEMBL -> UniProt: Swiss-Prot

TrEMBL= "translated EMBL"

National Center for Biotechnology Information www.ncbi.nlm.nih.gov

European Bioinformatics Institute www.ebi.ac.uk

GenBank

- Direct submissions from individual laboratories
- Bulk submissions from large-scale sequencing centers.

Direct submissions to GenBank

- **BankIt**, which is a Web-based form
- **Sequin** stand-alone submission program

Upon receipt of a sequence submission

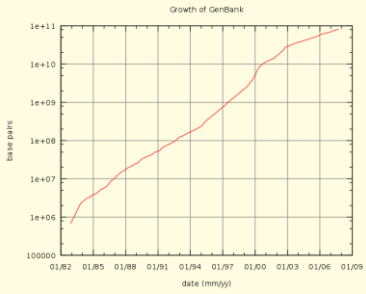
- GenBank staff examines the originality of the data
- Assigns an accession number to the sequence
- Performs quality assurance checks

GenBank

- **GenBank has become an important database for research in biological fields**
- **Ii has grown in recent years at an exponential rate by doubling roughly every 18 months**
- Entries are retrievable by Entrez
- Downloadable by FTP

Bulk submissions

- Expressed Sequence Tag (EST)
- Sequence-tagged site (STS)
- Genome Survey Sequence (GSS)
- High-Throughput Genome Sequence (HTGS)



Growth in GenBank base pairs, 1982 to 2007, on a semi-log scale

GenBank

An expressed sequence tag (EST) is a short sub-sequence of a cDNA sequence

A **sequence-tagged site** (or **STS**) is a short (200 to 500 base pair) DNA sequence that has a single occurrence in the genome

Genome Survey Sequences (GSS) are nucleotide sequences similar to EST's that the only difference is that most of them are genomic in origin, rather than mRNA

Complementary DNA

- **Complementary DNA (cDNA) is DNA synthesized from a single stranded RNA (e.g., messenger RNA (mRNA)**
- **DNA is derived from mRNA, so it contains only exons, with no introns.**
- **cDNA is also produced naturally by retroviruses (such as HIV-1, HIV-2, simian immunodeficiency virus, etc**

Complementary DNA

- **cDNA is often used to clone eukaryotic genes in prokaryotes.**
- **When scientists want to express a specific protein in a cell that does not normally express that protein, they will transfer the cDNA that codes for the protein to the recipient cell.**
- **cDNA is also produced naturally by retroviruses (such as HIV-1, HIV-2, simian immunodeficiency virus, etc**

Top organisms in GenBank (Release 191)[9]

Organism	base pairs
Homo sapiens	16,310,774,187
Mus musculus	9,974,977,889
Rattus norvegicus	6,521,253,272
Bos taurus	5,386,258,455
Zea mays	5,062,731,057
Sus scrofa	4,887,861,860
Danio rerio	3,120,857,462
Strongylocentrotus purpuratus	1,435,236,534
Macaca mulatta	1,256,203,101
Oryza sativa Japonica Group	1,255,686,573
Nicotiana tabacum	1,197,357,811
Xenopus (Silurana) tropicalis	1,249,938,611
Drosophila melanogaster	1,119,965,220
Pan troglodytes	1,008,323,292
Arabidopsis thaliana	1,144,226,616
Canis lupus familiaris	951,238,343
Vitis vinifera	999,010,073
Gallus gallus	899,631,338
Glycine max	906,638,854
Triticum aestivum	898,689,329

```
ID   AF115338   standard; DNA; PRO; 591 BP.
AC   AF115338;
SV   AF115338.1
DT   03-JUN-1999 (Rel. 59, Created)
DT   23-AUG-1999 (Rel. 60, Last updated, Version 2)
DE   Pseudomonas fluorescens ECF sigma factor SigX (sigX) gene, complete cds.
KW   .
OS   Pseudomonas fluorescens
OC   Bacteria; Proteobacteria; gamma subdivision; Pseudomonadaceae; Pseudomonas.
RN   [1]
RP   1-591
RX   MEDLINE; 99369842.
RA   Brinkman F.S., Schoofs G., Hancock R.E., De Mot R.;
RT   "Influence of a putative ECF sigma factor on expression of the major outer
RT   membrane protein, OprF, in Pseudomonas aeruginosa and Pseudomonas
RT   fluorescens";
RL   J. Bacteriol. 181(16):4746-4754(1999).
RN   [2]
RP   1-591
RA   De Mot R.;
RT   ;
RL   Submitted (04-DEC-1998) to the EMBL/GenBank/DDBJ databases.
RL   F.A. Janssens Laboratory of Genetics, Applied Plant Sciences, K.
RL   Mercierlaan 92, Heverlee B-3001, Belgium
DR   SPTREMBL; Q9X4L7; Q9X4L7.
FH   Key             Location/Qualifiers
FH
FT   source 1..591
FT                   /db_xref="taxon:294"
FT                   /organism="Pseudomonas fluorescens"
FT                   /strain="M114"
FT   CDS 1..591
FT                   /codon_start=1
FT                   /db_xref="SPTREMBL:Q9X4L7"
FT                   /transl_table=11
FT                   /gene="sigX"
FT                   /product="ECF sigma factor SigX"
FT                   /protein_id="AAD34329.1"
FT                   /translation="MNKAQTLSTRYDPRELSDEELVARSHTELFHVTRAYEELMRRYQR
FT                   TLFNVCARYLGNDRDADDVCQEVMLKVLYGLKNLEGKSKFKTWLYSITYNECITQYRKK
FT                   RRKRRLMDALSLDPLEEASEEKALQPEEKGGLDRWLVYVNPIDRGILVLRFVAELEFQE
FT                   IADIMHMGLSATKMRYKRALDKLREKFAGETET"
SQ   Sequence 591 BP; 157 A; 133 C; 170 G; 131 T; 0 other;
     atgaataaag cccaaacgct atccacgcgc tacgaccccc gcgagctctc tgatgaggag        60
     ttggtcgcgc gctcgcatac cgagcttttt cacgtaacgc gcgcctatga agaactgatg       120
     cggcgttacc agcgaacatt atttaacgtt tgtgcgagat atcttgggaa cgatcgcgac       180
     gcagacgatg tctgtcagga agtcatgttg aaggtgctgt atggcctgaa gaacctcgag       240
     gggaaatcga agttcaaaac gtggctctac agcatcacgt acaacgaatg tattacgcag       300
     tatcggaagg aacggcgaaa gcgtcgcttg atggacgcat tgagtcttga ccccctcgag       360
     gaagcgtccg aagaaaaggc gcttcaaccc gaggagaagg gcgggcttga tcgctggctg       420
     gtgtatgtga acccgattga ccgtggaatt ctggtgcttc gatttgtcgc agagctggaa       480
     tttcaggaga tcgcagacat catgcacatg ggtttgagtg cgacaaaaat gcgttacaaa       540
     cgtgctctag ataaattgcg tgagaaattt gcaggcgaga ctgaaactta g                591
//
```

GenBank Flat File

Header
- Title
- Taxonomy
- Citation

Features (AA seq)

DNA Sequence


```
ID   AF115338   standard; DNA; PRO; 591 BP.
AC   AF115338;
SV   AF115338.1
DT   03-JUN-1999 (Rel. 59, Created)
DT   23-AUG-1999 (Rel. 60, Last updated, Version 2)
DE   Pseudomonas fluorescens ECF sigma factor SigX (sigX) gene, complete cds.
KW   .
OS   Pseudomonas fluorescens
OC   Bacteria; Proteobacteria; gamma subdivision; Pseudomonadaceae; Pseudomonas.
RN   [1]
RP   1-591
RX   MEDLINE; 99369842.
RA   Brinkman F.S., Schoofs G., Hancock R.E., De Mot R.;
RT   "Influence of a putative ECF sigma factor on expression of the major outer
RT   membrane protein, OprF, in Pseudomonas aeruginosa and Pseudomonas
RT   fluorescens";
RL   J. Bacteriol. 181(16):4746-4754(1999).
RN   [2]
RP   1-591
RA   De Mot R.;
RT   ;
RL   Submitted (04-DEC-1998) to the EMBL/GenBank/DDBJ databases.
RL   F.A. Janssens Laboratory of Genetics, Applied Plant Sciences, K.
RL   Mercierlaan 92, Heverlee B-3001, Belgium
DR   SPTREMBL; Q9X4L7; Q9X4L7.
FH   Key             Location/Qualifiers
FH
FT   source 1..591
FT                   /db_xref="taxon:294"
FT                   /organism="Pseudomonas fluorescens"
FT                   /strain="M114"
FT   CDS 1..591
FT                   /codon_start=1
FT                   /db_xref="SPTREMBL:Q9X4L7"
FT                   /transl_table=11
FT                   /gene="sigX"
FT                   /product="ECF sigma factor SigX"
FT                   /protein_id="AAD34329.1"
FT                   /translation="MNKAQTLSTRYDPRELSDEELVARSHTELFHVTRAYEELMRRYQR
FT                   TLFNVCARYLGNDRDADDVCQEVMLKVLYGLKNLEGKSKFKTWLYSITYNECITQYRKK
FT                   RRKRRLMDALSLDPLEEASEEKALQPEEKGGLDRWLVYVNPIDRGILVLRFVAELEFQE
FT                   IADIMHMGLSATKMRYKRALDKLREKFAGETET"
SQ   Sequence 591 BP; 157 A; 133 C; 170 G; 131 T; 0 other;
     atgaataaag cccaaacgct atccacgcgc tacgaccccc gcgagctctc tgatgaggag        60
     ttggtcgcgc gctcgcatac cgagcttttt cacgtaacgc gcgcctatga agaactgatg       120
     cggcgttacc agcgaacatt atttaacgtt tgtgcgagat atcttgggaa cgatcgcgac       180
     gcagacgatg tctgtcagga agtcatgttg aaggtgctgt atggcctgaa gaacctcgag       240
     gggaaatcga agttcaaaac gtggctctac agcatcacgt acaacgaatg tattacgcag       300
     tatcggaagg aacggcgaaa gcgtcgcttg atggacgcat tgagtcttga ccccctcgag       360
     gaagcgtccg aagaaaaggc gcttcaaccc gaggagaagg gcgggcttga tcgctggctg       420
     gtgtatgtga acccgattga ccgtggaatt ctggtgcttc gatttgtcgc agagctggaa       480
     tttcaggaga tcgcagacat catgcacatg ggtttgagtg cgacaaaaat gcgttacaaa       540
     cgtgctctag ataaattgcg tgagaaattt gcaggcgaga ctgaaactta g                591
//
```


EMBL Flat File

Header
- Title
- Taxonomy
- Citation

Features (AA seq)

DNA Sequence

UniProt

The UniProt consortium

European Bioinformatics Institute (EBI), UK
Swiss Institute of Bioinformatics (SIB) , Switzerland
Protein Information Resource (PIR), USA

- EBI and SIB
  - Swiss-Prot and TrEMBL databases (**European Molecular Biology Laboratory**)
- PIR
  - Protein Sequence Database (PIR-PSD)

UniProt

- **UniProtKB**
  - **Swiss-Prot**
  - **TrEMBL)**
- **UniParc**
- **UniRef.**

UniProt

- **UniProtKB**
  - **Swiss-Prot (reviewed, manually annotated entries) 1.93 B Amino acid, 1914.3**
  - **TrEMBL (unreviewed, automatically annotated entries) 14 M Amino Acid, 1914.3**

UniProt

- **UniProtKB (UniProt Knowledgebase)**
  - Information extracted from scientific literature and biocurator-evaluated computational analysi
  - To provide all known relevant information about a particular protein

UniProtKB (Annotation)

- **Protein and gene names**
- **Function**
- **Enzyme-specific**
- **Subcellular location**
- **Protein-protein interactions**
- **Pattern of expression**
- **Ion-, substrate- and cofactor-binding sites**
- **Protein variant forms produced by natural genetic variation, RNA editing,**

UniProt: Swiss-Prot – An example of curated, reviewed annotation

- Incorporates:
  - ✓Function of the protein
  - ✓Subcellular localization of protein
  - ✓Post-translational modification
  - ✓Domains and sites
  - ✓Secondary structure
  - ✓Quaternary structure
  - ✓Similarities to other proteins
  - ✓Diseases associated with deficiencies in the protein
  - ✓Sequence conflicts, variants, etc.

24

Swiss-Prot

- Reliable protein sequences associated
  - With a high level of annotation
  - Description of the function of a protein
  - Its domain structure
  - Post-translational modifications
  - Variants
  - Minimal level of redundancy
  - High level of integration with other databases.

TrEMBL (Translated EMBL Nucleotide Sequence Data Library)

was created to provide automated annotations for those proteins not in Swiss-Prot

UniProt Archive (UniParc)

- It is a **comprehensive** and **non-redundant database**,
- **Proteins may exist in several different source databases**, and in **multiple copies** in the same database
- Information extracted from **scientific literature** and biocurator-evaluated **computational analysi**
- Each sequence is given a stable and unique identifier (UPI)

UniProt Archive (UniParc)

- INSDC **EMBL-Bank/DDBJ/GenBank** nucleotide sequence databases
- Ensembl
- European Patent Office **(EPO)**
- **FlyBase:** the primary repository of genetic and molecular data for the insect family Drosophilidae(FlyBase)
- H-Invitational Database **(H-Inv)**
- International Protein Index **(IPI)**
- Japan Patent Office (JPO)

Explain Post-Translational Modification

UniRef

The **UniProt Reference** Clusters (UniRef) consist of three databases

- **Protein sequences** from UniProtKB
- **Selected UniParc** records

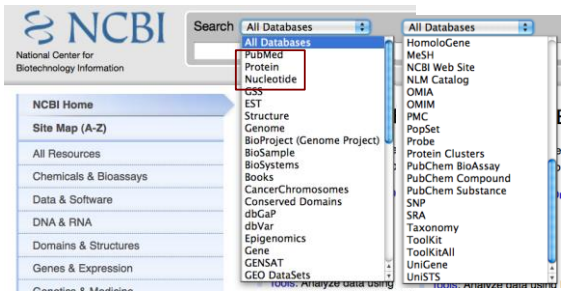
The Protein Data Bank(PDB)

- **Crystallographic database** for the three-dimensional structural data of large biological molecules,
  - Proteins
  - Nucleic acids.
- **The data, typically obtained by**
  - X-ray Crystallography
  - NMR spectroscopy,
  - Cryo-electron microscopy

The Protein Data Bank(PDB)

Experimental Method	Proteins	Nucleic Acids	Protein/Nudeic Acid complexes	Other	Total
X-ray diffraction	106595	1820	5471	4	113890
NMR	10296	1190	241	8	11735
Electron microscopy	1021	30	367	0	1418
Hybrid	99	3	2	1	105
Other	181	4	6	13	204
Total:	118192	3047	6087	26	127352

NCBI maintains both primary and derivative databases
We'll look at three of them



PubMed is the premier literature database in the world

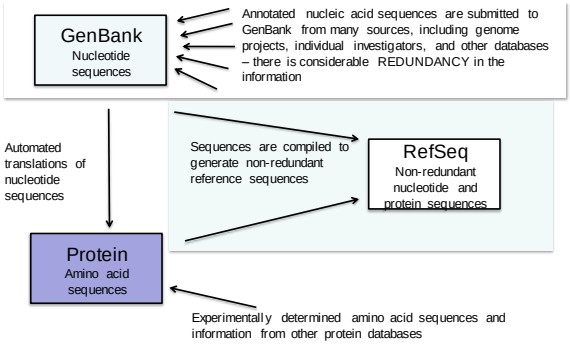
Questions for today:

What is a database?

Which databases are important for molecular cell biology research?

How is information processed in databases?

Curators are responsible for data flow between the NCBI databases

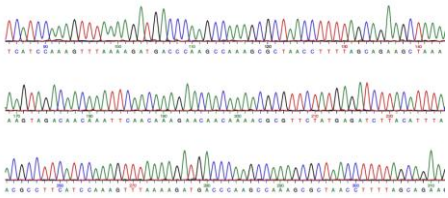


Most records in the Protein database have been derived by automated translation of nucleotide sequences

On a larger scale: Genome projects have produced the reference sequences in nucleotide databases

(robots and computers do much of the work)

1. Pieces of chromosomal DNA are sequenced, each ~1000 bp long



S. cerevisiae genome is ~12 Mbp – how many reads would be necessary to cover each base pair in the genome once?

3. Chromosomal sequences are analyzed for the presence of potential transcripts (open reading frames; ORFs)

ORFs are characterized by an under-representation of stop codons

ORF-finding computer algorithms look for sequences that
- begin with a methionine
- **methionine is separated from a stop codon in the same reading frame by a large number of amino acids** (often 100, equiv. to 300bp)

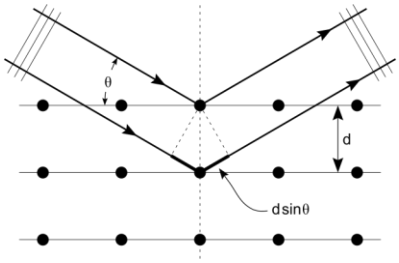
GenBank NM_######## records are predicted ORFs

4. Protein sequences are computationally predicted from ORF sequences

GenBank NP_####### records

X-ray crystallography

For determining the atomic and molecular structure of a crystal, in which the crystalline atoms cause a beam of incident X-rays to diffract into many specific directions



Nuclear magnetic resonance spectroscopy of proteins (NMR)

- It is a field of structural biology
- NMR spectroscopy is used
- To obtain information about the structure and dynamics of
  - proteins, and also nucleic acids,

Entrez
Global Query Cross-Database Search System
CDS
Coding sequence;

The locus name:
Tthe first three characters usually designated the organism
Accession Number [ACCN]
Sequence Length
Sequence Length [SLEN] 2500:2600[SLEN]
Molecule Type
The type of molecule that was sequenced. In this example, the molecule type is DNA
Properties [PROP]
biomol_genomic, biomol_mRNA {PROP}
GenBank Division
gbdiv_pri, gbdiv_est, Properties [PROP]
human[ORGN] NOT gbdiv_est[PROP]

1) To retrieve records within a range of lengths, use the colon as the range operator, e.g., 2500:2600[SLEN].
2) To retrieve all sequences shorter than a certain number, use 2 as the lower bound, e.g., 2:100[SLEN].
3) To retrieve all sequences longer than a certain number, use a series of 9's as the upper bound, e.g., 325000:99999999[SLEN].

Molecule Type
The type of molecule that was sequenced. In this example, the molecule type is DNA

- genomic DNA
- genomic RNA
- precursor RNA
- mRNA (cDNA)
- ribosomal RNA
- transfer RNA,
- small nuclear RNA
- small cytoplasmic RNA.

Modification Date
Modification Date [MDAT]
To retrieve records modified between two dates, 1999/07/25:1999/07/31[MDAT].
DEFINITION
Brief description of sequence
Coding region (CDS),
Word [TITL]
ACCESSION
The unique identifier for a sequence record
VERSION
If there is any change to the sequence data (even a single base), the version number will be increased, e.g., U12345.1 ? U12345.2,
KEYWORDS
Word or phrase describing the sequence
SOURCE
Free-format information including an abbreviated form of the organism name, sometimes followed by a molecule type.

Organism
The formal scientific name for the source organism
REFERENCE
Publications by the authors
AUTHORS
List of authors in the order in which they appear in the cited article
TITLE
Title of the published work or tentative title of an unpublished work.
< symbol indicates partial on the 5' end. Example: **<1..206**

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

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

FASTA format

- It is a text-based format for representing
  - Nucleotide sequences
  - Peptide sequences
- Nucleotides or amino acids are represented using single-letter codes
- The format also allows for sequence names and comments to precede the sequences.

- ❑ The simplicity of FASTA format
- ❑ Makes it easy to manipulate and parse sequences using text-processing tools and scripting languages
  - ❑ R programming language
  - ❑ Python
  - ❑ Ruby,
  - ❑ Perl

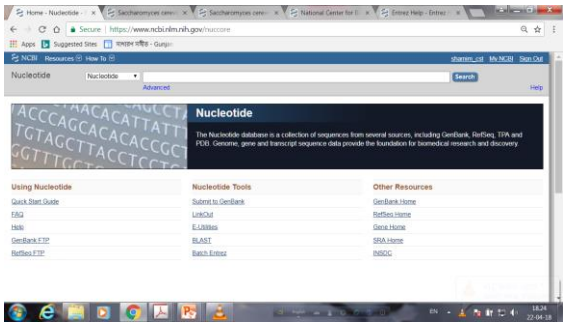
- A sequence in FASTA format is represented as a series of lines
- Each of which should be no longer than 120 characters and usually do not exceed 80 characters.
- The first line in a FASTA file starts either with a ">" (greater-than) symbol or, less frequently, a ";" (semicolon) and was taken as a comment.

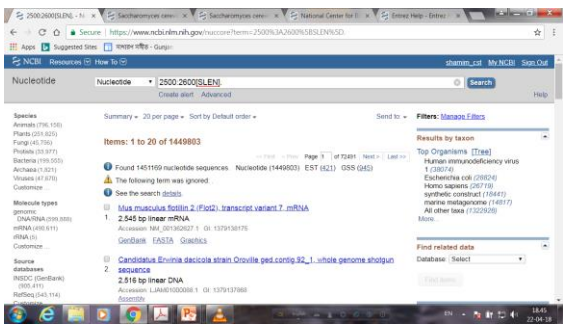
```
;LCBO - Prolactin precursor - Bovine
; a sample sequence in FASTA format
MDSKGSSQKGSRLLLLLVVSNLLLCQGVVSTPVCPNGPGNCQVSLRDLFDRAVMVSHYIHDLSS
EMFNEFDKRYAQGKGFITMALNSCHTSSLPTPEDKEQAQQTHHEVLMSLILGLLRSWNDPLYHL
VTEVRGMKGAPDAILSRAIEIEEENKRLLEGMEMIFGQVIPGAKETEPYPVWSGLPSLQTKDED
ARYSAFYNLLHCLRRDSSKIDTYLKLLNCRIIYNNNC*

>MCHU - Calmodulin - Human, rabbit, bovine, rat, and chicken
ADQLTEEQIAEFKEAFSLFDKDGDGTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTID
FPEFLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTNLGEKLTDEEVDEMIREA
DIDGDGQVNYEEFVQMMTAK*

>gi|5524211|gb|AAD44166.1|  cytochrome b [Elephas maximus maximus]
LCLYTHIGRNIYYGSYLYSETWNTGIMLLLITMATAFMGYVLPWGQMSFWGATVITNLFSAIPYIGTNLV
EWIWGGFSVDKATLNRFFAFHFILPFTMVALAGVHLTFLHETGSNNPLGLTSDSDKIPFHPYYTIKDFLG
LLILILLLLLLALLSPDMLGDPDNHMPADPLNTPLHIKPEWYFLFAYAILRSVPNKLGGVLALFLSIVIL
GLMPFLHTSKHRSMMLRPLSQALFWTLTMDLLTLTWIGSQPVEYPYTIIGQMASILYFSIILAFLPIAGX
IENY
```

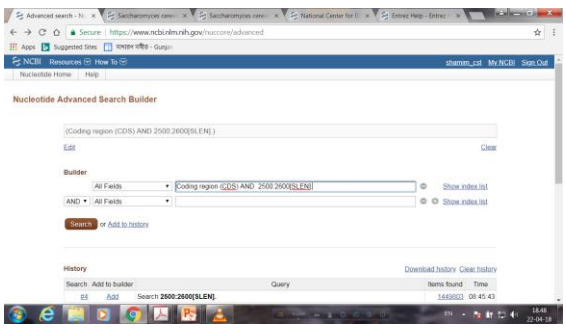
```
>SEQUENCE_1
MTEITAAMVKELRESTGAGMMDCKNALSETNGDFDKAVQLLREKGLGKAAKKADRLAAEG
LVSVKVSDDFTIAAMRPSYLSYEDLDMTFVENEYKALVAELEKENEERRRLKDPNKPEHK
IPQFASRKQLSDAILKEAEEKIKEELKAQGKPEKIWDNIIPGKMNSFIADNSQLDSKLTL
MGQFYVMDDKKTVEQVIAEKEKEFGGKIKIVEFICFEVGEGLEKKTEDFAAEVAAQL
>SEQUENCE_2
SATVSEINSETDFVAKNDQFIALTKDTTAHIQSNSLQSVEELHSSTINGVKFEEYLKSQI
ATIGENLVVRRFATLKAGANGVVNGYIHTNGRVGVVIAAACDSAEVASKSRDLLRQICMH
```

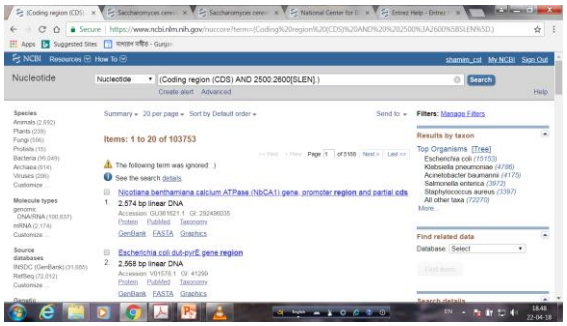
Nucleic Acid Code	Meaning	Mnemonic
A	A	Adenine
C	C	Cytosine
G	G	Guanine
T	T	Thymine
U	U	Uracil
R	A or G	puRine
Y	C, T or U	pYrimidines
K	G, T or U	bases which are Ketones
M	A or C	bases with aMino groups
S	C or G	Strong interaction
W	A, T or U	Weak interaction
B	not A (i.e. C, G, T or U)	B comes after A
D	not C (i.e. A, G, T or U)	D comes after C
H	not G (i.e., A, C, T or U)	H comes after G
V	neither T nor U (i.e. A, C or G)	V comes after U
N	A C G T U	Nucleic acid
-	gap of indeterminate length	

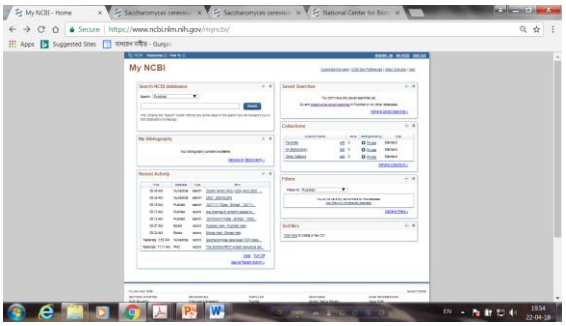
Amino Acid Code	Meaning
A	Alanine
X	any
R	Arginine
N	Asparagine
D	Aspartic acid
B	Aspartic acid (D) or Asparagine (N)
C	Cysteine
-	gap of indeterminate length
E	Glutamic acid
Z	Glutamic acid (E) or Glutamine (Q)
Q	Glutamine
G	Glycine
H	Histidine
I	Isoleucine
L	Leucine
J	Leucine (L) or Isoleucine (I)
K	Lysine
M	Methionine
F	Phenylalanine
P	Proline
O	Pyrrolysine
U	Selenocysteine
S	Serine
T	Threonine
*	translation stop
W	Tryptophan
Y	Tyrosine
V	Valine

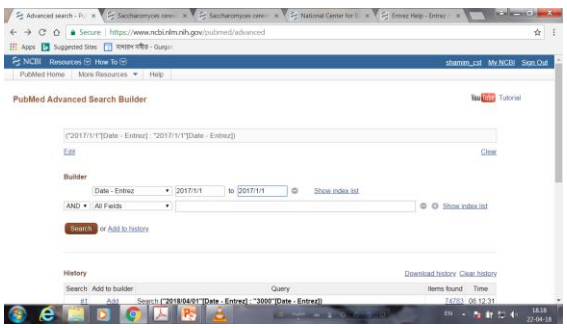


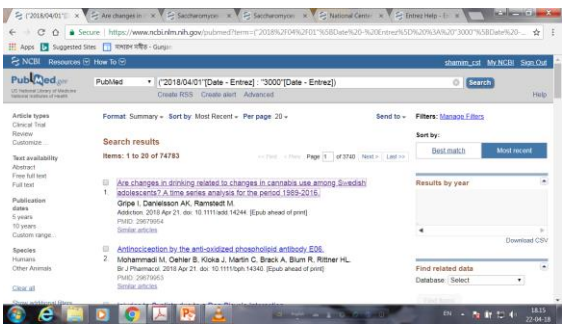


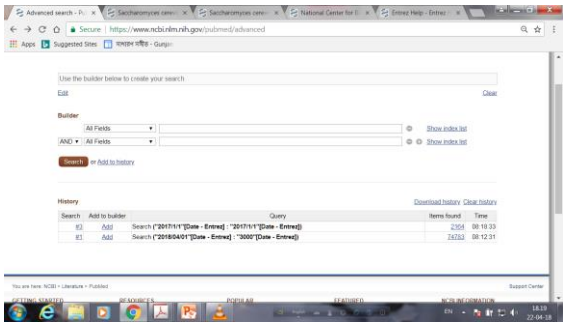


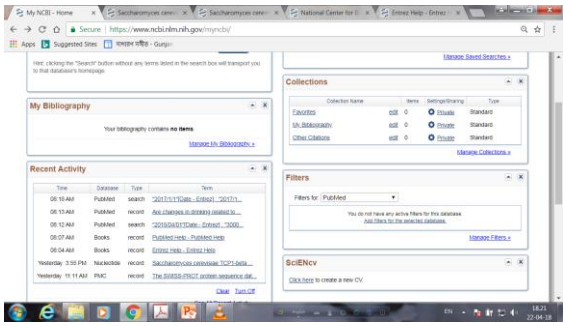

Nucleotide
(Coding region (CDS) AND 2500:2600[SLEN])
Search
Items: 1 to 20 of 103753
Nicotiana benthamiana calcium ATPase (NbCA1) gene, promoter region and partial cds
2,574 bp linear DNA
Escherichia coli dut-pyrE gene region
2,568 bp linear DNA
Results by taxon
Top Organisms [Tree]
Escherichia coli (15153)
Klebsiella pneumoniae (4786)
Acinetobacter baumannii (4175)
Salmonella enterica (3972)
Staphylococcus aureus (3397)
All other taxa (72270)
Filters: Manage Filters


My NCBI


PubMed Advanced Search Builder
("2017/1/1"[Date - Entrez] : "2017/1/1"[Date - Entrez])
Builder
Date - Entrez
2017/1/1
2017/1/1
All Fields
Search
Add to history
History


PubMed
("2018/04/01"[Date - Entrez] : "3000"[Date - Entrez])
Search results
Items: 1 to 20 of 74783
Are changes in drinking related to changes in cannabis use among Swedish adolescents? A time series analysis for the period 1989-2016.
Antinociception by the anti-oxidized phospholipid antibody, E06.
Results by year





Entrez Global Query page

- Literature
- Health
- Genomes
- Genes
- Proteins
- Chemicals.

Nucleotide

The Nucleotide database is a collection of sequences from several sources,
GenBank
RefSeq
TPA
PDB
Genome
Gene
transcript sequence data

RefSeq: NCBI Reference Sequence Database

- A comprehensive
- Integrated
- non-redundant
- well-annotated set of reference sequences including genomic, transcript, and protein.

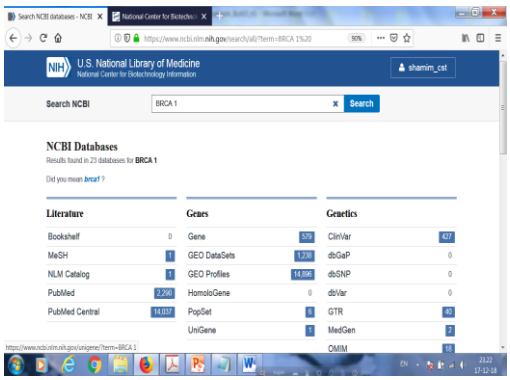
BioSystems

The NCBI BioSystems Database provides integrated access to biological systems and their component

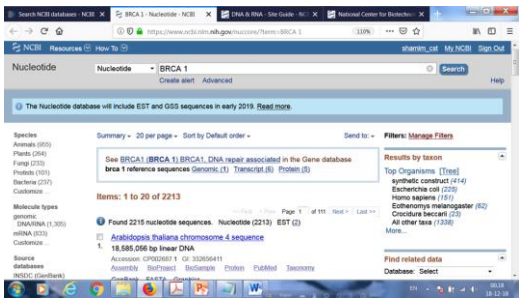
- Genes
- Proteins
- Small molecules
- Literature describing those biosystems
- Other related data throughout Entrez.

Gene

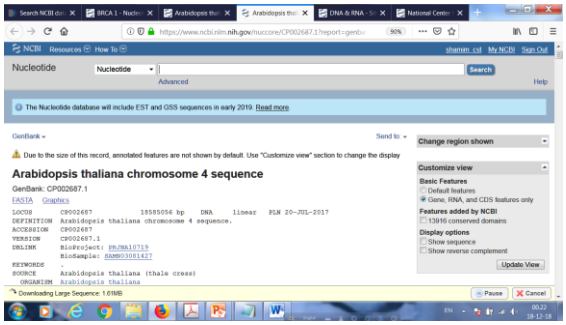
- Gene integrates information from a wide range of species. A record may include
- Nomenclature
- Reference Sequences (RefSeqs)
- Maps
- Pathways
- Variations
- Phenotypes
- Links to genome-, phenotype-



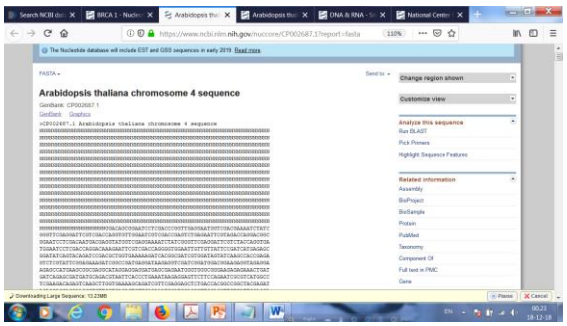
Nucleotide is selected



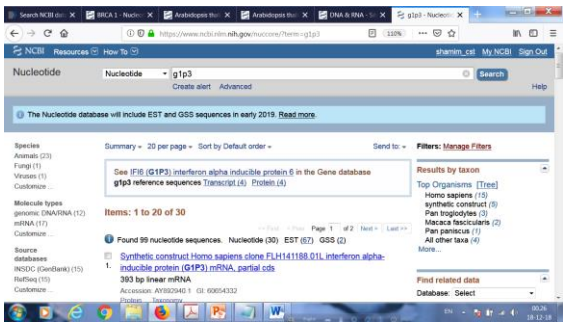
Genbank is selected

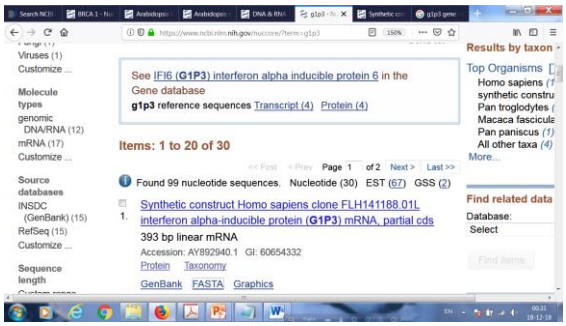


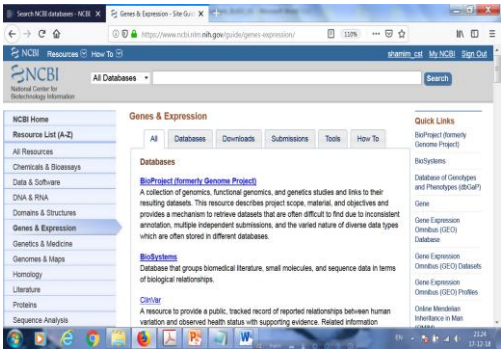
FASTA format is selected

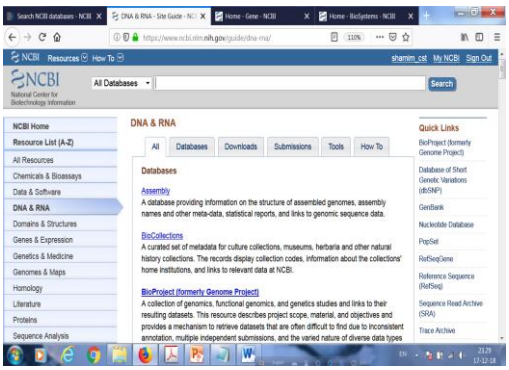


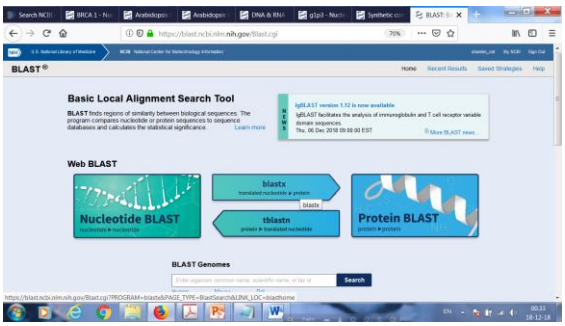
Search for gip3 (**G1P3, an interferon inducible gene**)



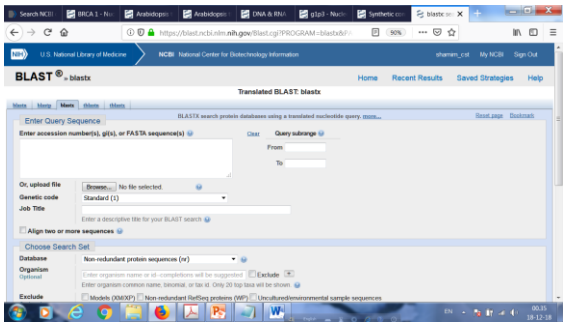

See IFI6 (G1P3) interferon alpha inducible protein 6 in the Gene database
g1p3 reference sequences Transcript (4) Protein (4)
Items: 1 to 20 of 30
Found 99 nucleotide sequences. Nucleotide (30) EST (67) GSS (2)
Synthetic construct Homo sapiens clone FLH141188.01L interferon alpha-inducible protein (G1P3) mRNA, partial cds
393 bp linear mRNA
Accession: AY892940.1 GI: 60654332
Protein Taxonomy
GenBank FASTA Graphics
Results by taxon
Find related data


Genes & Expression
Databases
BioProject (formerly Genome Project)
BioSystems
ClinVar
Quick Links


DNA & RNA
Databases
Assembly
BioCollections
BioProject (formerly Genome Project)
Quick Links

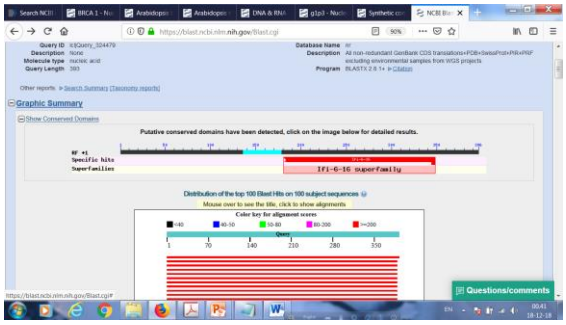

BLAST®
Basic Local Alignment Search Tool
Web BLAST
Nucleotide BLAST
blastx
tblastn
Protein BLAST
BLAST Genomes

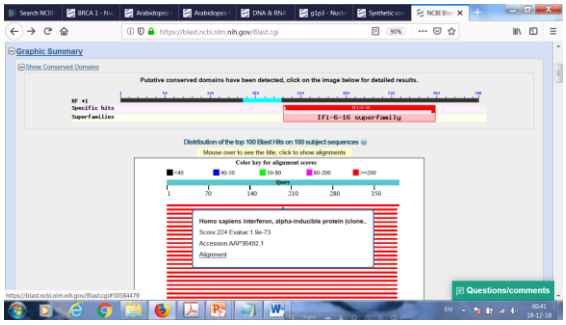
Select Blastx

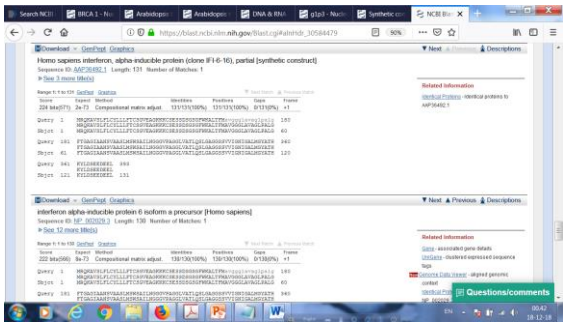


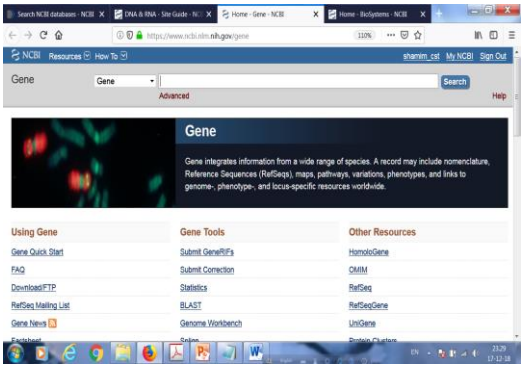
Gip3 sequence found from FASTA

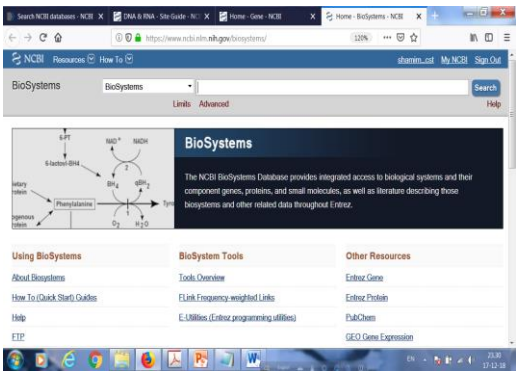
ATGCGGCAGAAGGCGGTATCGCTTTTCTTGTGCTACCTGCTGCTCTTCACTTGCAGTG
GGGTGGAGGCAG
GTAAGAAAAAGTGCTCGGAGAGCTCGGACAGCGGCTCCGGGTTCTGGAAGGCCCTGA
CCTTCATGGCCGT
CGGAGGAGGACTCGCAGTCGCCGGGCTGCCCGCGCTGGGCTTCACCGGCGCCGGCATCG
CGGCCAACTCG
GTGGCTGCCTCGCTGATGAGCTGGTCTGCGATCCTGAATGGGGGCGGCGTGCCCGCCG
GGGGGCTAGTGG
CCACGCTGCAGAGCCTCGGGGCTGGTGGCAGCAGCGTCGTCATAGGTAATATTGGTGC
CCTGATGGGCTA
CGCCACCCACAAGTATCTCGATAGTGAGGAGGATGAGGAGTTG

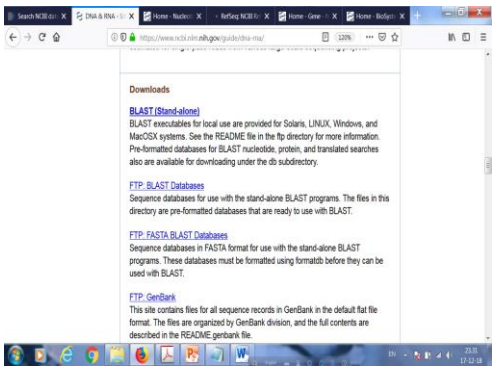


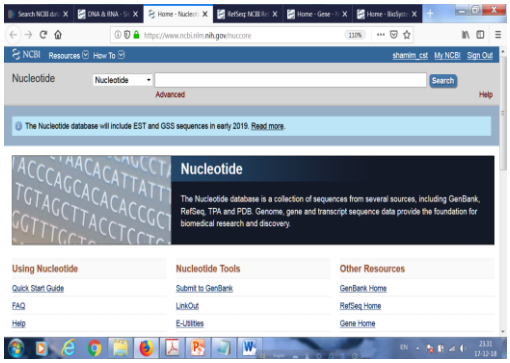



Homo sapiens interferon, alpha-inducible protein (clone IFI-6-16), partial [synthetic construct]
interferon alpha-inducible protein 6 isoform a precursor [Homo sapiens]
Questions/comments


Gene
Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.
Using Gene
Gene Tools
Other Resources


BioSystems
The NCBI BioSystems Database provides integrated access to biological systems and their component genes, proteins, and small molecules, as well as literature describing those biosystems and other related data throughout Entrez.
Using BioSystems
BioSystem Tools
Other Resources


Downloads
BLAST (Stand-alone)
BLAST executables for local use are provided for Solaris, LINUX, Windows, and MacOSX systems. See the README file in the ftp directory for more information. Pre-formatted databases for BLAST nucleotide, protein, and translated searches also are available for downloading under the db subdirectory.
FTP: BLAST Databases
Sequence databases for use with the stand-alone BLAST programs. The files in this directory are pre-formatted databases that are ready to use with BLAST.
FTP: FASTA BLAST Databases
Sequence databases in FASTA format for use with the stand-alone BLAST programs. These databases must be formatted using formatdb before they can be used with BLAST.
FTP: GenBank
This site contains files for all sequence records in GenBank in the default flat file format. The files are organized by GenBank division, and the full contents are described in the README.genbank file.


NCBI Resources How To
shamim_cst My NCBI Sign Out
Nucleotide
Search
Advanced
Help
The Nucleotide database will include EST and GSS sequences in early 2019. Read more.
Nucleotide
The Nucleotide database is a collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB. Genome, gene and transcript sequence data provide the foundation for biomedical research and discovery.
Using Nucleotide
Quick Start Guide
FAQ
Help
Nucleotide Tools
Submit to GenBank
LinkOut
E-Utilities
Other Resources
GenBank Home
RefSeq Home
Gene Home


NCBI Resources How To
shamim_cst My NCBI Sign Out
RefSeq
Search
RefSeq: NCBI Reference Sequence Database
A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.
Using RefSeq
About RefSeq
Human Reference Genome
Prokaryotic RefSeq Genomes
FAQ
NCBI Handbook
Factsheet
RefSeq Access
Human Genome Resources and Download
RefSeq FTP
RefSeq genomes FTP
New RefSeq genomic (last 30 days)
New RefSeq transcripts (last 30 days)
New RefSeq proteins (last 30 days)
RefSeq projects
Consensus CDS (CCDS)
RefSeq Functional Elements
RefSeqGene
Targeted Loci
Virus Variation