NCBI Molecular Biology Resources

USDA
AFRS

NCBI

April 24, 2002

NCBI Resources

- About NCBI
- NCBI Sequence Databases
  - Primary Database – GenBank
  - Derivative Databases - RefSeq
- Entrez Databases and Text Searching
- BLAST Services
- Genomic Resources

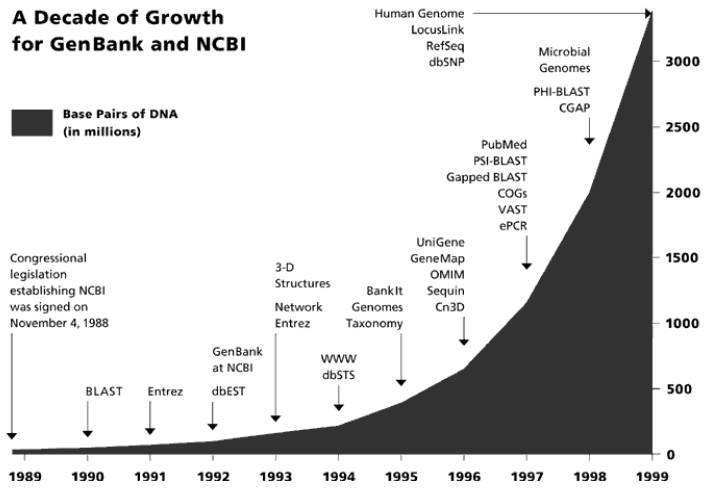
NCBI

The National Center for Biotechnology Information (NCBI)

- **Created as a part of NLM in 1988**
  - **Establish public databases**
  - **Research in computational biology**
  - **Develop software tools for sequence analysis**
  - **Disseminate biomedical information**
- **Tools: BLAST(1990), Entrez (1992)**
- **GenBank (1992)**
- **Free MEDLINE (PubMed, 1997)**
- **Human genome (2001)**

NCBI

NCBI History



NCBI

Molecular Databases

- Primary Databases
  - Original submissions by experimentalists
  - Database staff organize but don't add additional information
    - **Example: GenBank**
- Derivative Databases
  - Human curated
    - compilation and correction of data
    - **Example: SWISS-PROT, NCBI RefSeq mRNA**
  - Computationally Derived
    - **Example: UniGene**
  - Combinations
    - **Example: NCBI Genome Assembly**

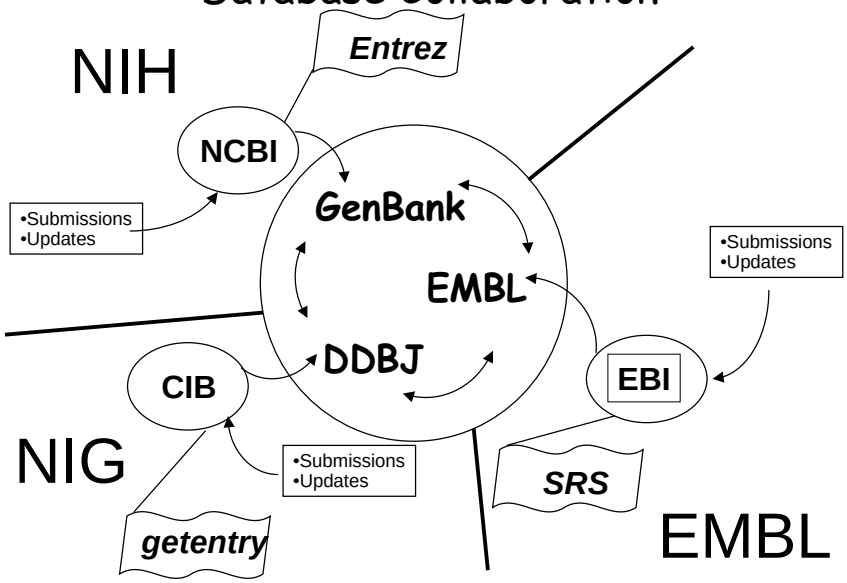
NCBI

What is GenBank? NCBI's Primary Sequence Database

- **Nucleotide only sequence database**
- **Archival in nature**
- **GenBank Data**
  - **Direct submissions individual records (BankIt, Sequin)**
  - **Batch submissions via email (EST, GSS, STS)**
  - **ftp accounts sequencing centers**
- **Data shared three collaborating databases**
  - **GenBank**
  - **DNA Database of Japan (DDBJ).**
  - **European Molecular Biology Laboratory Database (EMBL) at EBI.**

NCBI

The International Sequence Database Collaboration



NCBI

GenBank: NCBI's Primary Sequence Database

Release 128	February 2002
15,465,325	Records
17,089,143,893	Nucleotides
110,000 +	Species

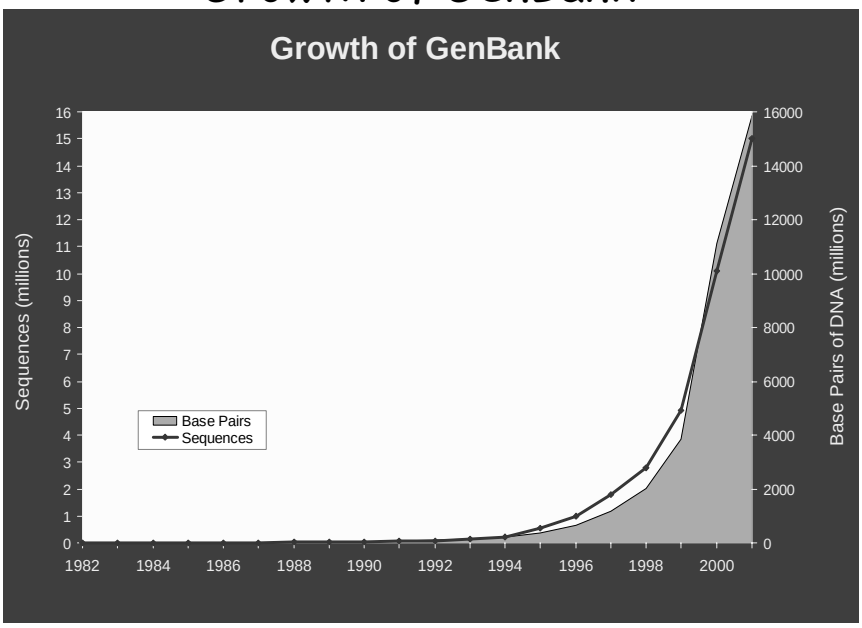
- full release every two months
- incremental and cumulative updates daily
- available only through internet

ftp://ftp.ncbi.nih.gov/genbank/

60 Gigabytes of data

NCBI

Growth of GenBank



NCBI

GenBank Divisions

Bulk Sequence Divisions

PAT	Patent
EST	Expressed Sequence Tags (142 files)
STS	Sequence Tagged Sites
GSS	Genome Survey Sequences (48 files)
HTG	High Throughput Genome (26 files)
HTC	High Throughput cDNA
CON	Contig

Traditional Divisions

BCT INV MAM PHG PLN PRI
ROD SYN UNA VRL VRT

A Traditional GenBank Record

```
LOCUS       AF153828                1586 bp    mRNA    linear   PLN 18-APR-2000
LOCUS   AF153828   1586 bp  mRNA  linear PLN   18-APR-2000
            AF153828.1    I:7532798
SOURCE
  ORGANISM                                  ica
            Eukaryota, V                    ptophyta; Embr
            Spermatophyt                    eudicotyledons, core eudicots;
            Rosidae; eur                    Rosaceae; Maloideae; Malus.
REFERENCE   1  (bases 1
  AUTHORS   Wegrzyn,T.,                     i,G., Murphy,P., Newcomb,R.,
            Gardner,R. and MacRae,E.
  TITLE     A novel alpha-amylase gene is transiently
  JOU
  MED
   PU
REFERENCE   2  (bases                       to 1586)
  AUTH                                      illy,K., Cipriani,G., Murphy,P.  Newcomb,R.,
                                            MacRae,E.
  TITLE     Direct Submission
  JOURNAL   Submitted (25-MAY-1999) Postharvest and Food, HortResearch, Private
            Bag 92169, Auckland, New Zealand
```

Locus Name

length

molecule type
mRNA==cDNA
DNA==gDNA

topology

gb division

modification date

Accession Number

```
ACCESSION   AF153828
VERSION     AF153828.1  GI:7532799
```

Version Number

GI Number

GenBank Record: Feature Table

```
FEATURES             Location/Qualifiers
     source          1..1586
                     /organism="Malus x domestica"
                     /cultivar="Granny Smith"
                     /db_xref="taxon:3750"
                     /tissue_type="fruit"
                     /note="isolated from young 8 week-old fruit and fruit
                     after 6 days in cold storage at 0.5 degrees Celsius."
                     /protein_id="AAC16332.2"
                     /db_xref="GI:7144485"
                                                 arch"
                     /note="alpha-amylase by similarity"
                     /codon_start=1
                     /product="alpha-amylase"
                     /protein_id="AAF63239.1"
                     /db_xref="GI:7532799"
                     /translation="MGYGSNDSRENAQQTDIGAAVRNGREILLQAFNWESHKHDWWRN

                       CPAGREWTLATCGHRYAVWNK"
BASE COUNT      474 a    311 c    370 g    431 t
ORIGIN
        1 tgcaatccgg ggccgagttg ggaaactaca tcctgagtca aatgggttac ggaagtaatg
     1561 tagtgcccta aaaaaaaaaa aaaaaa
//
```

GenPept Protein IDS

EST Division: Expressed Sequence Tags

```
>IMAGE:275615 5' mRNA sequence
GACAGCATTCGGGCCGAGATGTCTCGCTCCGTGGCCTTAGCTGTGCTCGCGCTACTCTCTCTTTCTGGCC
TGGAGGTATCCAGCGTACTCCAAAGATTCAGGTTTACTCACGTCATCCAGCAGAGAATGGAAAGTCAAAT
TTCCTGAATTGCTATGTGTCTGGGTTTCATCCATCCGACATTGAAGTTGACTTACTGAAGAATGGAGAGA
GAATTGAAAAAGTGGAGCATTCAGACTTGTCTTTCAGCAAGGACTGGTCTTTCTATCTCTTGTACTACAC
TGAATTCACCCCCACTGAAAAAGATGAGTATGCCTGCCGTGTTGAACCATGTNGACTTTGTCACAGNCCC
AAGTTNAGTTTAAGTGGGNATCGAGACATGTAAGGCAGGCATCATGGGAGGTTTTGAAGNATGCCGCNTT
TTGGATTGGGATGAATTCCAAATTTCTGGTTTGCTTGNTTTTTTAATATTGGATATGCTTTTG
```

```
>IMAGE:275615 3', mRNA sequence
NNTCAAGTTTTATGATTTATTTAACTTGTGGAACAAAAATAAACCAGATTAACCACAACCATGCCTTACT
TTATCAAATGTATAAGANGTAAATATGAATCTTATATGACAAAATGTTTCATTCATTATAACAAATTTCC
AATAATCCTGTCAATNATATTTCTAAATTTTCCCCCAAATTCTAAGCAGAGTATGTAAATTGGAAGTTAA
CTTATGCACGCTTAACTATCTTAACAAGCTTTGAGTGCAAGAGATTGANGAGTTCAAATCTGACCAAGAT
GTTGATGTTGGATAAGAGAATTCTCTGCTCCCCACCTCTANGTTGCCAGCCCTC
```

NCBI

make cDNA library

80-100,000 unique cDNA clones in library

What is UniGene?

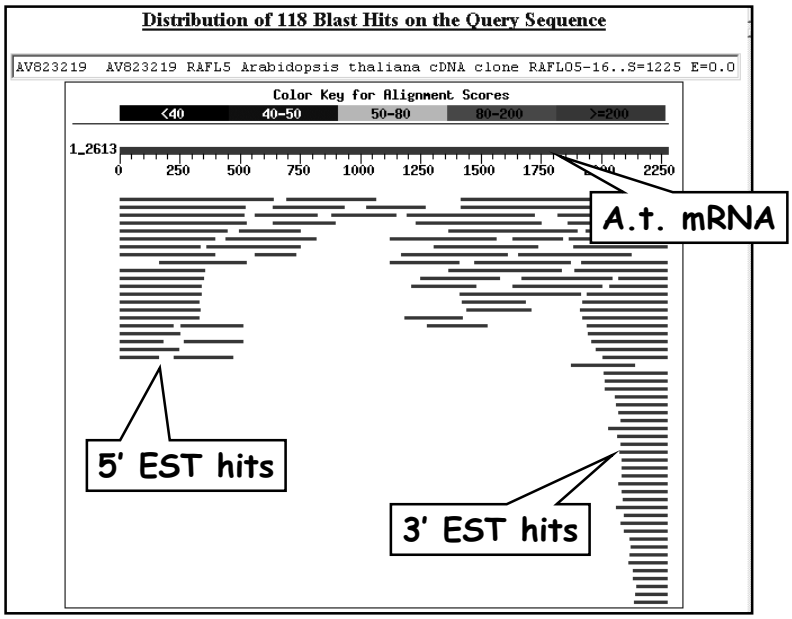
A gene-oriented view of sequence entries

- MegaBlast based automated sequence clustering
- Nonredundant set of gene oriented clusters
- Each cluster a unique gene
- Information on tissue types and map locations
- Includes well-characterized genes and novel ESTs
- Useful for gene discovery and selection of mapping reagents

NCBI

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

EST hits A.t. serine protease mRNA



NCBI

Arabidopsis UniGene Statistics

UniGene Build 14
Apr. 9th, 2002

```
    39,855     mRNAs + gene CDSs
    87,006     EST, 3'reads
    42,137     EST, 5'reads
+   32,571     EST, other/unknown
----------
   201,569    total sequences in clusters

Final Number of Clusters (sets)
==============================
26,808    sets total

25,474    s                    t one known gene
17,654    s                    t one EST
16,326    sets contain both genes and ESTs
```

115,000,000 bp
25,498 expected genes
5% uncharacterized transcripts

NCBI

Hs UniGene Statistics

UniGene Build 148
Apr. 8th, 2002

```
    73,419     mRNAs + gene CDSs
 1,181,855     EST, 3'reads
 1,461,928     EST, 5'reads
+  616,609     EST, other/unknown
----------
 3,333,811    total sequences in clusters

Final Number of Clusters (sets)
===============================
98,816   sets total

22,431   s                    one known gene
97,618   s                    one EST
21,233  sets contain both genes and ESTs
```

3,000,000 base pairs
30 K expected genes
80% uncharacterized transcripts

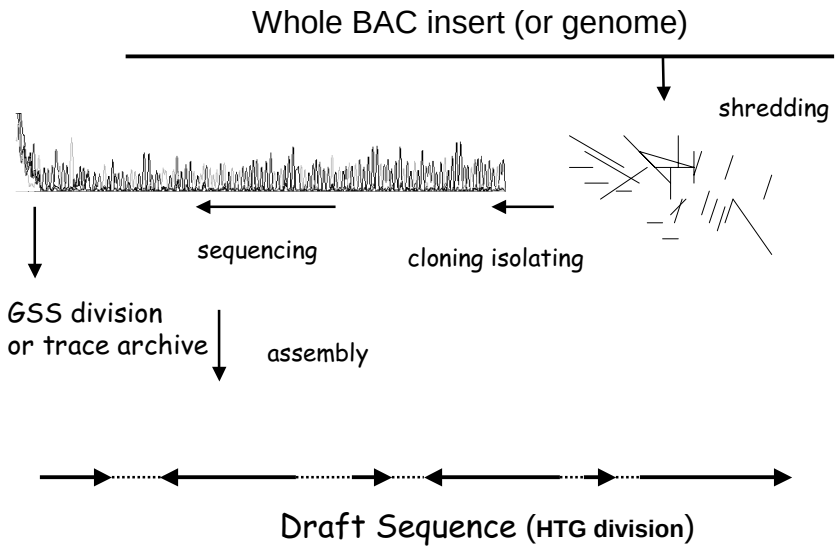
NCBI

UniGene Collections *Apr, 2002*

		Sequences	Clusters
Animals			
Homo sapiens	human	3,333,811	98,816
Mus musculus	mouse	2,274,640	86,897
Rattus norvegicus	rat	308,877	59,882
Danio rerio	zebrafish	159,261	14,893
Bos taurus	cow	122,503	9,303
Xenopus laevis	frog	120,489	16,489
Anopholes gambiae	mosquito	42,590	2,414
Plants			
Arabidopsis thaliana	thale cress	202,099	26,794
Oryzia sativa	rice	77,376	15,283
Triticum aestivum	wheat	35,387	3,091
Hordeum vulgare	barley	108,658	6,984
Zea mays	maize (corn)	108,030	9,889

NCBI

Genome Sequencing

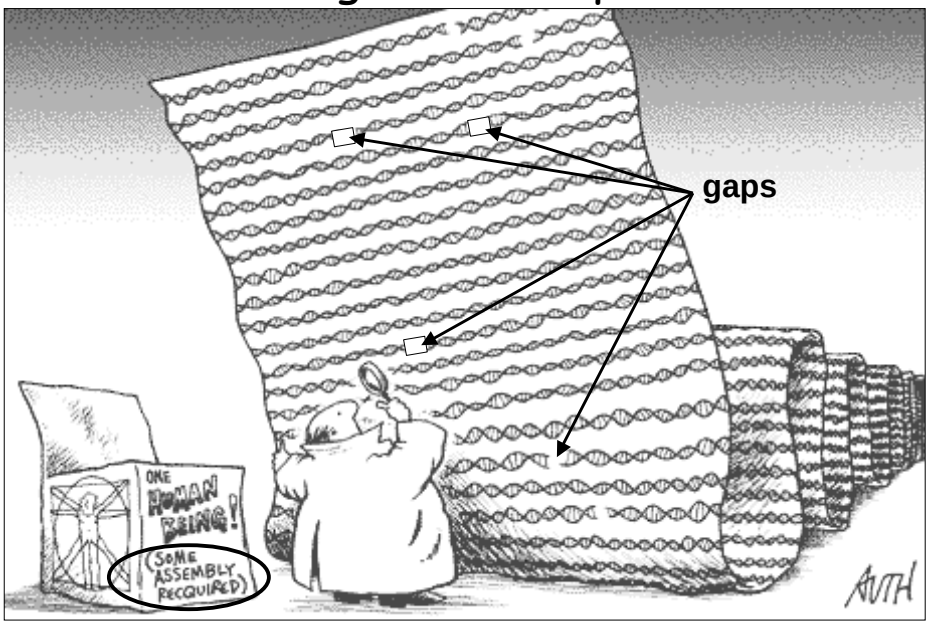


NCBI

```
LOCUS       BH245187                 195 bp    DNA     linear   GSS 13-NOV-2001
DEFINITION  AUIDA66TF AUID Arabidopsis thaliana genomic clone AUIDA66, DNA
            sequence.
ACCESSION   BH245187
VERSION     BH245187.1  GI:16922701
KEYWORDS    GSS.
SOURCE      thale cress.
  ORGANISM  Arabidopsis thaliana
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
            Rosidae; eurosids II; Brassicales; Brassicaceae; Arabidopsis.
REFERENCE   1  (bases 1 to 195)
  AUTHORS   Town,C.D., Whitelaw,C.A., Pai,G., Van Aken,S.E., Utterback,T.V.,
            Feldblyum,T.V. and Fraser,C.M.
  TITLE     Survey sequencing of Arabidopsis thaliana BAC F5I22
  JOURNAL   Unpublished (2001)
COMMENT     Contact: Chris Town
            TIGR
            9712 Medical Center Drive, Rockville, MD 20850, USA.
            Tel: 301-838-3523
            Fax: 301-838-0208
            Email: cdtown@tigr.org
            From Wash. U contig 720.
            Seq primer: TF
            Class: sheared ends.
FEATURES             Location/Qualifiers
     source          1..195
                     /organism="Arabidopsis thaliana"
                     /strain="Columbia"
                     /db_xref="taxon:3702"
                     /clone="AUIDA66"
                     /clone_lib="AUID"
                     /note="Vector: pHOS2; Site_1: BstXI; 2-3 kb sheared BAC
                     DNA inserted into pHOS2 using BstXI linkers"
```

NCBI

Working Draft Sequence



HTG Division: High Throughput Genome

phase 1 → ← → → ← HTG

```
ACCESSION   AC006228
VERSION     AC006228.1  GI:4056404
KEYWORDS    HTG; HTGS_PHASE1.
```

phase 2 → → → HTG

```
ACCESSION   AC006228
VERSION     AC006228.2  GI:4309686
KEYWORDS    HTG; HTGS_PHASE2.
```

phase 3 → PLN

```
ACCESSION   AC006228
VERSION     AC006228.4  GI:4580732
KEYWORDS    HTG.
```

40,000 to > 350,000 bp

NCBI

RefSeq: NCBI's Derivative Sequence Database

- **Curated transcripts and proteins**
  - reviewed
  - human, mouse, rat, fruit fly, zebrafish, arabidopsis
- **Human model transcripts and proteins**
- **Assembled Genomic Regions (contigs)**
  - draft human genome
  - mouse genome
- **Chromosome records**
  - microbial
  - organelle

NCBI

The RefSeq Accession Numbers

NCBI Reference Sequences

mRNAs and Proteins

NM_123456 Curated mRNA
NP_123456 Curated Protein
XM_123456 Predicted Transcript (human)
XP_123456 Predicted Protein (human)

human
mouse
rat
fruit fly
zebrafish
Arabidopsis

Gene Records

NG_123456 Reference Genomic Sequence (human)

Assemblies

NT_123456 Contig (Mouse and Human)
NC_123455 Chromosome (Microbial, Arabidopsis)

NCBI

GenBank Sequences: human CFTR

Show: 100 Items 1-83 of 83 One page.

1: M86631 Related Sequences, PubMed, Taxonomy
Homo sapiens (clone ST-18-5(9/16)) cystic fibrosis transmembrane conductance regulator (CFTR) gene, 3' end intron 17B; complete exon 18; complete intron 18
gi|180296|gb|M86631.1|HUMCFTR[180296]

2: S64699 Related Sequences, OMIM, Protein, PubMed, Taxonomy
Homo sapiens cystic fibrosis transmembrane conductance regulator isoform 36 (CFTR) mRNA, partial cds
gi|408285|gb|S64699.1|S64699[408285]

3: AL121762 Related Sequences, Taxonomy
Human DNA sequence from clone RP4-610C12 on chromosome 20 Contains the 3' end of a novel gene, a putative novel gene, a pseudogene similar to part of the cystic fibrosis transmembrane conductance regulator (CFTR), four CpG islands, ESTs, STSs and GSSs, complete sequence
gi|8574423|emb|AL121762.13|HS610C12[8574423]

4: AH006034 OMIM, Protein, PubMed, Taxonomy
Human cystic fibrosis transmembrane conductance regulator (CFTR) gene
gi|306537|gb|AH006034.1|SEG_HUMCFTRA[306537]

5: M55131 Related Sequences, ProbeSet, OMIM, Protein, PubMed, Taxonomy
Human cystic fibrosis transmembrane conductance regulator (CFTR) gene, exon 24
gi|306536|gb|M55131.1|HUMCFTRA26[306536]

6: M55130 Related Sequences, Protein, PubMed, Taxonomy
Human cystic fibrosis transmembrane conductance regulator (CFTR) gene, exon 23

NCBI

Curated RefSeq Records: **NM_, NP_**

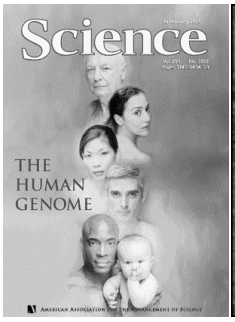
```
LOCUS       NM_000492    6159 bp    mRNA            PRI      26-JUL-1999
DEFINITION  Homo sapiens cystic fibrosis transmembrane conductance
            regulator(CFTR) mRNA
```

REFSEQ: This reference sequence was derived from M28668.1, M55131.1.
On Feb 17, 2000 this sequence version replaced gi:4502784.
Summary: Cystic fibrosis transmembrane conductance regulator is member 7 of the ATP-binding cassete sub-family C. The protein functions as a chloride channel and controls the regulation of other transport pathways. Mutations in this gene cause the autosomal recessive disorder, cystic fibrosis (CF) and congenital bilateral aplasia of the vas deferens (CBAVD). Alternative splice variants have been described, many of which result from mutations in the CFTR gene.
COMPLETENESS: full length.

Reviewed

```
COMMENT     REFSEQ: This reference sequence was derived from M55131.
            PROVISIONAL RefSeq: This is a provisional reference sequence
            record that has not yet been subject to human review. The final
            curated reference sequence record may be somewhat different from
            this one.
```

The Draft Human Genome



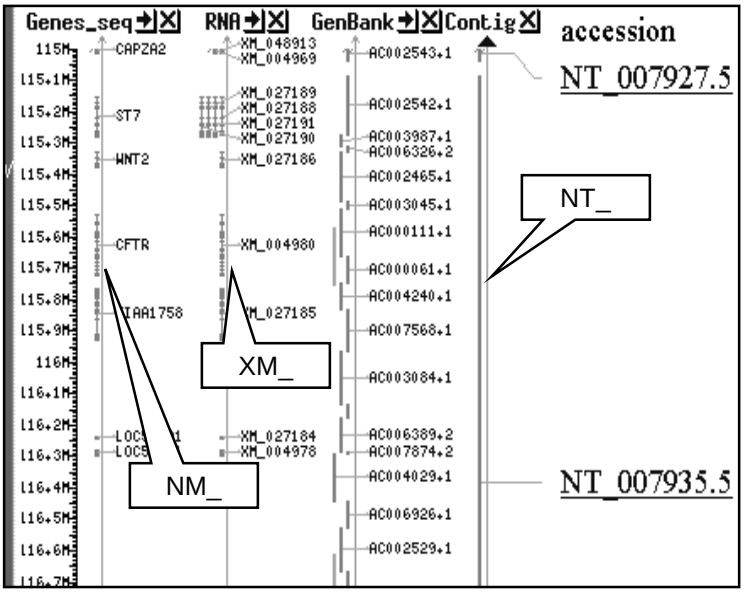
NCBI

RefSeq Human Contig: **NT_**

```
LOCUS       NT_007935 1888399 bp    DNA             CON       16-NOV-2000
DEFINITION  Homo sapiens chromosome 7 working draft sequence segment,
            complete sequence.
ACCESSION   NT_007935

     mRNA            complement(join(1255889..1257642,1258986..1259091,
CONTIG      join(AC073042.3:1155..2680,gap(100),AC074390.2:119526..151445,
            gap(100),AC074390.2:1..5245,gap(100),
            complement(AC074390.2:17705..23645),gap(100),
            AC074390.2:97658..119425,AC073042.3:106479..121155,
            AC074390.2:164226..165036,AC073042.3:70628..79503,gap(100),
            AC073042.3:4627..6382,gap(100),AC073042.3:2781..4526,gap(100),
            complement(AC073042.3:183627..[illegible]083),gap(100),
            AC073042.3:79604..88622,gap(100),[illegible]73042.3:139234..160437,
            gap(100),complement(AC073042.3:648[illegible]),gap(100),
            complement(AC073042.3:39354..45372),ga[illegible]
            complement(AC073042.3:21461..24064[illegible]
            AC074390.2:156347..160294,gap(100)[illegible]
            complement(AC074390.2:5346..10750),gap(100),
            complement(AC074390.2:153911..156246),gap(100),
            complement(AC074390.2:23746..32402),gap(100),
            complement(AC074390.2:151546..153810),gap(100),
            complement(AC074390.2:57277..75275),gap(100),
            complement(AC074390.2:75376..97557),gap(100),
                        /gene="CFTR"
```

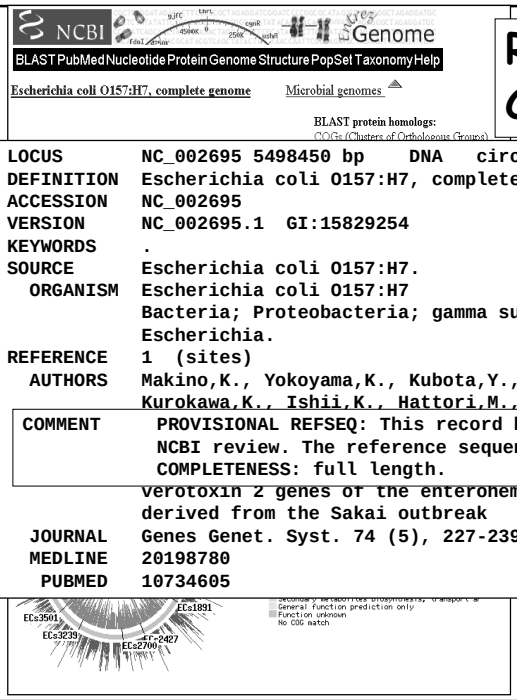
Reordering draft sequence

Map View of RefSeqs



NCBI

RefSeq Bacterial Chromosomes: NC_




```
LOCUS       NC_002695 5498450 bp    DNA   circular  BCT       02-OCT-2001
DEFINITION  Escherichia coli O157:H7, complete genome.
ACCESSION   NC_002695
VERSION     NC_002695.1  GI:15829254
KEYWORDS    .
SOURCE      Escherichia coli O157:H7.
  ORGANISM  Escherichia coli O157:H7
            Bacteria; Proteobacteria; gamma subdivision; Enterobacteriaceae;
            Escherichia.
REFERENCE   1  (sites)
  AUTHORS   Makino,K., Yokoyama,K., Kubota,Y., Yutsudo,C.H., Kimura,S.,
            Kurokawa,K., Ishii,K., Hattori,M., Tatsuno,I., Abe,H., Iida,T.,
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from BA000007.
            COMPLETENESS: full length.
            verotoxin 2 genes of the enterohemorrhagic Escherichia coli O157:H7
            derived from the Sakai outbreak
  JOURNAL   Genes Genet. Syst. 74 (5), 227-239 (1999)
  MEDLINE   20198780
   PUBMED   10734605
```

NCBI

RefSeq Plant Chromosomes: NC_

1: NC_003076
Arabidopsis thaliana chromosome 5, complete seq
gi|18426882|ref|NC_003076.2|[18426882]

2: NC_003074

```
LOCUS       NC_003076           26689408 bp    DNA     linear   PLN 10-JAN-2002
DEFINITION  Arabidopsis thaliana chromosome 5, complete sequence.
ACCESSION   NC_003076
VERSION     NC_003076.2  GI:18426882
KEYWORDS    HTG.
SOURCE      thale cress.
  ORGANISM  Arabidopsis thaliana
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
            Rosidae; eurosids II; Brassicales; Brassicaceae; Arabidopsis.
REFERENCE   1  (bases 1 to 26689408)
  AUTHORS   Town,C.D., Haas,B.J., Wu,D., Maiti,R., Hannick,L.I., Chan,A.P.,
            Tallon,L.J., Rooney,T., Utterback,T.R., VanAken,S.E.,
            Feldblyum,T.V., White,O. and Fraser,C.M.
  TITLE     Arabidopsis thaliana chromosome 5 genomic sequence
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 26689408)
  AUTHORS   Town,C.D. and Kaul,S.
  TITLE     Direct Submission
  JOURNAL   Submitted (10-JAN-2002) The Institute for Genomic Research, 9712
            Medical Center Dr, Rockville, MD 20850, USA, cdtown@tigr.org
COMMENT     PROVISIONAL REFSEQ: This record has not yet been subject to final
            NCBI review. The reference sequence was derived from AE502093.
            On Jan 30, 2002 this sequence version replaced gi:15237134.
            Address all correspondence to:at@tigr.org
```

Provisional record

NCBI

Other NCBI Derivative Databases

UniGene - gene oriented expressed sequence clusters

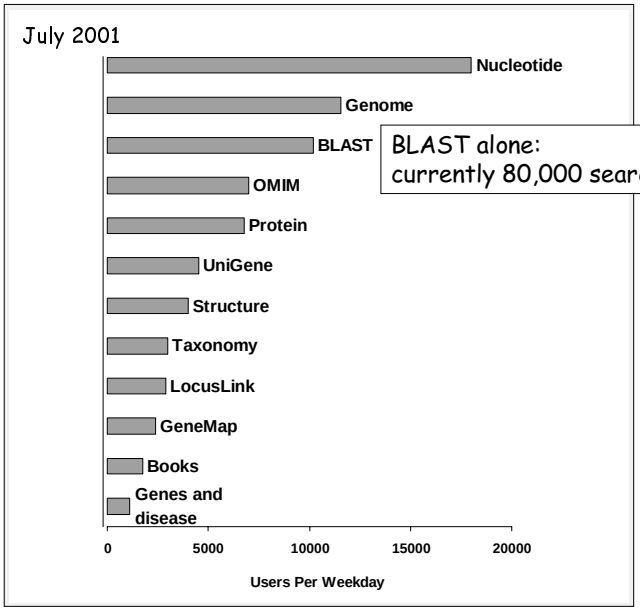
LocusLink - central resource and interface for known genes

NCBI

Integrated WWW Access: BLAST and Entrez

NCBI

Some Web Statistics



BLAST alone:
currently 80,000 searches per day

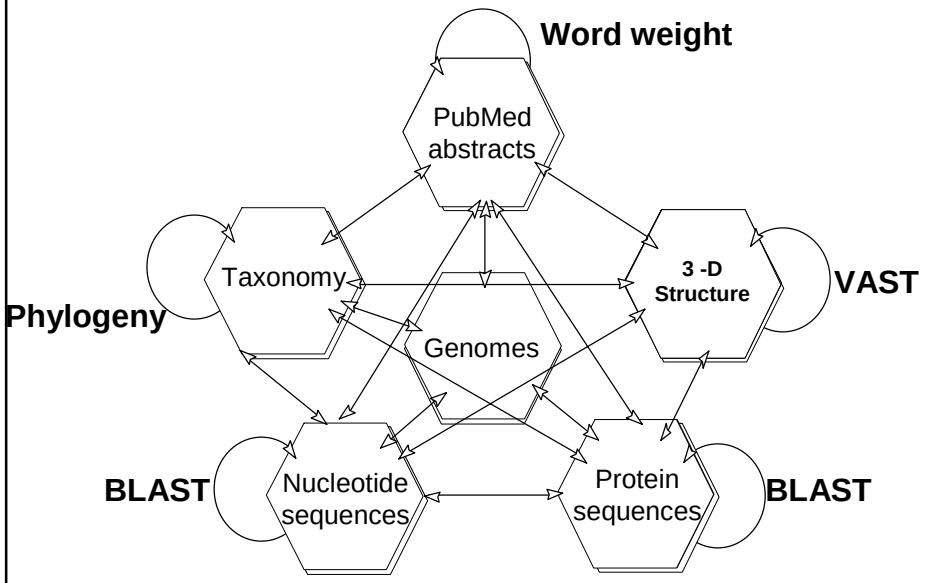
NCBI

Using Entrez

An integrated database
search and retrieval system

NCBI

Entrez: Database Integration



WWW Entrez

PubMed: The biomedical literature

- All of MEDLINE plus others
- Abstracts
- Links to online Journals

Nucleotide database

GenBank, EMBL, DDBJ
RefSeq, PDB

Protein sequence database

GenBank, DDBJ, EMBL translations
PDB, PIR, SWISS-PROT, PRF, RefSeq

Structure: three-dimensional molecular structure data structures

NCBI's MMDB - derived from PDB

Genome: complete genomes

Reference Genomes:
Graphical views, assembled sequence and mapping data

PopSet: population study data sets

OMIM: Online Mendelian Inheritance in Man

Taxonomy: organisms in GenBank

NEW Books: online books

NEW ProbeSet: Gene Expression Omnibus (GEO)

NEW 3D Domains: domains from Entrez Structure

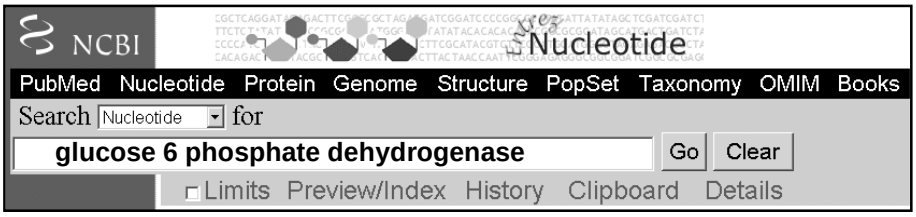
NCBI

Database Searching with Entrez

- Using limits and field restriction to find plant g6pdh
- Linking and neighboring with g6pdh

NCBI

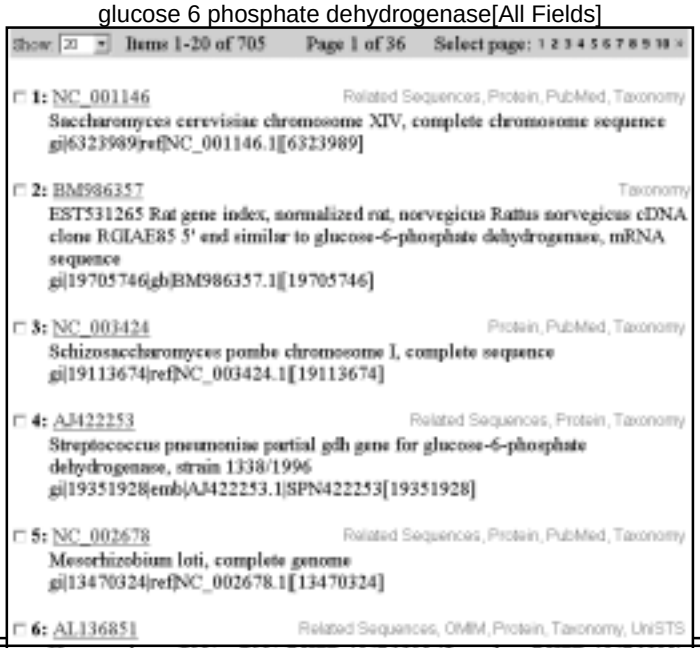
Entrez Nucleotides



NCBI

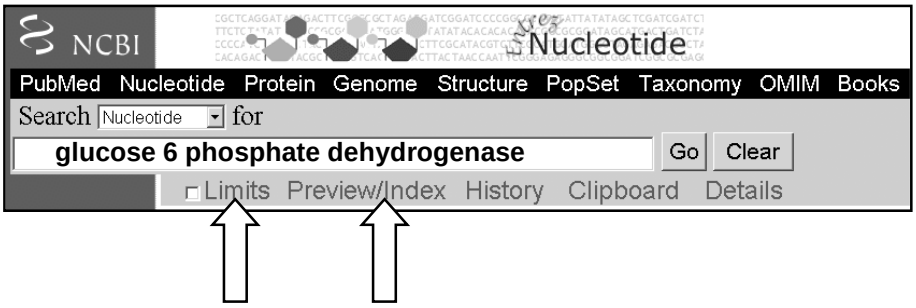
Document Summaries:

glucose 6 phosphate dehydrogenase[All Fields]



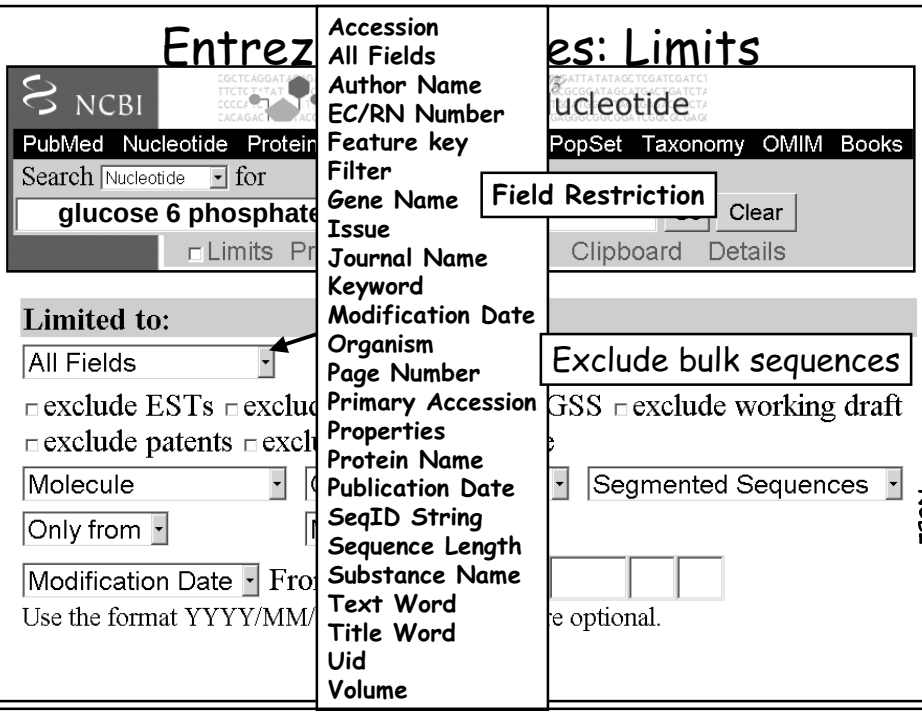
NCBI

Entrez Nucleotides: Limits & Preview/Index



NCBI

Entrez Nucleotides: Limits



NCBI

Entrez Nucleotides: Limits

NCBI Entrez Nucleotide

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM Books

Search Nucleotide for

glucose 6 phosphate dehydrogenase | Go | Clear

Clipboard Details

Limited to:

Title

exclude ESTs exclude STSs exclude GSS exclude working draft

exclude patents exclude all of the above

mRNA | Genomic DNA/RNA | Segmented Sequences

Only from | Modification Date

Modification From To

Use the and day are optional.

Title == Definition

Exclude Bulk Sequences

Nuclear gene

mRNA molecule type

NCBI

Document Summaries: Limits

Field: Title, Limits: exclude all of the above, Genomic DNA/RNA, mRNA

Display Summary Save Text Clip Add

Show 20 Items 1-20 of 61 Page 1 of 4 Select page: 1 2 3 4

1: AY078072 Related Sequences, Proteins, Taxonomy
Oryza sativa glucose-6-phosphate dehydrogenase (g6pdh) mRNA, complete cds
gi|19071786|gb|AY078072.1|[19071786]

2: XM_049337 Related Sequences, Proteins, Taxonomy
Homo sapiens glucose-6-phosphate dehydrogenase (G6PD), mRNA
gi|14768486|ref|XM_049337.1|[14768486]

3: NM_019468 Related Sequences, Proteins, PubMed, Taxonomy, UniSTS
Mus musculus glucose-6-phosphate dehydrogenase 2 (G6pd2), mRNA
gi|13937388|ref|NM_019468.1|[13937388]

4: NM_008062 Related Sequences, Proteins, PubMed, Taxonomy, UniSTS
Mus musculus glucose-6-phosphate dehydrogenase X-linked (G6pdx), mRNA
gi|6996916|ref|NM_008062.1|[6996916]

5: U09036 Proteins, Taxonomy
Hyalophora cecropia glucose-6-phosphate dehydrogenase mRNA, partial cds
gi|606629|gb|U09036.1|HCU09036|[606629]

6: AY056232 Related Sequences, Proteins, Taxonomy
Arabidopsis thaliana putative Glucose-6-phosphate dehydrogenase (F14J9.8) mRNA, complete cds
gi|15810386|gb|AY056232.1|[15810386]

7: BC000337 Related Sequences, OMIM, Proteins, Taxonomy

NCBI

Adding Terms: Preview/Index

Accession
All Fields

Add Term(s) to Query or View Index:

- Enter a term in the text box; use the pull-down menu to specify a search field.
- Click Preview to add terms to the query box and see the number of search results, or click Index to view terms within a field.
- Multiple terms selected from Index will be ORed; click AND to add to search.

Organism | green plants | Preview | Index

Click AND | OR | NOT to add terms selected from Index to the query box.

green nts(2774214)
green son berry(1)
green rcelain crab(11)
green pygmy goose(1)
green rice leafhopper(1)
green ringtail possum(1)
green rock lobster(2)
green sea mat(3)
green seaturtle(105)
green striped grasshopper(5)

Up
Down

NCBI

Plant cytosolic g6pdh mRNAs

Display | Summary | Save | Text | Clip Add

Items 1-21 of 21 | One page.

1: AY078072
Oryza sativa glucose-6-phosphate

2: AY056232
Arabidopsis thaliana putative Gluco

3: AB029456
Triticum aestivum g6pdh mRNA f
clone:Tagpd3

4: AB029455
Triticum aestivum g6pdh mRNA f
clone:Tagpd2

5: AB029454
Triticum aestivum g6pdh mRNA f
clone:Tagpd1

6: AJ279688
Beta vulgaris partial mRNA for G

7: AB001770
Nicotiana tabacum mRNA for cyto

8: AB001769 — Related Sequences, Protein, PubMed, Taxonomy
Nicotiana tabacum mRNA cytosolic glucose-6-phosphate dehydrogenase TC06

15: AJ000183 — Related Sequences, Protein, Taxonomy
Spinacia oleracea mRNA for glucose-6-phosphate dehydrogenase, clone O30A5
gi|2276345|emb|AJ000183.1|SO000183[2276345]

16: AJ000182 — Related Sequences, Protein, Taxonomy
Spinacia oleracea mRNA for glucose-6-phosphate dehydrogenase, clones O28FA14 & O30A4
gi|2276343|emb|AJ000182.1|SO000182[2276343]

17: X99405 — Related Sequences, Protein, Taxonomy
N.tabacum mRNA for chloroplast glucose-6-phosphate dehydrogenase
gi|1480343|emb|X99405.1|NTG6PD[1480343]

18: U18238 — Protein, PubMed, Taxonomy
Medicago sativa glucose-6-phosphate dehydrogenase mRNA, complete cds
gi|603218|gb|U18238.1|MSU18238[603218]

19: X83923 — Protein, PubMed, Taxonomy
S.tuberosum mRNA for glucose-6-phosphate dehydrogenase
gi|1197384|emb|X83923.1|STG6PDHPI[1197384]

20: X74421 — Protein, PubMed, Taxonomy
S.tuberosum mRNA for glucose-6-phosphate dehydrogenase
gi|471344|emb|X74421.1|STG6PDH[471344]

21: X84229 — Related Sequences, Protein, Taxonomy
A.thaliana mRNA for glucose-6-phosphate dehydrogenase (clone E5)
gi|1166404|emb|X84229.1|ATG6PDHE5[1166404]

NCBI

Plant cytosolic g6pdh mRNAs

Summary
Brief
GenBank
ASN.1
FASTA
GI list
LinkOut
PubMed Links
Protein Links
Nucleotide Neighbors
PopSet Links
Structure Links
Genome Links
Taxonomy Links
OMIM Links

Formats

Links and neighbors
(related records)

NCBI

Entrez GenBank / GenPept

1: U18238. Medicago sativa g...[gi:603218] Protein, PubMed, Taxonomy, LinkOut

```
LOCUS       MSU18238
DEFINITION  Medicago sativa gluco
            cds.
ACCESSION   U18238
VERSION     U18238.1  GI:603218
KEYWORDS    .
SOURCE      alfalfa.
  ORGANISM  Medicago sativa subsp
            Eukaryota; Viridiplan
            Spermatophyta; Magno
            Rosidae; eurosids I;
            Medicago.
REFERENCE   1  (bases 1 to 1785)
  AUTHORS   Fahrendorf,T., Ni,W.,
  TITLE     Stress responses in a
            Transcriptional activ
            genes at the onset o
  JOURNAL   Plant Mol. Biol. 28
  MEDLINE   95367649
   PUBMED   7640360
REFERENCE   2  (bases 1 to 1785)
  AUTHORS   Fahrendorf,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-DEC-199
            Foundation, Plant Bio
            Ardmore, OK 73402, US
FEATURES             Location/Qua
     source          1..1785
                     /organism="
                     /cultivar="
                     /sub_species
                     /db_xref="ta
                     /clone="G10"
                     /tissue_type
     CDS             38..1585
                     /codon_start
                     /product="gl
                     /protein_id=
                     /db_xref="GI
                     /translation
                     KTFPALFHLYKQELLPPDEVHIFGYARSKISDDELKNKLKSYLVPEKGASPKQLDDVS
```

1: AAB41552. glucose-6-phospha...[gi:603219] BLink, Nucleotide, Related Sequences, PubMed, Taxonomy, LinkOut

```
LOCUS       AAB41552                 515 aa            linear   PLN 30-JAN-1997
DEFINITION  glucose-6-phosphate dehydrogenase.
ACCESSION   AAB41552
PID         g603219
VERSION     AAB41552.1  GI:603219
DBSOURCE    locus MSU18238 accession U18238.1
KEYWORDS    .
SOURCE      Medicago sativa subsp. sativa.
  ORGANISM  Medicago sativa subsp. sativa
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
            Rosidae; eurosids I; Fabales; Fabaceae; Papilionoideae; Trifolieae;
            Medicago.
REFERENCE   1  (residues 1 to 515)
  AUTHORS   Fahrendorf,T., Ni,W., Shorrosh,B.S. and Dixon,R.A.
  TITLE     Stress responses in alfalfa (Medicago sativa L.) XIX.
            Transcriptional activation of oxidative pentose phosphate pathway
```

```
                     /db_xref="taxon:56147"
                     /clone="G10"
                     /tissue_type="cell suspension culture"
     Protein         1..515
                     /product="glucose-6-phosphate dehydrogenase"
     CDS             1..515
                     /coded_by="U18238.1:38..1585"
ORIGIN
        1 mgtnewhver rdsigtespv arevletgtl sivvlgasgd lakkktfpal fhlykqellp
       61 pdevhifgya rskisddelr nklrsylvpe kgaspkqldd vskflqlvky vsgpydsedg
      121 frlldkeise heylknskeg ssrrlfylal ppsvypsvck miktccmnks dlggwtrvvv
      181 ekpfgrdles aeelstqige lfeepqiyri dhylgkelvq nmlvlrfanr fflplwnhnh
      241 idnvqivfre dfgtdgrggy fdqygiirdi ipnhllqvlc liamekpvsl kpehirdekv
      301 kvlesvlpir ddevvlgqye gytddptvpd dsntptfatt ilrihnerwe gvpfivkagk
      361 alnsrkaeir vqfkdvpgdi frskkqgrne fvirlqpsea iymkltvkqp glemsavqse
      421 ldlsygqryq gitipeayer lildtirgdq qhfvrrdelk aswqiftpll hkidrgelkp
      481 vpynpgsrgp aeadelleka gyvqtpgyiw ipptl
//
```

NCBI

FASTA Format

```
>gi|603218|gb|U18238.1|MSU18238 Medicago sativa glucose-6-phosphate dehyd
CCACCAGATATAATTAAGTAGATCAGAGTAGAAGAAGATGGGAACAAATGAATGGCATGTAGAAAGAAGA
```

FASTA Definition Line

>gi|603218|gb|U18238.1|MSU18238

gi number

Locus Name

Accession number

Database Identifiers	
gb	GenBank
emb	EMBL
dbj	DDBJ
sp	SWISS-PROT
pdb	Protein Databank
pir	PIR
prf	PRF
ref	RefSeq

NCBI

Abstract Syntax Notation: ASN.1

```
Seq-entry ::= set {
  level 1 ,
  class nuc-
  descr {
    title "Medicago sativa glucose-6-phosphate dehydrogenase mRNA, and
 translated products" ,
    source {
      org {
        taxname "Medicago sativa subsp. sativa" ,
        db {
          {
            db "taxon" ,
            tag
              id 56147 } } ,
        orgname {
          nam
            b
                icago" ,
                tiva" ,
              subspecies "subsp. sativa" } ,
          mod {
```

GenPept

GenBank

ASN.1

FASTA Protein

FASTA Nucleotide

NCBI

NCBI Toolbox

```
/*****************************************************************************
*
*   asn2ff.c
*         convert an ASN.1 entry to flat file format, using the FFPrintArray.
*
****************
#include <accent
#include "asn2ff
#include "asn2ff
#include "ffprin
#include <subuti
#include <objall
#include <objcod
#include <lsqfet
#include <explor

#ifdef ENABLE_ID
#include <accid1
#endif

FILE *fpl;

Args myar
```

Toolbox Sources

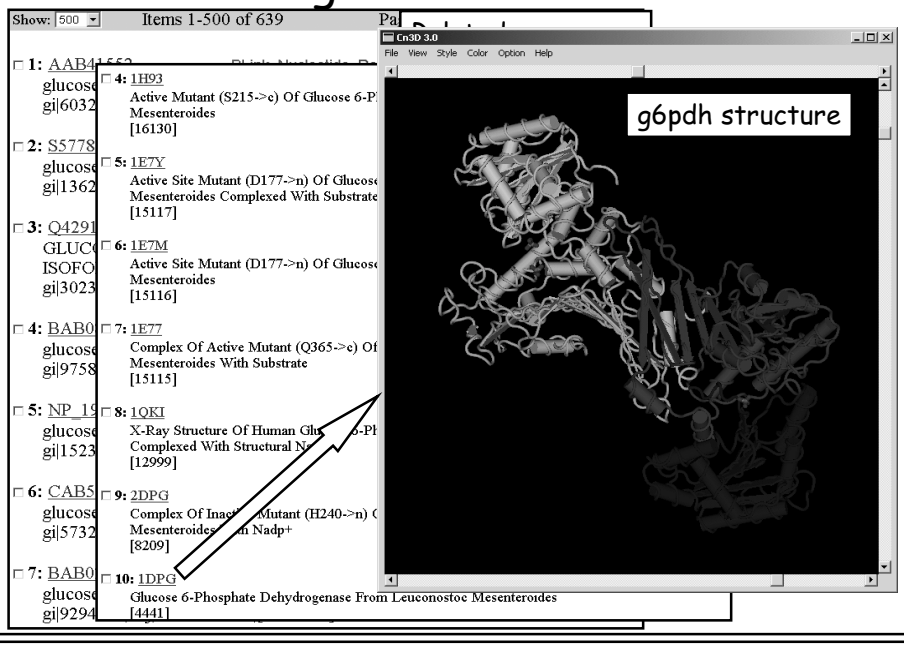
```
ftp> open ftp.ncbi.nih.gov
.
.
ftp> cd toolbox
ftp> cd ncbi_tools
```

ftp://ftp.ncbi.nlm.gov/toolbox/ncbi_tools

```
{"Input asnfile in binary mode","F",NULL,NULL,TRUE,'b',ARG_BOOLEAN,0.0,0,NULL},
{"Output Filename","stdout", NULL,NULL,TRUE,'o',ARG_FILE_OUT,0.0,0,NULL},
{"Show Sequence?","T", NULL ,NULL ,TRUE,'h',ARG_BOOLEAN,0.0,0,NULL},
```

NCBI

Protein Neighbors-Structure Links



NCBI

Advanced Neighbors: BLink

☐ **1:** AAB41552 BLink, Nucleotide, Related Sequences, PubMed, Taxonomy
glucose-6-phosphate dehydro[illegible]ase [Medicago sativa subsp. sativa]
gi|603219|gb|AAB41552.1|[60[illegible]19]

NCBI

BLink

BLAST PubMed Nucleotide Protein Genome Structure Taxonomy Help

Query: gi|603219 glucose-6-phosphate dehydrogenase
Matching gi: 1362053, 3023815

515 aa

SCORE	P	ACCESSION	GI	N	ORGANISM
2250	14	BAB08837	9758370	14	Arabidopsis thaliana
350	3	AAL18424	16416802	- 1	*Aedeomyia squamipennis*
347	3	AAL18427	16416808	- 1	*Anopheles stephensi*
340	3	AAL18435	16416824	- 1	*Anopheles coustani*
339	3	AAL18426	16416806	- 1	*Toxorhynchites amboinensis*
337	3	AAL18443	16416840	- 1	*Anopheles albimanus*
336	3	AAL18425	16416804	- 1	*Uranotaenia sapphirina*
335	3	AAL18442	16416838	- 1	*Anopheles albitarsis*
332	3	AAB02777	606624	- 1	*Ctenolepisma lineata*
331	3	AAL18439	16416832	- 1	*Bironella gracilis*
330	3	AAL18432	16416818	- 1	*Anopheles mattogrossensis*
327	3	AAL18441	16416836	- 1	*Armigeres subalbatus*
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]*opheles quadrimaculatus*
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]*opheles bellator*
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]*opheles cruzii*
[illegible]	[illegible]	AAB02786	606642	- 1	*Salina tristani*
[illegible]	3	AAB02787	606644	- 1	*Zerene cesonia*
317	3	AAL18431	16416816	- 1	*Anopheles neivai*
316	3	AAL18430	16416814	- 1	*Orthopodomyia alba*
313	3	AAB02781	606632	- 1	*Sarcophaga bullata*
304	3	AAB02778	606626	- 1	*Culex pipiens*
295	3	BAA78540	4996576	- 1	*Bactrocera dorsalis*
293	3	BAA78545	4996586	- 1	*Bactrocera scutellata*
287	3	BAA78548	4996592	- 1	*Bactrocera kandiensis*
286	3	CAB61333	6453559	- 1	*Laminaria digitata*
286	3	BAA78543	4996582	- 1	*Bactrocera cucurbitae*

Hits to NAD binding domain

NCBI

PubMed Link

1: Plant Mol Biol 1995 Aug;28(5):885-900 Related Articles, Nucleotide, Protein, NEW Books, LinkOut

Stress responses in alfalfa (Medicago sativa L.) XIX. Transcriptional activation of oxidative pentose phosphate pathway genes at the onset of the isoflavonoid phytoalexin response.

Fahrendorf T, Ni W, Shorrosh BS, Dixon RA.

Plant Biology Division, Samuel Roberts Noble Foundation, Ardmore, OK 73401, USA.

We have isolated cDNA clones encoding the pentose phosphate pathway enzymes 6-phosphogluconate dehydrogenase (6PGD, EC 1.1.1.44) and glucose 6-phosphate dehydrogenase (G6PDH, EC 1.1.1.49) from alfalfa (Medicago sativa L.). These exhibit extensive nucleotide and amino acid sequence similarity to the corresponding genes from bacteria, Drosophila and mammals. Transcripts encoding both enzymes are expressed at high levels in roots and nodules. Exposure of alfalfa suspension cells to an elicitor from yeast cell walls results in co-ordinated increases in transcription rates for both genes, followed by increased steady state transcript levels but only slightly increased extractable enzyme activities, at the onset of accumulation of isoflavonoid phytoalexins. Levels of NADPH and NADP remain relatively constant in alfalfa cells following elicitation. The rapid transcriptional activation of 6PGDH and G6PDH does not therefore appear to be a response to altered pyridine nucleotide redox state. These genes appear to respond to early events in elicitor-mediated signalling rather than to subsequent elicitor-induced changes in secondary metabolism. Hydrogen peroxide, a potential signal for elicitation of anti-oxidative genes in biologically stressed plant cells did not induce 6PGDH or G6PDH transcripts or enzymatic activity.

NCBI

Online Books

NCBI Books

PubMed Nucleotide Protein

Search Books for cDNA clones

Limits Preview/Index

About Entrez

Books

Overview

Using the books

Information for authors and publishers

Contact us

Project background

FAQ

Cubby

Privacy Policy

Display Books Save

20 items in Molecul
Lodish, Harvey; Berk, A
James E.
New York: W H Freema

11 items in C. elega
Riddle, Donald L.; Blum
Plainview (NY): Cold Sp

9 items in Molecula
Alberts, Bruce; Bray, D
New York and London:

6 items in Introduc
Griffiths, Anthony J.F.;
New York: W H Freema

6 items in Cancer
Bast, Robert C.; Kufe,
Frei, Emil, editors.
Canada: BC Decker Inc;

3 items in Retroviru
Coffin, John M.; Hughes
Plainview (NY): Cold Sp

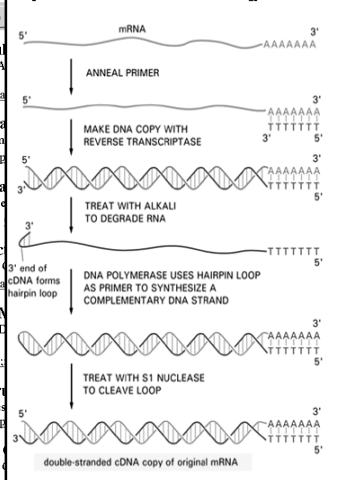
3 items in Modern
Griffiths, Anthony J.F.;
New York: W H Freema

MOLECULAR BIOLOGY OF THE CELL @NCBI

PUBMED NCBI HOMEPAGE MOLECULAR BIOLOGY OF THE CELL HOMEPAGE

Short Contents | Full Contents Other books @ NCBI

Molecular Biology of the Cell, 3rd edn. ➔ Part II Molecular Genetics ➔ **Chapter 7. Recombinant DNA Technology** ➔ DNA Cloning 15



Navigation

7. Recombinant DNA Technology

Introduction

The Fragmentation, Separation, and Sequencing of DNA Molecules

Nucleic Acid Hybridization

DNA Cloning

DNA Engineering

References

Search

Go

This book All books

NCBI

The Tax Browser

NCBI

Taxonomy Browser

PubMed | Entrez | BLAST | OMIM | Taxonomy | Structure

Search for | As complete name | lock | Go | Clear

Display levels 3 | nucleotides | proteins | structures | genome records | Accept

Lineage (abbreviated): root; Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots; Rosidae; eurosids I; Rosales; Rosaceae; Rosoideae

Fragaria [301 154] *Click on name to get more information.*

- **Fragaria chiloensis** [14]
  - **Fragaria chiloensis subsp. lucida** [2]
  - **Fragaria chiloensis subsp. pacifica** [2]
- **Fragaria daltoniana** [2]
- **Fragaria gracilis** [2]
- **Fragaria iinumae** [9 2]
- **Fragaria moschata** [5 1]
- **Fragaria nilgerrensis** [9 1]
- **Fragaria nipponica** [2]
- **Fragaria nubicola** [8 6]
- **Fragaria orientalis** [6]
- **Fragaria pentaphylla** [4]
- **Fragaria vesca** [59 25]
  - **Fragaria vesca subsp. americana** [2]
  - **Fragaria vesca subsp. bracteata** [2 2]
  - **Fragaria vesca subsp. vesca** [10 6]
    - **Fragaria vesca f. alba** [2]
- **Fragaria virginiana** [14]
  - **Fragaria virginiana subsp. glauca** [2]
  - **Fragaria virginiana subsp. platypetala** [4]
  - **Fragaria virginiana subsp. virginiana** [6]
- **Fragaria viridis** [5 1]
- **Fragaria x ananassa** (strawberry) [160
- **Fragaria sp. CFRA 538** [2]

Strawberry sequences

NCBI

TaxBrowser: Rose Family

Rosaceae

Taxonomy ID: 3745
Rank: **family**
Genetic code: Translation table
Mitochondrial genetic code: T

Other names:
rose family[common nam

Lineage(abbreviated)
Eukaryota; Viridiplantae;
Magnoliophyta; eudicotyl

Nucleotide (6509) Prote

NCBI

Entrez Nucleotide

PubMed | Nucleotide | Protein | Genome | Structure | PopSet | Taxonomy | OMIM | Books

Search Nucleotide for `txid3745[Organism]` | Go | Clear

Limits | Preview/Index | History | Clipboard | Details

Display | Summary | Save | Text | Clip Add

Show 20 | **Items 1-20 of 6509** | **Page 1 of 326** | **Select page:** 1 2 3 4 5 6 7 8 9 10 »

1: AF454000 — Protein, Taxonomy
Prunus dulcis cultivar Anxaneta RNase (S) gene, S2 allele, partial cds
gi|18092541|gb|AF454000.1|[18092541]

2: AY061960 — Related Sequences, Protein, Taxonomy
Prunus dulcis cultivar Ferragnes RNase (S) gene, S1 allele, partial cds
gi|17266291|gb|AY061960.1|[17266291]

3: AF323507 — Protein, Taxonomy
Malus x domestica sorbitol dehydrogenase (SDH4) mRNA, complete cds
gi|17225199|gb|AF323507.1|[17225199]

4: AF323506 — Protein, Taxonomy
Malus x domestica sorbitol dehydrogenase (SDH3) mRNA, complete cds
gi|17225197|gb|AF323506.1|[17225197]

5: AF323505 — Protein, Taxonomy
Malus x domestica sorbitol dehydrogenase (SDH2) mRNA, complete cds
gi|17225195|gb|AF323505.1|[17225195]

6: AF323504 — Related Sequences, Protein, Taxonomy
Malus x domestica sorbitol dehydrogenase (SDH1) mRNA, complete cds

NCBI

Entrez Structures

Molecular Modeling Database (MMDB)
and Cn3D

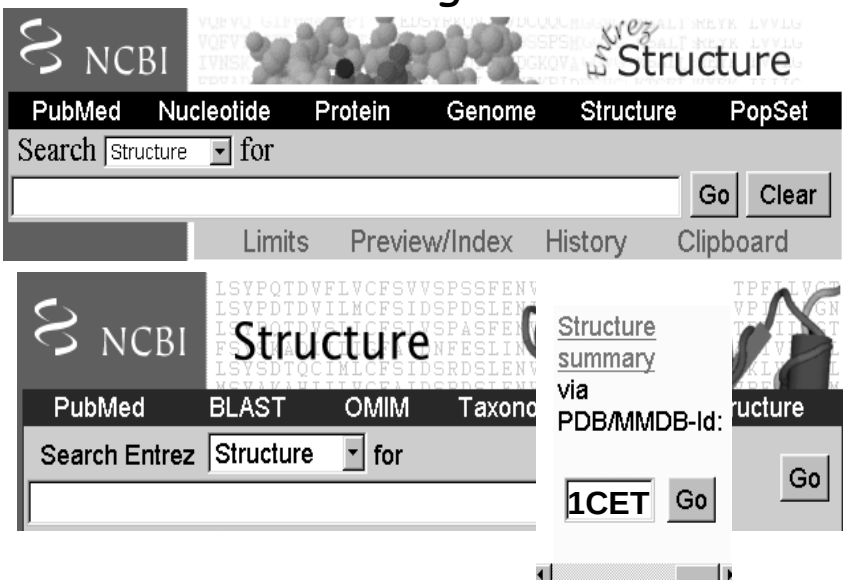
NCBI

MMDB: Molecular Modeling Data Base

- Derived from experimentally determined PDB records
- Value added to PDB records including:
  - Addition of explicit chemical graph information
  - Validation
  - Inclusion of Taxonomy, Citation, and other information
  - Conversion to ASN.1 data description language
- Structure neighbors determined by Vector Alignment Search Tool (VAST)

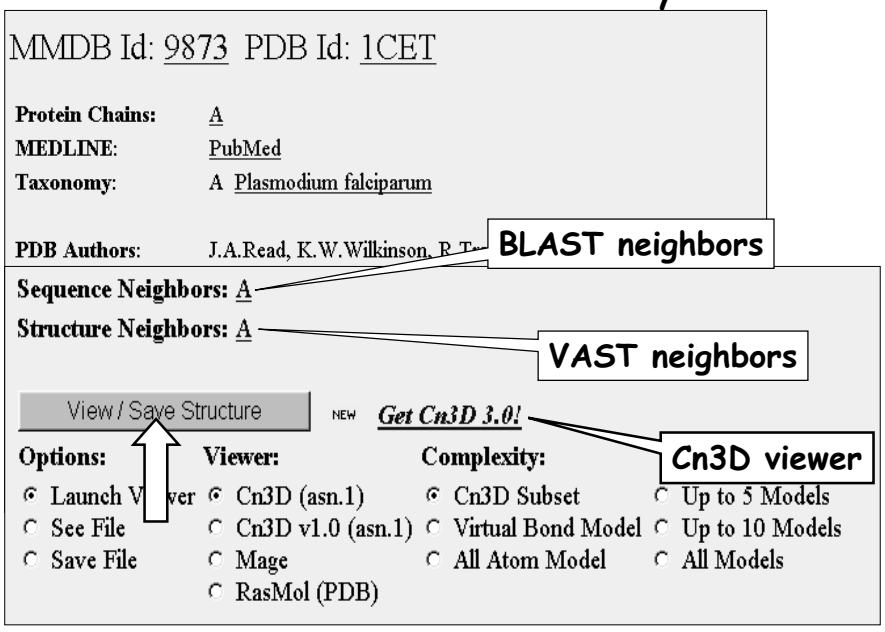
NCBI

Searching MMDB



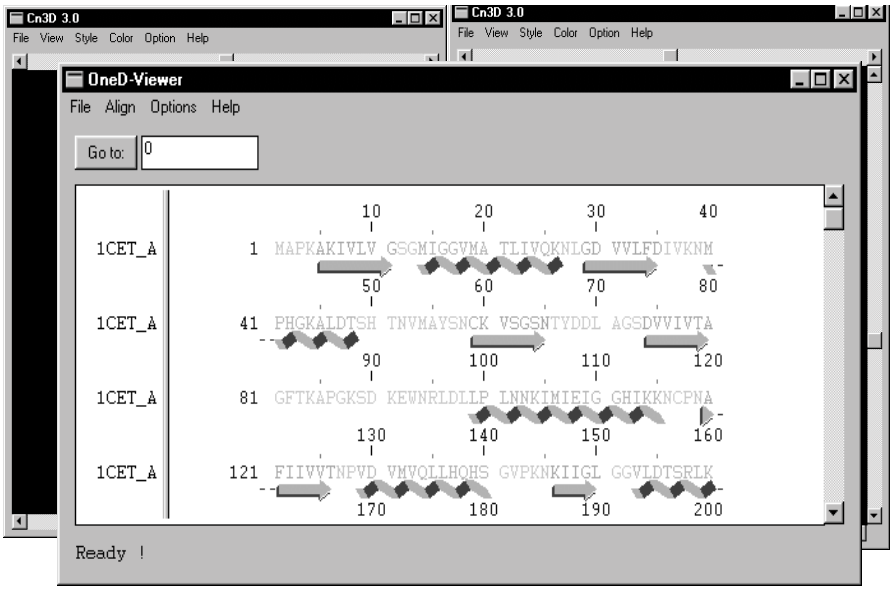
NCBI

Structure Summary



NCBI

Cn3D : Displaying Structures



NCBI

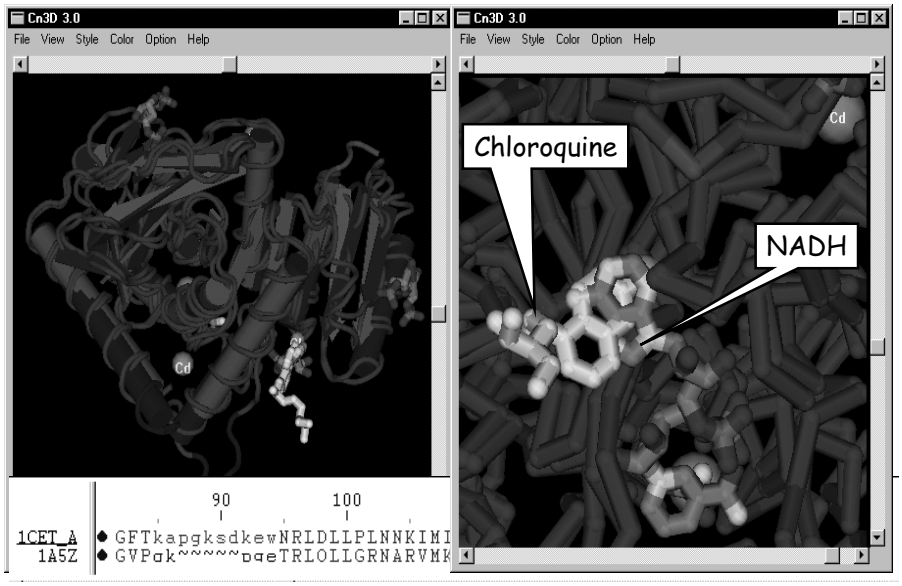
Structure Neighbors

Structure neighbors 1-20 out of 208 displayed. Page 1 of 11.

	PDB	C	D	RMSD	NRES	%Id	Description
☐	1LLD	A		1.8	294	28.9	L-Lactate Dehydrogenase (E.C.1.1.1.27) (T-State) Mutant With Cys 199 Replaced By Ser (C199s) Complex With Nadh
☐	1LDG			0.6	300	100.0	Plasmodium Falciparum L-Lactate Dehydrogenase Complexed With Nadh And Oxamate
☐	1CEQ	A		0.2	304	99.7	Chloroquine Binds In The Cofactor Binding Site Of Plasmodium Falciparum Lactate Dehydrogenase.
☐	6LDH			1.7	298	27.8	M4 Apo-Lactate Dehydrogenase (E.C.1.1.1.27)
☑	1A5Z			1.6	300	36.3	Lactate Dehydrogenase From Thermotoga Maritima (Tmldh)
☐	1LDN	A		1.6	300	28.7	L-Lactate Dehydrogenase (E.C.1.1.1.27) Complexed With Nadh, Oxamate, And Fructose-1,6-Bisphosphate

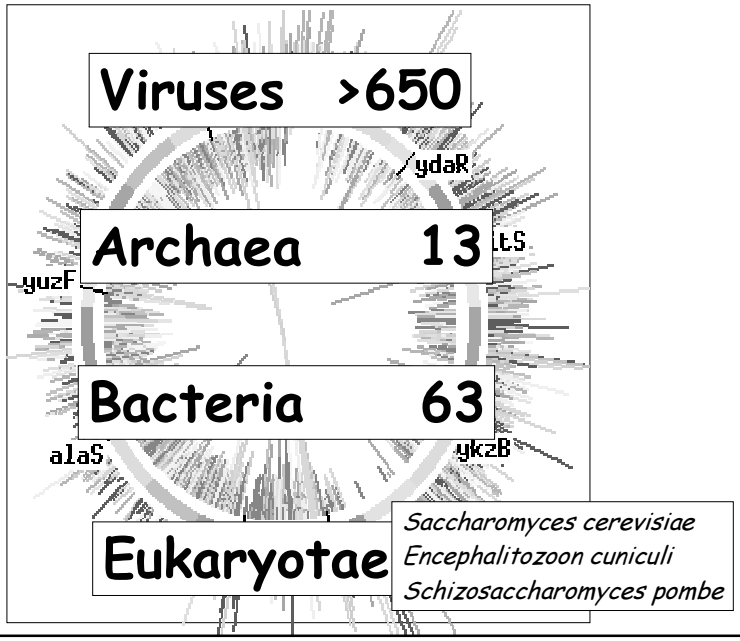
NCBI

Structural Alignments



NCBI

Microbial Genomes in GenBank



arch 26, 2002

Bacterial Genomes

Bacteria Complete Genomes Taxonomy / List 50

Agrobacterium tumefaciens	NC_003062	2841
	NC_003063	2034
Aquifex aeolicus	NC_000918	1551
Bacillus halodurans	NC_002570	4202
Bacillus subtilis	NC_000964	4214
Borrelia burgdorferi	NC_001318	910
Buchnera sp. APS	NC_002528	640
Campylobacter jejuni	NC_002163	1641
Caulobacter crescentus	NC_002696	4016
Chlamydia muridarum	NC_002620	1069
Chlamydia trachomatis	NC_000117	1042
Chlamydophila pneumoniae AR39	NC_002179	1229
Chlamydophila pneumoniae CWL029	NC_000922	1230
Chlamydophila pneumoniae J138	NC_002491	1226
Clostridium acetobutylicum	NC_003030	3940
Deinococcus radiodurans	NC_001263	2648
	NC_001264	412
Escherichia coli K12	NC_000913	4639
Escherichia coli O157:H7	NC_002695	5498
Escherichia coli O157:H7 EDL933	NC_002655	5528
Haemophilus influenzae Rd	NC_000907	1830
Helicobacter pylori 26695	NC_000915	1667
Helicobacter pylori J99	NC_000921	1643
Lactococcus lactis subsp. lactis	NC_002662	2365
Mesorhizobium loti	NC_002678	7036
Mycobacterium leprae	NC_002677	3268203 bp Feb 20 2001
Mycobacterium tuberculosis CDC1551	NC_002755	4403836 bp Apr 25 2001
Mycobacterium tuberculosis H37Rv	NC_000962	4411529 bp Sep 7 2001
Mycoplasma genitalium	NC_[illegible]	580074 bp Jan 8 2001
Mycoplasma pneumoniae	NC_[illegible]12	816394 bp Apr 2 2001
Mycoplasma pulmonis	NC_[illegible]71	963879 bp May 14 1981
Neisseria meningitidis MC58	NC_[illegible]03	2272351 bp Feb 25 2000
Neisseria meningitidis Z2491	NC_003203	2184406 bp Mar 30 2000
Pasteurella multocida	NC_002663	2257487 bp Sep 10 2001
Pseudomonas aeruginosa	NC_002516	6264403 bp Sep 10 2001
Rickettsia conorii	NC_003103	1268755 bp Apr 26 2001
Rickettsia prowazekii	NC_000963	1111523 bp Sep 10 2001
Sinorhizobium meliloti	NC_003047	3654135 bp Jul 30 2001
Staphylococcus aureus subsp. aureus Mu50	NC_002758	2878040 bp Jun 1 2001
Staphylococcus aureus subsp. aureus N315	NC_002745	2813641 bp Apr 21 2001
Streptococcus pneumoniae R6	NC_003098	2038615 bp Jul 27 2001
Streptococcus pneumoniae TIGR4	NC_003028	2160837 bp Jul 25 2001
Streptococcus pyogenes M1 GAS	NC_002737	1852441 bp Apr 10 2001
Synechocystis PCC6803	NC_000911	3573470 bp Mar 21 1997
Thermotoga maritima	NC_000853	1860725 bp Sep 10 2001
Treponema pallidum	NC_000919	1138011 bp Sep 7 2001
Ureaplasma urealyticum	NC_002162	751719 bp Jan 18 2000
Vibrio cholerae	NC_002505	2961149 bp Sep 10 2001
	NC_002506	1072315 bp Sep 10 2001

NCBI

M. tuberculosis Complete Genome

Protein coding genes distribution map
To see map locations of genes, click on a region in

Mycobacterium tuberculosis H37Rv complete genome Microbial genomes

GenBank: NC_000962
Total Bases: 4411529 bp
Completed: Jun 11, 1998.

Feature table:
Protein coding genes: 3918
Structural RNAs: 54

BLAST protein homologs:
COGs (Clusters of Orthologous Groups)
3D Structure (Sequences with known structure)
TaxMap (Sequences grouped by superkingdom)
TaxPlot (3-way genome comparison)
CDD(Conserved Domain Database)

Contributor: Sanger Centre **See genome at** Sanger Centre
Download chromosome sequence data from NCBI FTP site

NEW **BLAST your query sequence against the genome** NEW **BLAST against protein sequences**

Organism: Mycobacterium tuberculosis
Genetic Code: 11
Lineage: Eubacteria; Firmicutes; Actinomycetes; Mycobacteria; Mycobacteriaceae; Mycobacterium.

NCBI

Coding Regions

Mycobacterium tuberculosis H37Rv complete genome

Download [Table], [FASTA protein], [FASTA nucleotide] from NCBI FTP site

◆ - GenBank record including protein ◇ - DNA region in flatfile format ◆ - DNA and protein in FASTA format

	Location	Strand	Length	PID	Gene	COG	Synonym	Product
◆ ◇ ◆	1..1524	+	508	2808711	dnaA	COG0593	Rv0001	dnaA
◆ ◇ ◆	2052..3260	+	403	3261513	dnaN	COG0592	Rv0002	dnaN
◆ ◇ ◆	3280..4437	+	386	1552556	recF	COG1195	Rv0003	recF
◆ ◇ ◆	4434..4997	+	188	1552557	Rv0004			hypothetical protein Rv0004
◆ ◇ ◆	5123..7267	+	715	1552558	gyrB	COG0187	Rv0005	gyrB
◆ ◇ ◆	7302..9818	+	839	1552559	gyrA	COG0188	Rv0006	gyrA
◆ ◇ ◆	9914..10828	+	305	1552560	Rv0007			hypothetical protein Rv0007
◆ ◇ ◆	11874..12311	-	146	1552562	Rv0008c			hypothetical protein Rv0008c
◆ ◇ ◆	12468..13016	+	183	1552563	ppiA	COG0652	Rv0009	ppiA
◆ ◇ ◆	13133..13558	-	142	1552565	Rv0010c			hypothetical protein Rv0010c
◆ ◇ ◆	13714..13995	-	94	1552566	Rv0011c			hypothetical protein Rv0011c
◆ ◇ ◆	14089..14877	+	263	3261646	Rv0012			hypothetical protein Rv0012
◆ ◇ ◆	14914..15612	+	233	1552568	pabA	COG0512	Rv0013	pabA
◆ ◇ ◆	15590..17470	-	627	1552569	pknB	COG0515	Rv0014c	pknB
◆ ◇ ◆	17467..18762	-	432	1552570	pknA	COG0515	Rv0015c	pknA
◆ ◇ ◆	18759..20234	-	492	1552571	pbpA	COG0768	Rv0016c	pbpA
◆ ◇ ◆	20231..21640	-	470	1552572	rodA	COG0772	Rv0017c	rodA
◆ ◇ ◆	21637..23181	-	515	1552573	ppp	COG0631	Rv0018c	ppp
◆ ◇ ◆	23270..23737	-	156	1552574	Rv0019c	COG1716		hypothetical protein Rv0019c
◆ ◇ ◆	23861..25444	-	528	1552575	Rv0020c	COG1716		hypothetical protein Rv0020c
◆ ◇ ◆	25913..26881	-	323	1552576	Rv0021c	COG2070		hypothetical protein Rv0021c
◆ ◇ ◆	27023..27442	-	140	1552577	Rv0022c			hypothetical protein Rv0022c
◆ ◇ ◆	27595..28365	+	257	1552578	Rv0023	COG1396		hypothetical protein Rv0023

NCBI

Genome Annotations

Mycobacterium tuberculosis H37Rv complete genome 818239..868238

Rv0726c fucARv0728c xylB Rv0730 secY adk map sigL

31 Archaea | 94 Bacteria | 68 Metazoa | 0 Fungi | 6 Plants | 0 Viruses | 1 Other Eukaryotae

Keep only No filter Cut-Off 100 Select Reset

390 aa

SCORE	P	ACCESSION	GI	N	ORGANISM
1146	3	AAK23331	13422699	- 7	*Caulobacter crescentus*
1017	2	AAF12736	6465955	- 21	*Homo sapiens*
1001	2	BAB26255	12844144	- 13	*Mus musculus*
989	3	AAG04135	9946632	- 13	*Pseudomonas aeruginosa*
676	7	CAB61663	6468707	- 12	*Streptomyces coelicolor A3(2)*
657	4	BAB07517	10176423	- 4	*Bacillus halodurans*
653	3	AAK18172	13310129	- 1	*Pseudomonas putida*
647	4	CAA89868	853760	- 7	*Bacillus subtilis*
644	3	AAF10828	6459002	- 7	*Deinococcus radiodurans*
640	7	CAB61531	6465864	- 1	*Streptomyces lividans*
597	4	AAA95968	1055219	- 1	*Clostridium acetobutylicum*

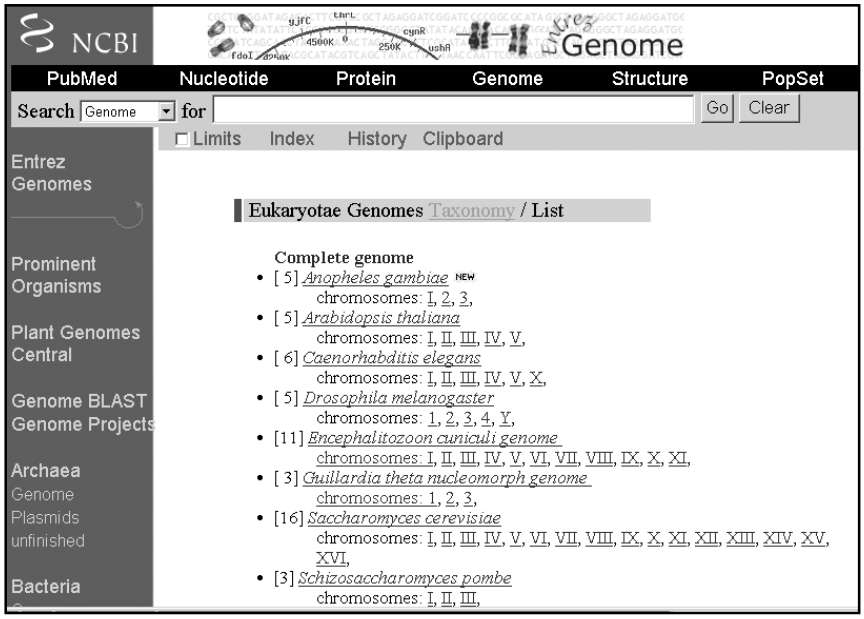
NCBI

M. tuberculosis vs. E.coli COGS

Code	COGs	Description	Code	COGs	Description
J	140	Translation, ribosomal structure and biogenesis	J	167	Translation, ribosomal structure and biogenesis
K	165	Transcription	K	238	Transcription
L	166	DNA replication, recombination and repair	L	207	DNA replication, recombination and repair
D	28	Cell division and chromosome partitioning	D	26	Cell division and chromosome partitioning
O	86	Posttranslational modification, protein turnover, chaperones	O	117	Posttranslational modification, protein turnover, chaperones
M	115	Cell envelope biogenesis, outer membrane	M	190	Cell envelope biogenesis, outer membrane
N	32	Cell motility and secretion	N	119	Cell motility and secretion
P	115	Inorganic ion transport and metabolism	P	184	Inorganic ion transport and metabolism
T	92	Signal transduction mechanisms	T	140	Signal transduction mechanisms
C	192	Energy production and conversion	C	266	Energy production and conversion
G	117	Carbohydrate transport and metabolism	G	324	Carbohydrate transport and metabolism
E	200	Amino acid transport and metabolism	E	343	Amino acid transport and metabolism
F	68	Nucleotide transport and metabolism	F	89	Nucleotide transport and metabolism
H	111	Coenzyme metabolism	H	115	Coenzyme metabolism
I	206	Lipid metabolism	I	86	Lipid metabolism
R	440	General function prediction only	R	356	General function prediction only
S	188	Function unknown	S	290	Function unknown
-	1456	not in COGs	-	1031	not in COGs

NCBI

Entrez Genomes



NCBI

The Arabidopsis Map Viewer

NCBI Entrez Genome

PubMed | Nucleotide | Protein | Genome | Structure | PopSet | Taxonomy | OMIM | Help

Search for At5g44790 on chromosome(s) Find

Show linked entries Help FTP

Entrez Genomes

Prominent organisms

FTP SITE

Related Databases: TAIR TIGR MIPS KAOS

Sequencing Projects: SPP Consortium ESSA Genoscope

***Arabidopsis thaliana* genome view** **BLAST search** Arabidopsis genome

I II III IV V MT CHL

Hits: 1

Search results for query "At5g44790": 1 hit

Chr	Match	Map element	Type	Maps
V	At5g44790	NM_123847	GENE	Gene

were provided to NCBI by TIGR on behalf of the collaborators.

NCBI

Map View: NM

Map Viewer Help
Arabidopsis Maps
Help
FTP

Data As Table View

Maps&Options

Region Shown: 17,787K 17,796K Go

out zoom in

Clone

ideogram
master

***Arabidopsis thaliana* Map Vie**

Chromosome: I II III IV [V]

Query: At5g44790 [clear]

Master Map: Gene

Total Genes On Chromosome: 6000

Region Displayed: 17,787K-17,796K

bp Download/View Sequence/Evide

Genes Labeled: 1 Total Genes in Reg

Gene gene aliases

At5g44790 K23L20.14

Reverse Complement Strand

View on plus strand Protein coding genes Hide Toolbar

Search for gene Find CDS with gene and mRNA Refresh

Hide sequence

Legend: CDS RNA gene sequence fragment shown

Sequence:

NCBI

NM_123847 Maps and Options

Map Viewer Help
Arabidopsis Maps Help
FTP

Data As Table View
Maps&Options

Region Shown:
17805056
17805056

out zoom in

Clone

ideogram
master

Arabidopsis thaliana **Map View**

Chromosome: I II

Master Map: Gene
Total Genes On Ch
Region Displayed: 1
bp Download/View
Genes Labeled: 5 T

Contig Clone

Maps & Options - Netscape

Order of Maps Displayed (left to right):

Contig | Clone | At UniGene clusters
ruler | ruler | ruler
----Select map---- | ----Select map---- | ----Select map----
ruler | ruler
Zm UniGene clusters | ----Select
ruler | ruler

Master Map (rightmost map with desc

Gene | ruler **Chrom** | 17,796K

Options:

Show connections | Verbose Mode

Page Size: 20

Thumb default (clone)

Apply

Contig
Clone
Marker
Gene
At UniGene clusters
At ESTs
Barley UniGene clusters
Barley ESTs
Rice UniGene clusters
Rice ESTs
Wheat UniGene clusters
Wheat ESTs
Maize UniGene clusters
Maize ESTs

At5g44800

NCBI

ClusterAt.24482

Arabidopsis thaliana ATP dependent copper transporter (RAN1) mRNA, complete cds

SEE ALSO
HomoloGene: At.24482

SELECTED MO ORGANISM PROTEIN SIMILARITIES
organism, protein percent identity and length of aligned region

H.sapiens:	35670 - AT7B_HUMAN COPPER-TRANSPORTING ATPASE 2	**38 % / 932 aa** (see ProtEST)
M.musculus:	sp:Q64446 - AT7B_MOUSE COPPER-TRANSPORTING ATPASE 2 (COPPER PUMP 2) (WILSON DISEASE-ASSOCIATED PROTEIN HOMOLOG)	**38 % / 930 aa** (see ProtEST)
R.norvegicus:	sp:P70705 - AT7A_RAT Copper-transporting ATPase 1 (Copper pump 1) (Menkes disease-associated protein homolog)	**38 % / 934 aa** (see ProtEST)
A.thaliana:	sp:Q9S7J8 - AHM5_ARATH COPPER-TRANSPORTING ATPASE RAN1 (RESPONSIVE-TO-ANTAGONIST 1)	**100 % / 1000 a** (see ProtEST)
C.elegans:	ref:NP_499778.1 - Heavy-metal-associated domain containing proteins (2 domains), E1-E2 ATPases (2 domains) [Caenorhab	**37 % / 940 aa** (see ProtEST)
D.melanogaster:	sp:P22700 - ATC1_DROME CALCIUM-TRANSPORTING ATPASE SARCOPLASMIC/ENDOPLASMIC RETICULUM TYPE (CALCIUM PUMP)	**30 % / 166 aa** (see ProtEST)

EXPRESSION INFORMATION
cDNA sources: aboveground organs;roots;seedling hypocotyl

mRNA/GENE SEQUENCES (3)

AF082565	Arabidopsis thaliana ATP dependent copper transporter (RAN1) mRNA, complete cds	P
AF367273	Arabidopsis thaliana AT5g44790/K23L20_14 mRNA, complete cds	P
NM_123847	Arabidopsis thaliana chromosome 5 CHR5v12152001 genomic sequence	P

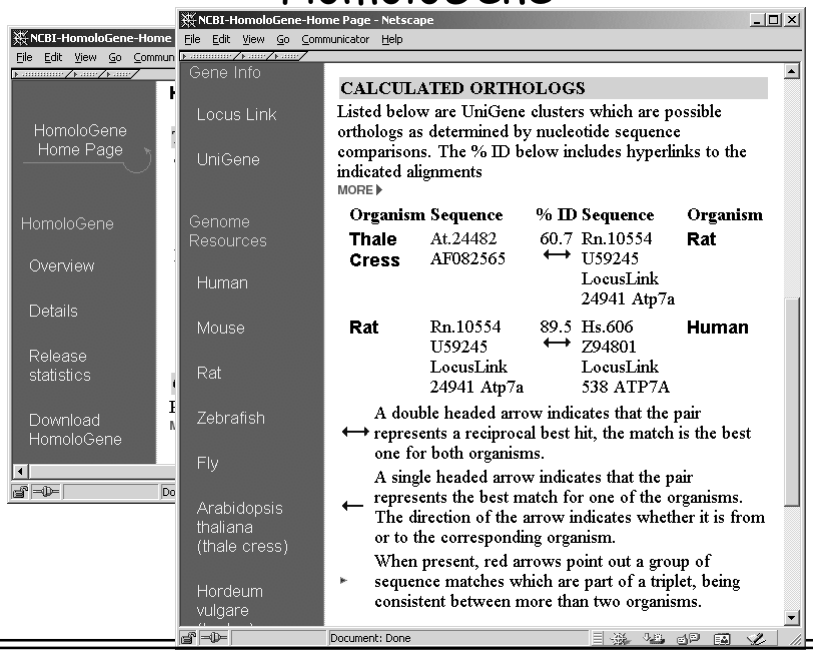
EST SEQUENCES (10 of 23)[Show all ESTs]

W43689	cDNA clone H3H1T7	seedling hypocotyl	P
AV803681	cDNA clone RAFL09-36-G17	3' read	
AV819225	cDNA clone RAFL11-03-A01	3' read	
AV817416	cDNA clone RAFL09-94-F19	3' read	
AV810868	cDNA clone RAFL09-65-L06	3' read	
AV805193	cDNA clone RAFL09-42-B02	3' read	
AV808638	cDNA clone RAFL09-55-L11	3' read	
AV820355	cDNA clone RAFL11-09-N07	3' read	
T46552	cDNA clone 135E16T7		P
AV827274	cDNA clone RAFL09-13-F01	5' read	

Key to Symbols
P Has similarity to known Proteins (after translation)
A Contains a poly-Adenylation signal
M Clone is putatively CDS-complete by MGC criteria

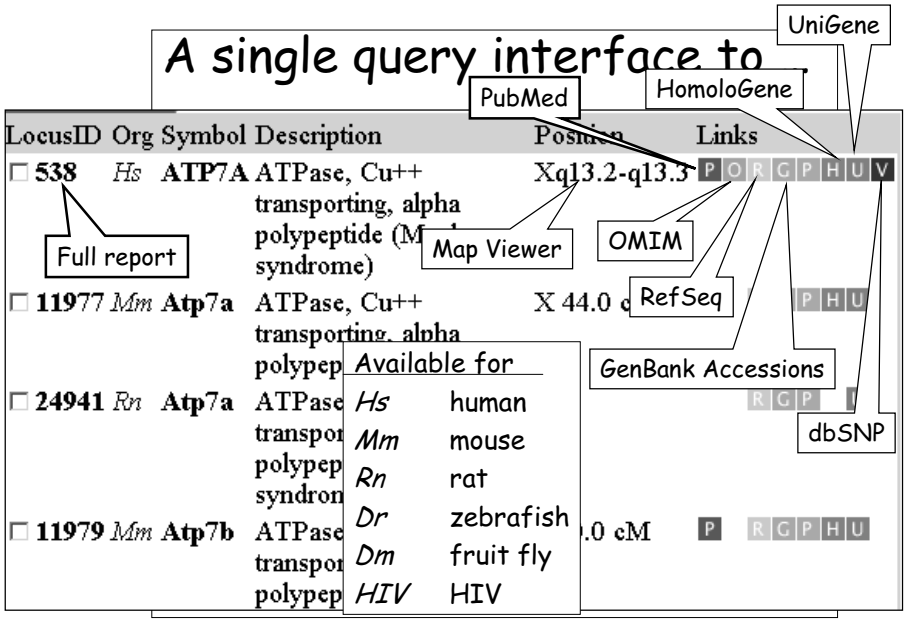
NCBI

HomoloGene



NCBI

LocusLink



LocusLink ATP7A

Click to Display mRNA-Genomic Alignments (spanning 78777 bps)

PUB OMIM UNIGENE MAP VAR HOMOL GDB HGMD e!
UCSC PROTEOME

Homo sapiens Official Gene Symbol and Name (HGNC)

ATP7A: ATPase, Cu++ transporting, alpha polypeptide (Menkes syndrome)

LocusID: 538

Overview ?

Proteome Summary: Copper transporting ATPase alpha polypeptide; important for copper efflux from cells; has six putative metal-binding motifs

Locus Type: gene with protein product, function known or inferred

Product: ATPase, Cu++ transporting, alpha polypeptide

Alternate Symbols: MK, MNK, OHS

Function Submit GeneRIF (All Pubs) ?

Phenotype:
- Cutis laxa, neonatal
- Menkes disease
- Occipital horn syndrome

EC Number: 3.6.1.36

Gene Ontology™:

Term	Evidence	Source	Pub
· Golgi apparatus	E	Proteome	pm
· copper ion transport	P	Proteome	pm
· copper-exporting ATPase	P	Proteome	pm
· integral plasma membrane protein	NR	Proteome	

Other Ontologies:

Term	Evidence	Source	Pub
· Golgi	NR	Proteome	pm
· Integral membrane	NR	Proteome	pm
· Unspecified membrane	P	Proteome	pm

Relationships ?

Mouse Homology Maps:

NCBI vs. MGD	X 44.00 cM	Atp7a	Hs Mm
UCSC vs. MGD	X 44.00 cM	Atp7a	Hs Mm

Links:

pm	PubMed
mv	MapViewer
sv	Sequence Viewer
ev	Evidence Viewer
BL	BLink

Map Information ?

Chromosome: X mv

Cytogenetic: Xq13.2-q13.3 HUGO

Markers:

Chr. X	CHLC.UTR_01549_L06133	mv
Chr. X	SGC31483	mv
Chr. -	AL023617	
Chr. -	GDB:435280	
Chr. -	GDB:435283	
Chr. -	GDB:435286	
Chr. -	GDB:435290	
Chr. -	GDB:435294	

NCBI Reference Sequences (RefSeq) ?

Category: PROVISIONAL

mRNA: NM_000052

Protein: NP_000043 ATPase, Cu++ transporting, alpha polypeptide BL

Domains: E1-E2 ATPase score: 532; Heavy-metal-associated domain score: 196

GenBank Source: L06133

Category: NCBI Genome Annotation

Genomic Contig: NT_030881 sv mv ev

Annotated transcripts/proteins for this locus:

Evidence: none

Model mRNA: XM_013141

Model Protein: XP_013141 BL

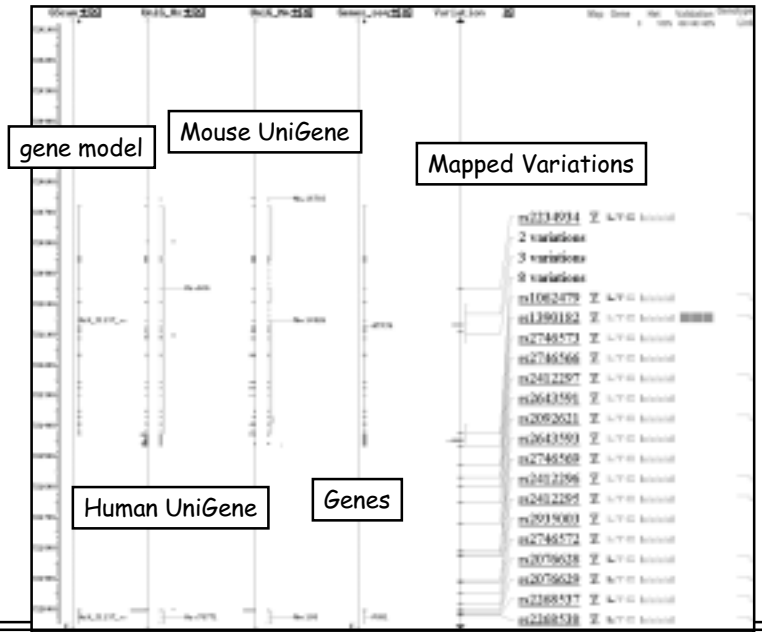
Domains: E1-E2 ATPase score: 589; Heavy-metal-associated domain score: 190; haloacid dehalogenase-like hydrolase score: 200

GenBank Sequences ?

Nucleotide	Type	Protein

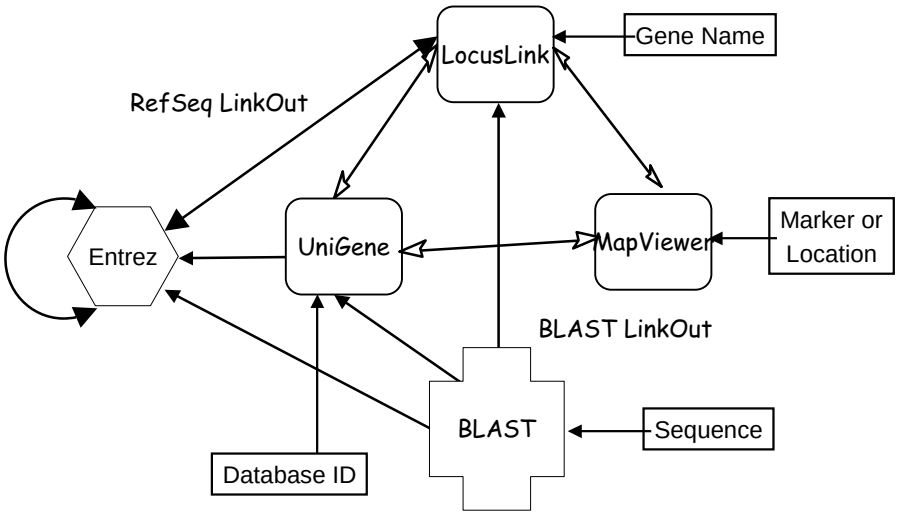
NCBI

Human Map View ATP7A



NCBI

Genome Resources Integration



NCBI

Why do we need similarity searching?

- ✦Identification and annotation
  - •Incomplete or no annotations (GenBank)
  - •Incorrectly annotated sequences
- ✦ Evolutionary relationships
  homologous molecules **may** have similar functions

but it ain't necessarily so!

NCBI

Basic Local Alignment Search Tool

- Widely used similarity search tool
- Heuristic approach based on Smith Waterman algorithm
- Finds best local alignments
- Provides statistical significance
- All combinations (DNA/Protein) query and database.
  - DNA vs DNA
  - DNA translation vs Protein
  - Protein vs Protein
  - Protein vs DNA translation
  - DNA translation vs DNA translation
- www, email server, standalone, and network clients

NCBI

How BLAST Works

- Make lookup table (hash table) for query
- Scan database for hits
- Ungapped extensions of hits
- Gapped extensions (no traceback)
- Gapped extensions (traceback)

NCBI

Look Up Table (Hash Table)

Query: **GTQITVEDLFYNIATRRKALKN**

GTQ

Word Size = 3

TQI

QIT

ITV -> **LTV,MTV,ISV,LSV,MSV**

TVE **IAV,LAV,MAV,ITL,etc.**

VED

EDL

DLF

LFY

FYN

Adjustable

2 or 3 for protein (3 default)

> 7 for blastn searches (11 default)

Neighborhood Words

Make table for both query and database

NCBI

Messy Details

ATCGCCATGCTTAATTGGGCTT

CATGCTTAATT exact word match

one hit

- Nucleotide BLAST looks for exact matches
- Protein BLAST requires two hits

GTQITVEDLFYNI

SEI **YYN** neighborhood words

two hits

NCBI

More Details (BLAST options)

```
-W  Word size
-f  Threshold for extending hits
-X  X dropoff value for gapped alignment (in bits)
-y  Dropoff (X) for blast extensions in bits
-Z  X dropoff value for final gapped alignment (in bits)
-A  Multiple Hits window size (zero for single hit algorithm)
-e  Expectation value (E)    default = 10.0
-q  Penalty for a nucleotide mismatch (blastn only) default = -3
-r  Reward for a nucleotide match (blastn only)    default = 1
-v  Number of database one-line descriptions
-b  Number of database alignments
```

NCBI

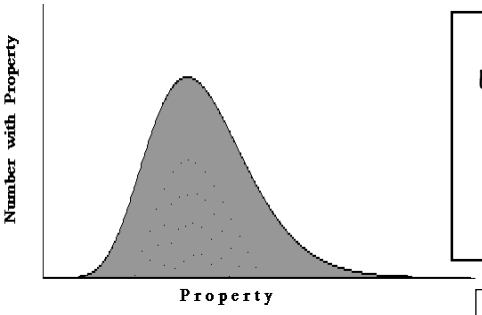
An alignment that BLAST can't find

```
  1 GAATATATGAAGACCAAGATTGCAGTCCTGCTGGCCTGAACCACGCTATTCTTGCTGTTG
    || | || || || |  || || ||   ||  |  ||| |||||| | | || | ||| |
  1 GAGTGTACGATGAGCCCGAGTGTAGCAGTGAAGATCTGGACCACGGTGTACTCGTTGTCG

 61 GTTACGGAACCGAGAATGGTAAAGACTACTGGATCATTAAGAACTCCTGGGGAGCCAGTT
    | || ||     ||  ||| ||  | |||||| || | ||||||  |||||  |     |
 61 GCTATGGTGTTAAGGGTGGGAAGAAGTACTGGCTCGTCAAGAACAGCTGGGCTGAATCCT

121 GGGGTGAACAAGGTTATTTCAGGCTTGCTCGTGGTAAAAAC
    |||| || ||||| ||  ||    |  | ||||  || |||
121 GGGGAGACCAAGGCTACATCCTTATGTCCCGTGACAACAAC
```

NCBI

Local Alignment Statistics

High scores of local alignments between two random sequences follow Extreme Value Distribution



For ungapped alignments:

Expected number with score S or greater

$$E = Kmne^{-\lambda S}$$

or

$$E = mn2^{-S'}$$

K = scale for search space
λ = scale for scoring system
S'= bitscore = $(\lambda S - \ln K)/\ln 2$

http://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html

NCBI

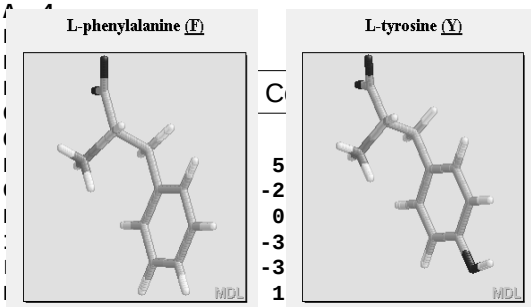
Scoring Systems

	A	G	C	T
A	+1	–3	–3	-3
G	–3	+1	–3	-3
C	–3	–3	+1	-3
T	–3	–3	–3	+1

- Nucleic acids identity matrix
- Proteins
  - Position Independent Matrices
    - PAM Matrices (Percent Accepted Mutation)
      - Implicit model of evolution
      - Higher PAM number all calculated from PAM1
      - PAM250 widely used
    - BLOSUM Matrices (BLOck SUbstitution Matrices)
      - Empirically determined from alignment of conserved blocks
      - Each includes information up to a certain level of identity
      - BLOSUM62 widely used
  - Position Specific Score Matrices (PSSMs)
    - PSI and RPS BLAST

NCBI

BLOSUM62



have low weights

have high weights

```
M -1 -1 -2 -3 -1  0 -2 -3 -2  1  2 -1  5
F -2 -3 -3 -3 -2 -3 -3 -3 -1  0  0 -3  0  6
P -1 -2 -2 -1 -3 -1 -1 -2 -2 -3 -3 -1 -2 -4  7
S  1 -1 -1 -1  0 -1  0  0  0 -1 -2 -2  0 -1 -2 -1  4
                                 1 -1 -1 -2 -1  1  5
                                 2 -3 -1  1 -4 -3 -2 11
Y -2 -2 -2 -3 -2 -1 -2 -3  2 -1 -1 -2 -1  3 -3 -2 -2  2  7
V  0 -3 -3                                -1 -2 -2  0 -3 -1  4
X  0 -1 -1                                -1 -2  0  0 -2 -1 -1 -1
   A  R  N                                 F  P  S  T  W  Y  V  X
```

Negative for less likely substitutions

Positive for more likely substitutions

NCBI

Position Specific Substitution Rates

```
                 210       220       230       240       250       260       270
         ....*....|....*....|....*....|....*....|....*....|....*....|....*....|
154 AT-CRR-AYSGG----------PAITDNMLCAGGLE--GGKDACQGDSGGPLVC-NDP-----RWVLVGI
707 KV-CNR-YEYLG----------GKVSPNELCAGHLA--GGIDSCQGDSGGPLVCfEKD-----KYILQGV
187 KGsCER-DAQYApgydkvkdisEVVTPRFLCTGGVSpyADPNTCRGDSGGPLIV-HKr----sRFIQVGV
166 SS-CKS-AYPG------------QITSNMFCAGYLE--GGKDSCQGDSGGPVVC-S--------GKLQGI
166 TN-CK--KYWG-----------TKIKDAMICAGA----SGVSSCMGDSGGPLVC-KKn----gAWTLVGI
160 SD-CNN-SYPG------------MITNAMFCAGYLE--GGKDSCQGDSGGPVVC-N--------GELQGV
167 KV-CNRyEFLNG-----------RVQSTELCAGHLa--GGTDSCQGDSGGPLVC-FEk----dKYILQGV
156 AT-CSKpGWWG-----------STVKTNMICAGGD---GIISSCNGDSGGPLNC-QGan---gQWQVHGI
156 AI-CSSsSYWG-----------STVKNSMVCAGGD---GVRSGCQGDSGGPLHC-LVn----gQYAVHGV
169 EH-CSQyDWWG                            TRSGCDGDSGGPLNC-
156 HI-CDAkYHLG                            -RDSCQGDSGGPLVC-
168 AT-CSQrDWWG-----------TTVKETMVCAGGE---GVISACNGDSGGPLNC-QAd----gQWDVRGI
165 EE-CS--QTWGN----------NMISDVMICAGA----AGATSCMGDSGGPLVC-QKd----nVWTLVGI
185 NQ-CR--QYWG-----------SSITDSMICAGG----AGASSCQGDSGGPLVC-QKg----nTWVLIGI
518 ER-CSS-PEVHGd---------AFLSG-MLCAGFLEg--gTDACQGDSGGPLVC-Edea-aehRLILRGI
320 DV-CNGaDFYGN-----------QIKPKMFCAGYPe--GGIDACQGDSGGPFVC-EDsisrtpRWRLCGI
183 DE-CEK-AHVQ------------KVTDFMLCVGHLE--GGKDTCVGDSGGPLMC-D--------GVLQGV
160 KN-CDD-AHIA------------NVTGTMLCAGDLA--GGKDTCVGDSGGPLIC-D--------GVLQGL
182 DM-CAR-AYSE------------KVTEFMLCAGLWT--GGKDTCGGDSGGPLVC-N--------GVLQGI
678 RF-CKE-RYKG------------LFTGRMLCAGNLQedNRVDSCQGDSGGPLMC-EKpd---eSWVVYGV
408 QR-CNSrYVYDN-----------LITPAMICAGFLq--GNVDSCQGDSGGPLVT-SNn----nIWWLIGD
475 TL-CNSrQLYDH-----------MIDDSMICAGNLQk-pGQDTCQGDSGGPLTC-EKd----gTYYVYGI
181 AH-CSRyDWWG-----------SLVTTSMVCAGGD---GVLASCNGDSGGPLNC-QNad---gSWDVHGV
176 AE-CAA-ALVNv----------vPVTEQMICAGYAAg--GKDSCQGDSGGPLVS-GD--------KLVGV
181 RE-CNChYQTILeq------ddEVIKQDMLCAGSE---G-HDSCQMDSGGPLVC-RWk----cTWIQVGV
```

Weakly conserved serine

Active site serine

NCBI

Position Specific Score Matrix (PSSM)

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
206	D	0	-2	0	2	-4	2	4	-4	-3	-5	-4	0	-2	-6	1	0	-1	-6	-4	-1
207	G	-2	-1	0	-2	-4	-3	-3	6	-4	-5	-5	0	-2	-3	-2	-2	-1	0	-6	-5
208	V	-1	1	-3	-3	-5	-1	-2	6	-1	-4	-5	1	-5	-6	-4	0	-2	-6	-4	-2
209	I	-3	3	-3	-4	-6	0	-1	-4	-1	2	-4	6	-2	-5	-5	-3	0	-1	-4	0
210	S	-2	-5	0	8	-5	-3	-2	-1	-4	-7	-6	-4	-6	-7	-5	1	-3	-7	-5	-6
211	S	4	-4	-4	-4	-4	-1	-4	-2	-3	-3	-5	-4	-4	-5	-1	4	3	-6	-5	-3
212	C	-4	-7	-6	-7	12	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-7	-4	-4	-5	0	-4
213	N	-2	0	2	-1	-6	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-5	-1	-3	-3	-4	-3
214	G	-2	-3	-3	-4	-4	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	-6	-3	-5	-6	-6	-6
215	D	-5	-5	-2	9	-7	-4	-1	-5	-5	-7	-7	-4	-7	-7	-5	-4	-4	-8	-7	-7
216	S	-2	-4	-2	-4	-4	-3	-3	-3	-4	-6	-6	-3	-5	-6	-4	7	-2	-6	-5	-5
217	G	-3	-6	-4	-5	-6	-5	-6	8	-6	-8	-7	-5	-6	-7	-6	-4	-5	-6	-7	-7
218	G	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6	-7	-7	-5	-6	-7	-6	-2	-4	-6	-7	-7
219	P	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	6	-6	-7	-4	-6	-7	9	-4	-4	-7	-7	-6
220	L	-4	-6	-7	-7	-5	-5	-6	-7	0	-1	6	-6	1	0	-6	-6	-5	-5	-4	0
221	N	-1	-6	0	-6	-4	-4	-6	-6	-1	3	0	-5	4	-3	-6	-2	-1	-6	-1	6
222	C	0	-4	-5	-5	10	-2	-5	-5	1	-1	-1	-5	0	-1	-4	-1	0	-5	0	0
223	Q	0	1	4	2	-5	2	0	0	0	-4	-2	1	0	0	0	-1	-1	-3	-3	-4
224	A	-1	-1	1	3	-4	-1	1	4	-3	-4	-3	-1	-2	-2	-3	0	-2	-2	-2	-3

Serine scored differently in these two positions

Active site nucleophile

NCBI

Gapped Alignments

- Gapping provides more biologically realistic alignments
- Statistical behavior not completely understood for gapped alignments
  - Gapped BLAST parameters must be found by simulations for each matrix
- Affine gap costs = -(a+bk)

  a = gap open penalty b = gap extend penalty

  A gap of length 1 receives the score -(a+b)

NCBI

Scores

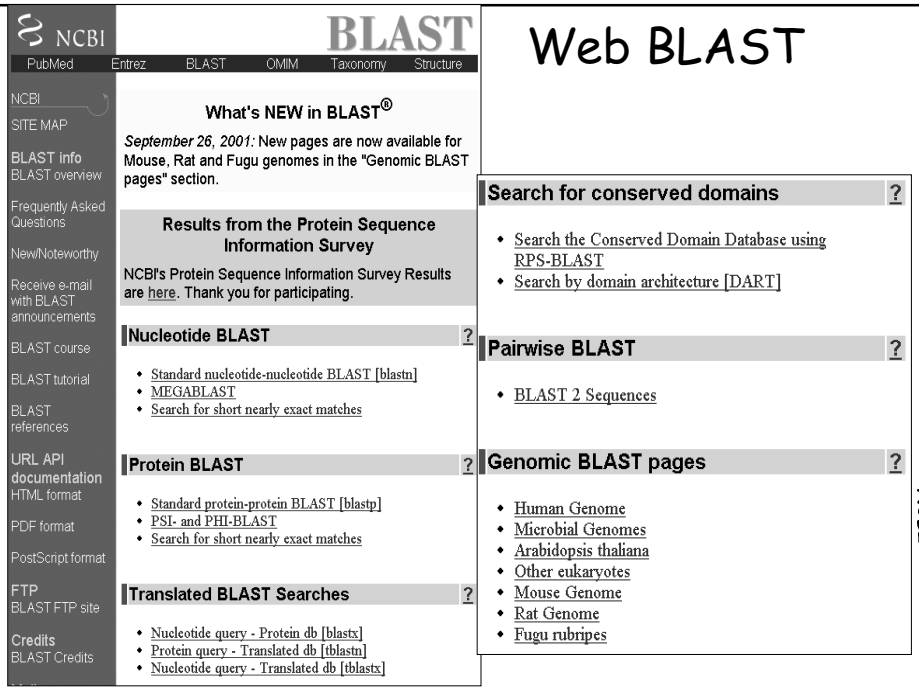
	V	D	S	–	C	Y	
	V	E	T	L	C	F	
BLOSUM62	+4	+2	+1	-12	+9	+3	7
PAM30	+7	+2	0	-10	+10	+2	11

NCBI

WWW BLAST

NCBI

Web BLAST



NCBI

BLAST Databases

Protein (nonredundant)

```
nr          Non-redundant GenBank CDS translations
            PDB+SwissProt+SPupdate+PIR
swissp      Non-redundant SwissProt sequences
pdb         equences
```

Entrez Proteins:
includes swissprot and PDB

Nucleotide (NOT nonredun dant)

Entrez Nucleotides:
without bulk division sequences

```
nr(nt)      GenBank+EMBL+DDBJ+PDB sequences
dbest       Expressed Sequence Tags (EST Division)
htgs        High-Throughput Genome Sequences
            (HTG Division)
chromosome NC_ RefSeqs
```

Higher Genomes

NCBI

Protein BLAST Page

NCBI protein–protein BLAST

Nucleotide | Protein | Translations | Retrieve results for an RID

Search

```
>Mutated in Colon Cancer
IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILER
VQQHIESKLLGSNSSRMYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSS
DKVYAHQMVRTDSREQKLDAFLQPLSKPLSS
```

Set subsequence From: To:

Choose database swissprot

Protein database

Do CD-Search ☑

Now: BLAST! or Reset query Reset all

NCBI

BLAST Formatting Page

NCBI formatting BLAST

Nucleotide | Protein | Translations | Retrieve results for an RID

Your request has been successfully submitted and put into the Blast Queue.

Query = Mutated in Colon Cancer (131 letters)

Putative conserved domains have been detected

Click on the image below for detailed CD-Search results

1 20 40 60 80 100 120 131

DNA_mis_repair

The request ID is 989277607-10255-30769

Format! or Reset all

The results are estimated to be ready in 4 seconds but may be done sooner.

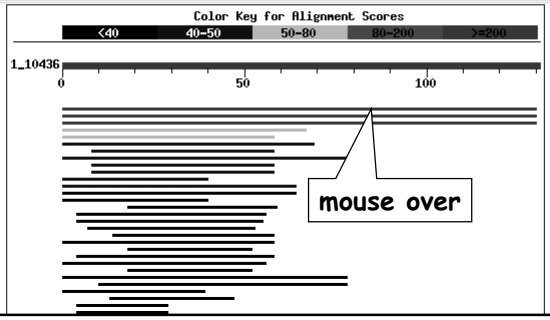
NCBI

BLAST Output: Graphic

Taxonomy reports

Distribution of 30 Blast Hits on the Query Sequence

P40692 DNA mismatch repair protein Mlh1 (MutL protein homolog 1..S= 233 E=8e-62



NCBI

BLAST Output: Descriptions

sorted by e values

4 X 10^{-56}

link to entrez

LocusLink

Default e value cutoff 10

Bacterial mismatch repair proteins

Sequences producing significant a		Score (bits)	E Value
gi|730028|sp|P40692|MLH1 HUMAN	DNA mismatch repair protein ...	233	8e-62
gi|13878583|sp|Q9JK91|MLH1 MOUSE	DNA m	214	4e-56
gi|13878571|sp|P97679|MLH1 RAT	DNA mis otein ...	212	1e-55
gi|1709056| P38920|MLH1 YEAST	MUTL p 1 (DNA...	72	7e-13
gi|1171080| 44494|MUTL HAEIN	DNA mismatch repair protein...	54	7e-08
gi|13431695 57886|MUTL PASMU	DNA mismatch repair protei...	5	1e-06
1035|MUTL THEMA	DNA mismatch repair protein...	48	4e-06
1|MUTL BACHD	DNA mismatch repair protei	46	1e-05
MUTL ECOLI	DNA mismatch repair prote	44	5e-05
gi|127553|sp|P14161|MUTL SALTY	DNA mismatch r	44	7e-05
gi|6225738|sp|Q9ZC88|MUTL RICPR	DNA mismatch	40	7e-04
gi|14194944|sp|Q9PJG5|MUTL CHLMU	DNA mismatch	40	0.001
gi|8928218|sp|O84579|MUTL CHLTR	DNA mismatch repair protein...	39	0.001
gi|20043258|sp|Q9KV13|MUTL VIBCH	DNA mismatch repair protei...	39	0.002
gi|13631230|sp|Q9RP66|MUTL CAUCR	DNA mismatch repair protei...	39	0.002
gi|8928214|sp|O51229|MUTL BORBU	DNA mismatch repair protein...	39	0.002
gi|1709188|sp|P49850|MUTL BACSU	DNA mismatch repair protein...	38	0.005
gi|8039787|sp|O83325|MUTL TREPA	DNA mismatch repair protein...	36	0.013
gi|19856116|sp|P14160|HEX		36	0.020
gi|3914082|sp|P70754|MUTL		35	0.020
gi|11386926|sp|P57633|MUT		35	0.026
gi|8928240|sp|Q9Z794|MUTL CHLPN	DNA misma		
gi|1709684|sp|P54280|PMS1 SCHPO	DNA misma		
gi|3914081|sp|O67518|MUTL AQUAE	DNA mismatch repair pr	34	0.063
gi|1709685|sp|P54278|PMS2 HUMAN	PMS1 protein homolog 2 (D	33	0.16
gi|1709686|sp|P54279|PMS2 MOUSE	PMS1 PROTEIN HOMOLOG 2 (DNA.	32	0.24
gi|8928222|sp|P73349|MUTL SYNY3	DNA mismatch repair protein...	31	0.60
gi|1709683|sp|P54277|PMS1 HUMAN	PMS1 protein homolog 1 (DNA...	31	0.85
gi|126232|sp|P02239|LGB1 LUPLU	Leghemoglobin I	30	1.2
gi|126238|sp|P02240|LGB2 LUPLU	Leghemoglobin II	28	4.1

TaxBLAST: Taxonomy Reports

```
 Haemophilus influenzae [g-proteobacteria] taxid 727
gi|1171080|sp|P44494|MUTL_HAEIN DNA mismatch repair protei...    53  7e-08

 Pasteurella multocida [g-proteobacteria] taxid 747
gi|13431695|sp|P57886|MUTL_PASMU DNA mismatch repair prote...    49  1e-06

 Thermotoga maritima [thermotogales] taxid 2336
gi|8926224|sp|P74926|MUTL_THEMA DNA mismatch repair protei...    47  4e-06

 Bacillus halodurans [low GC Gram+] taxid 86665
gi|20139217|sp|Q9KAC1|MUTL_BACHD DNA mismatch repair prote...    46  1e-05

 Escherichia coli [enterobacteria] taxid 562
gi|127552|sp|P23367|MUTL_ECOLI DNA mismatch repair protein...    44  5e-05

 Salmonella typhimurium [enterobacteria] taxid 602
gi|127553|sp|P14161|MUTL_SALTY DNA mismatch repair protein...    43  7e-05

 Rickettsia prowazekii [a-proteobacteria] taxid 782
gi|6225738|sp|Q9ZC68|MUTL_RICPR DNA mismatch repair protei...    40  7e-04

 Chlamydia muridarum (agent of mouse pneumonitis) [chlamydiae] taxid 83560
gi|14194944|sp|Q9PJ35|MUTL_CHLMU DNA mismatch repair prote...    40  0.001

 Chlamydia trachomatis [chlamydiae] taxid 813
gi|8926218|sp|O84579|MUTL_CHLTR DNA mismatch repair protei...    39  0.001

 Vibrio cholerae [g-proteobacteria] taxid 666
gi|20043250|sp|Q9KV13|MUTL_VIBCH DNA mismatch repair prote...    38  0.002

 Caulobacter vibrioides [a-proteobacteria] taxid 155892
gi|13631230|sp|Q9RP66|MUTL_CAUCR DNA mismatch repair prote...    38  0.002

 Borrelia burgdorferi (Lyme disease spirochete) [spirochetes] taxid 139
gi|8926214|sp|O51229|MUTL_BORBU DNA mismatch repair protei...    38  0.002
```

NCBI

BLAST Output: Alignments

```
>gi|127552|sp|P23367|MUTL_ECOLI  DNA mismatch repair protein mutL
          Length = 615

 Score = 44.3 bits (103), Expect = 5e-05
 Identities = 25/59 (42%), Positives = 33/59 (55%), Gaps = 8/59 (13%)

Query: 9   LPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHF-----LHE---ESILERVQQHIESKL 59
           L  +  P   L LEI P  VDVNVHP KHEV F     +H+   + +L  +QQ +E+ L
Sbjct: 280 LGADQQPAFVLYLEIDPHQVDVNVHPAKHEVRFHQSRLVHDFIYQGVLSVLQQQLETPL 338
```

NCBI

BLAST Output: Alignments

```
>gi|730028|sp|P40692|MLH1_HUMAN  DNA mismatch repair protein Mlh1 1)
          Length = 756

 Score =  233 bits (593), Expect = 8e-62
 Identities = 117/131 (89%), Positives = 117/131 (89%)

Query: 1   IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL 60
           IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL
Sbjct: 276 IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL 335

Query: 61  GSNSSRMYFTQTLLPGLAGPSGEMVKXXXXXXXXXXXXXXDKVYAHQMVRTDSREQKLDA 120
           GSNSSRMYFTQTLLPGLAGPSGEMVK              DKVYAHQMVRTDSREQKLDA
Sbjct: 336 GSNSSRMYFTQTLLPGLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQKLDA 395

Query: 121 FLQPLSKPLSS 131
           FLQPLSKPLSS
Sbjct: 396 FLQPLSKPLSS 406
```

low complexity sequence filtered

NCBI

Results from nr

```
Sequences producing significant alignments:                        (bits) Value

gi|604369|gb|AAA85687.1|  (U17857) hMLH1 gene product [Homo ...   233   3e-61
```

```
>gi|4557757|ref|NP_000240.1|   (NM_000249) mutL homolog 1; mutL (E. coli) homolog 1;
          coli) homolog 1 (colon cancer, nonpolyposis type 2)
          [Homo sapiens]
 gi|730028|sp|P40692|MLH1_HUMAN  DNA mismatch repair protein Mlh1 (MutL protein homolog 1)
 gi|631299|pir||S43085  DNA mismatch repair protein MLH1 - human
 gi|463989|gb|AAC50285.1|(U07343) hMLH1 [Homo sapiens]
 gi|1079787|gb|AAA82079.1|(U40978) DNA mismatch repair protein homolog [Homo sapiens]
 gi|13905126|gb|AAH06850.1|AAH06850   (BC006850) mutL (E. coli) homolog 1
          type 2) [Homo sapiens]
 gi|741682|prf||2007430A  DNA mismatch repair protein [Homo sapiens]
         Length = 756

 Score =  233 bits (593), Expect = 4e-61
 Identities = 117/131 (89%), Positives = 117/131 (89%)

Query: 1   IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL 60
           IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL
Sbjct: 276 IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL 335
```

```
gi|16272041|ref|NP_438240.1|  (NC_000907) DNA mismatch repai...    54   3e-07
gi|19173567|ref|NP_597370.1|  (NC_003232) DNA MISMATCH REPAI...    52   9e-07
gi|13543339|gb|AAH05833.1|AAH05833  (BC005833) Similar to mu...    50   5e-06
gi|15602769|ref|NP_245841.1|  (NC_002663) MutL [Pasteurella ...    50   6e-06
gi|15642797|ref|NP_227838.1|  (NC_000853) DNA mismatch repai...    48   2e-05
```

NCBI

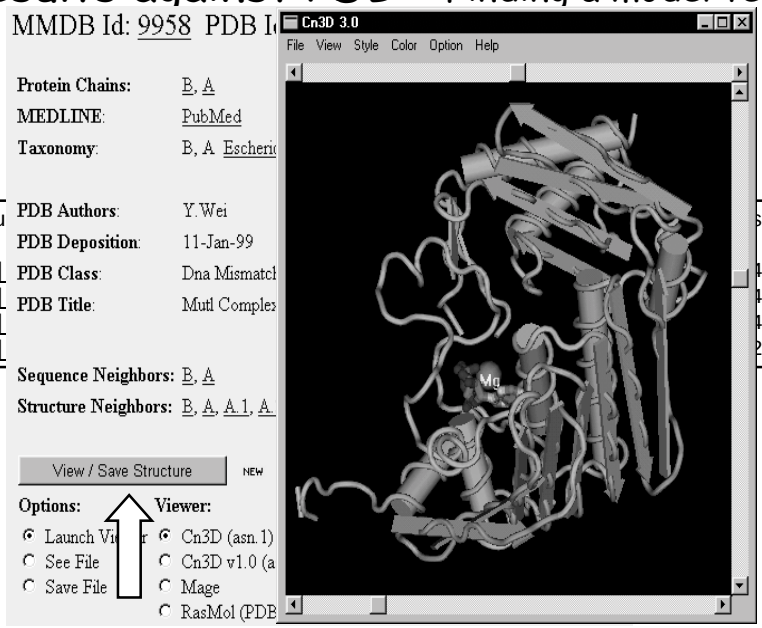
tblastn Results Against ESTs

```
>gi|12794555|emb|AL531062.1|AL531062 AL531062 LTI_NFL001_NBC4 Homo sapiens
 cDNA clone CS0DM005YM23 5
           prime.
          Length = 878

 Score =  167 bits (422), Expect(3) = 1e-42
 Identities = 81/82 (98%), Positives = 81/82 (98%)
 Frame = +2

Query: 1   IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL 60
           IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL
Sbjct: 512 IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL 691

Query: 61  GSNSSRMYFTQTLLPGLAGPSG 82
           GSNSSRMYFTQTLLPGLAGP G
Sbjct: 692 GSNSSRMYFTQTLLPGLAGPLG 757


 Score = 24.3 bits (51), Expect(3) = 1e-42
 Identities = 11/26 (42%), Positives = 11/26 (42%)
 Frame = +1

Query: 80  PSGEMVKXXXXXXXXXXXXXXDKVYA 105
           PSG MVK              DKVYA
Sbjct: 748 PSG*MVKSTTSLTSSSTSGSSDKVYA 825
```

combined expect for hits to multiple frames

NCBI

Results against PDB - Finding a model template

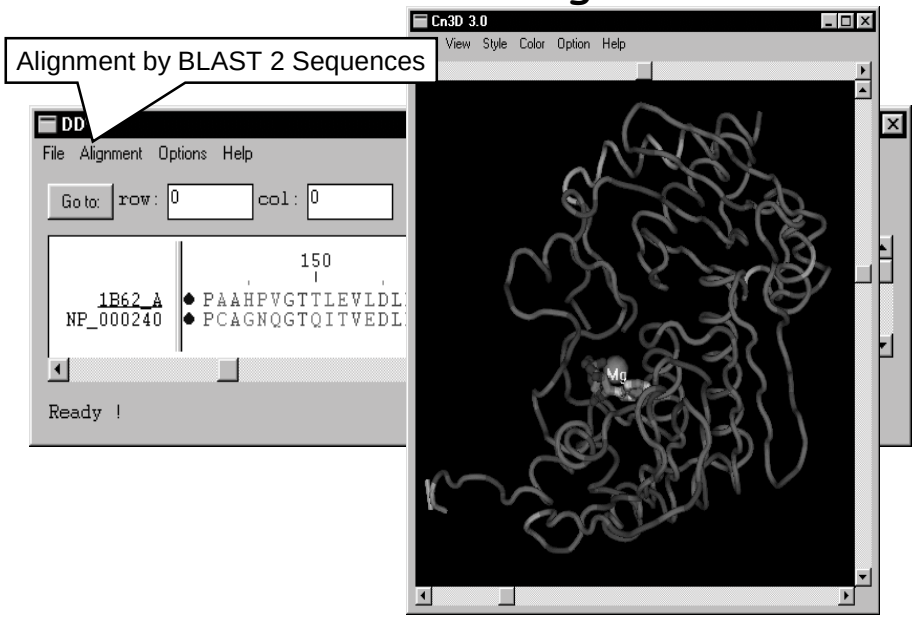


Sequ ... s) Value

pdb| ... 45 1e-05
pdb| ... 45 1e-05
pdb| ... 43 4e-05
pdb| ... 27 2.0

NCBI

Cn3D BLAST Alignment



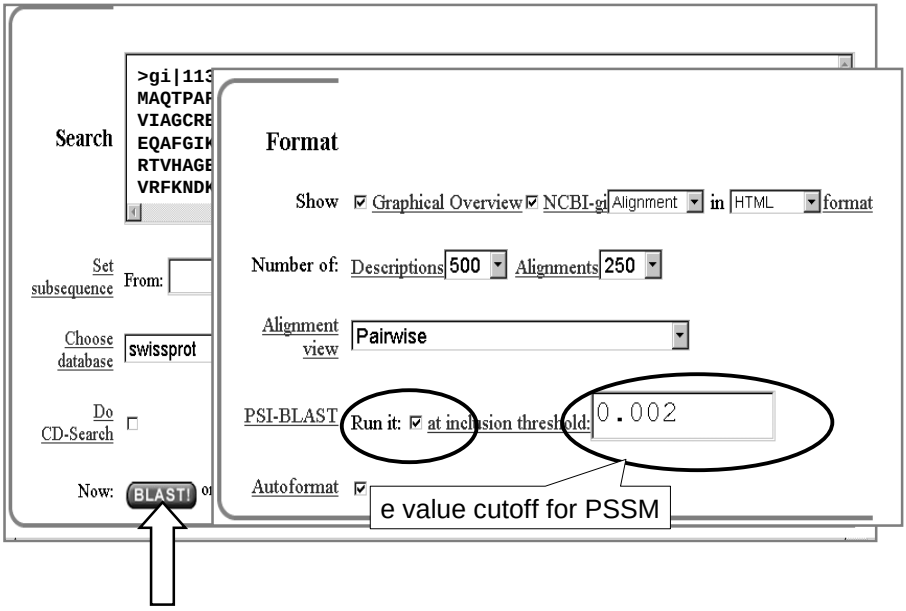
NCBI

PSI-BLAST

Confirming relationships of purine nucleotide metabolism proteins

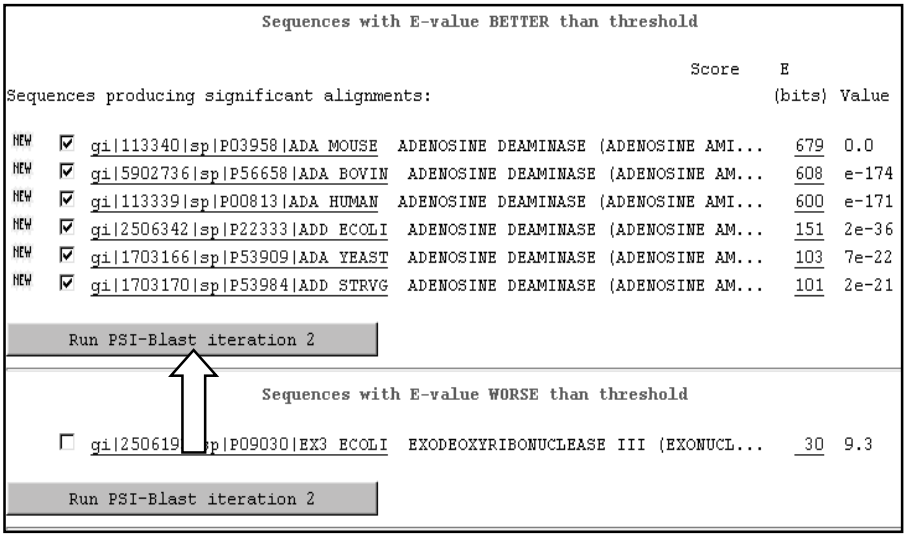
NCBI

PSI BLAST



NCBI

PSI RESULTS: Initial BLAST Run



NCBI

First PSSM Search

Sequences with E-value BETTER than threshold

Sequences producing significant alignments:	Score (bits)	E Value
gi|5902736|sp|P56658|ADA_BOVIN ADENOSINE DEAMINASE (ADENOSINE AM...	618	e-177
gi|113339|sp|P00813|ADA_HUMAN ADENOSINE DEAMINASE (ADENOSINE AMI...	615	e-176
gi|2506342|sp|P22333|ADD_ECOLI ADENOSINE DE...	452	e-127
gi|1703166|sp|P53909|ADA_YEAST ADENOSINE D... ENOSINE AM...	438	e-123
gi|1703170|sp|P53984|ADD_STRVG ADENOSINE ...INASE (ADENOSINE AM...	419	e-117
NEW gi|399033|sp|Q01432|AMD3_HUMAN AMP DEAMINASE 3 (AMP DEAMINASE IS...	53	1e-06
NEW gi|1351916|sp|P15274|AMDM_YEAST AMP DEAMINASE (MYOADENYLATE DEAM...	51	4e-06
NEW gi|2494044|sp|O09178|AMD3_RAT AMP DEAMINASE 3 (AMP DEAMINASE ISO...	49	1e-05
NEW gi|2494043|sp|O08739|AMD3_MOUSE AMP DEAMINASE 3 (AMP DEAMINASE I...	49	2e-05
NEW gi|113698|sp|P10759|AMD1_RAT AMP DEAMINASE 1 (MYOADENYLATE DEAMI...	46	8e-05
NEW gi|1703262|sp|P50998|AMDM_SCHPO AMP DEAMINASE (MYOADENYLATE DEAM...	46	9e-05
NEW gi|399031|sp|Q01433|AMD2_HUMAN AMP DEAMINASE 2 (AMP DEAMINASE IS...	46	1e-04
NEW gi|113697|sp|P23109|AMD1_HUMAN AMP DEAMINASE 1 (MYOADENYLATE DEA...	44	4e-04

Other purine nucleotide metabolizing enzymes not found by ordinary BLAST

Run PSI-Blast iteration 3

NCBI

Third PSSM Search: Convergence

Sequences with E-value BETTER than threshold

Sequences producing significant alignments:	Score (bits)	E Value
gi|5902736|sp|P56658|ADA_BOVIN ADENOSINE DEAMINASE (ADENOSINE AM...	581	e-166
gi|113339|sp|P00813|ADA_HUMAN ADENOSINE DEAMINASE (ADENOSINE AMI...	578	e-165
gi|113340|sp|P03958|ADA_MOUSE ADENOSINE DEAMINASE (ADENOSINE AMI...	569	e-162
gi|2506342|sp|P22333|ADD_ECOLI ADENOS...		
gi|1703166|sp|P53909|ADA_YEAST ADENOS...		
gi|1703170|sp|P53984|ADD_STRVG ADENOS...		
gi|399031|sp|Q01433|AMD2_HUMAN AMP DEAMINASE 2 (AMP DEAMINASE IS...		e-63
gi|1703262|sp|P50998|AMDM_SCHPO AMP DEAMINASE (MYOADENYLATE DEAM...		e-60
gi|2494044|sp|O09178|AMD3_RAT AMP DEAMINASE 3 (AMP DEAMINASE ISO...		-59
gi|399033|sp|Q01432|AMD3_HUMAN AMP DEAMINASE 3 (AMP DEAMINASE IS...	223	58
gi|2494043|sp|O08739|AMD3_MOUSE AMP DEAMINASE 3 (AMP DEAMINASE I...	222	8
gi|113698|sp|P10759|AMD1_RAT AMP DEAMINASE 1 (MYOADENYLATE DEAMI...	217	2e-5
sp|P25524|CODA_ECOLI CYTOSINE DEAMINASE (CYTOSINE AMINOHYDROLASE)	38	0.041
gi|113697|sp|P23109|AMD1_HUMAN AMP DEAMINASE 1 (MYOADENYLATE DEA...	208	2e-53
gi|399032|sp|Q02356|AMD2_RAT AMP DEAMINASE 2 (AMP DEAMINASE ISOF...	134	2e-31
gi|586396|sp|P38150|YB9Z_YEAST HYPOTHETICAL 92.9 KD PROTEIN IN S...	131	1e-30
gi|731934|sp|P40361|YJH0_YEAST HYPOTHETICAL 104.3 KD PROTEIN IN ...	129	6e-30

Just below threshold, another nucleotide metabolism enzyme

NCBI

PHI BLAST

Search

```
>gi|231729|sp|P30429|CED4_CAEEL CELL DEATH PROTEIN 4
MLCEIECRALSTAHTRLIHDFEPRDALTYLEGKNIFTEDHSELISKMSTRLERIANFLRIYRRQASE
LIDFFNYNNQSHLADFLEDYIDFAINEPDLLRPVVIAPQFSRQMLDRKLLLGNVPKQMTCYIREYHV
IKKLDEMCDLDSFFLFLHGRAGSGKSVIASQALSKSDQLIGINYDSIVWLKDSGTAPKSTFDLFTDI
LKSEDDLLNFPSVEHVTSVVLKRMICNALIDRPNTLFVFDDVVQEETIRWAQELRLRCLVTTRDVEI
ASQTCEFIEVTSLEIDECYDFLEAYGMPMPVGEKEEDVLNKTIELSSGNPATLMMFFKSCEPKTFEK
```

Set subsequence From: To:

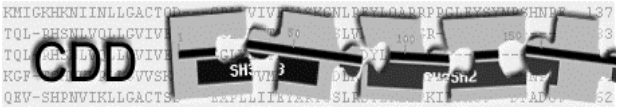
Phi pattern [GA]xxxxGK[ST]

Do CD-Search

Now: BLAST! or Reset query Reset all

NCBI

Conserved Domain Search



NCBI CD-Search Entrez ?

Search the Conserved Domain Database with Reverse Position Specific BLAST

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

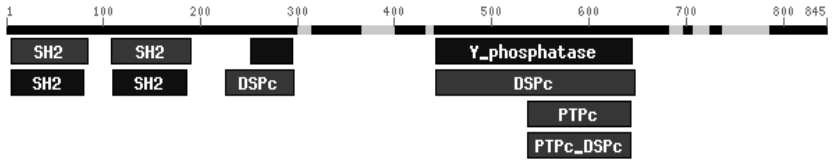
Search Database: All

Enter query as **Protein** Sequence in FASTA format Submit Query Reset

```
>gi|7290263|gb|AAF45724.1| CG3954 gene product [alt 2] [Drosop
MSSRRWFHPTISGIEAEKLLQEQGFDGSFLARLSSSNPGAFTLSVRRGNEVTHIKIQNNGDF
FDLYGGEKFATLPELVQYYMENGELKEKNGQAIELKQPLICAEPTTERWFHGNLSGKEAEKL
ILERGKNGSFLVRESQSKPGDFVLSVRTDDKVTHVMIRWQDKKYDVGGGESFGTLSELIDHY
KRNPMVETCGTVVHLRQPFNATRITAAGINARVEQLVKGGFWEEFESLQQDSRDTFSRNEGY
KQENRLKNRYRNILPYDHTRVKLLDVEHSVAGAEYINANYIRLPTDGDLYNMSSSSESLNSS
VPSCPACTAAQTQRNCSNCQLQNKTCVQCAVKSAILPYSNCATCSRKSDSLSKHKRSESSAS
```

NCBI

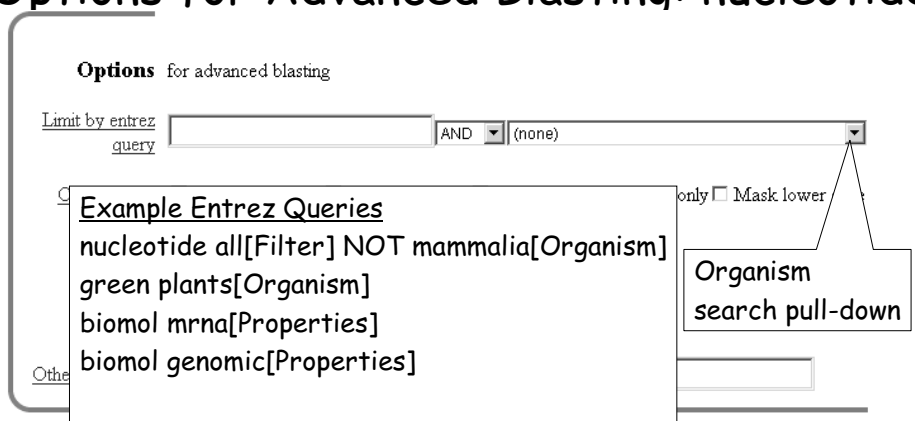
CDD Results




```
                                                                    Score   E
Sequences producing significant alignments:                        (bits) value

gnl|Pfam|pfam00102 Y_phosphatase, Protein-tyrosine phosphatase      236  3e-63
gnl|Pfam|pfam00102 Y_phosphatase, Protein-tyrosine phosphatase     55.4  1e-08
gnl|Smart|DSPc     Dual specificity phosphatase, catalytic domain   236  3e-63
gnl|Smart|DSPc     Dual specificity phosphatase, catalytic domain  70.2  4e-13
gnl|Smart|PTPc     Protein tyrosine phosphatase, catalytic domain   102  9e-23
gnl|Smart|PTPc_DSPcProtein tyrosine phosphatase, catalytic domain, un...102  9e-23
gnl|Smart|SH2      Src homology 2 domains; Src homology 2 domains bi...88.2  1e-18
gnl|Smart|SH2      Src homology 2 domains; Src homology 2 domains bind 76.9  4e-15
gnl|Pfam|pfam00017 SH2, Src homology domain 2                      78.0  2e-15
gnl|Pfam|pfam00017 SH2, Src homology domain 2                      71.8  1e-13
```

NCBI

Options for Advanced Blasting: nucleotide



Example Entrez Queries

nucleotide all[Filter] NOT mammalia[Organism]

green plants[Organism]

biomol mrna[Properties]

biomol genomic[Properties]

OtherAdvanced

-W 7 word size to 7

–e 10000 expect value

-v 2000 descriptions

-b 2000 alignments

NCBI

Options for Advanced Blasting: protein

Options for advanced blasting

Limit by entrez query | AND | (none)

Composition-based statistics ☑

PHI pattern

Organism search pull-down

Example Entrez Queries

proteins all[Filter] NOT mammalia[Organism]

green plants[Organism]

srcdb refseq[Properties]

OtherAdvanced

-W 2 word size to 2

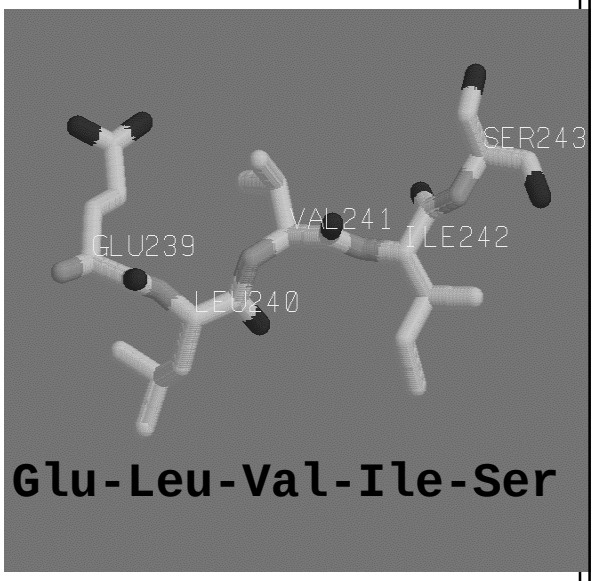
–e 10000 expect value

-v 2000 descriptions

-b 2000 alignments

NCBI

The Elvis Problem: Short Sequences BLAST



Finding Hits with Short Sequences

Composition-based statistics ☐

Choose filter ☐ Low complexity ☐ Mask for lookup table only ☐ Mask lower case

Expect 20000

Word Size 2

Matrix PAM30 Gap Costs Existence: 9 Extension: 1

245 Elvises

```
Sequences producing sign                                    (bits) Value

gi|1750206|gb|AAC47412.1                          ply...    18   7266
gi|18485040|ref|XP_08055                          hil...    18   7266
gi|14521730|ref|NP_12720                          ote...    18   7266
gi|15604448|ref|NP_220966.1|  (NC_000963) TRANSCRIPTION-REPA...    18   7266
gi|18311275|ref|NP_563209.1|  (NC_003366) probable 8-oxoguan...    18   7266
gi|7299701|gb|AAF54883.1|  (AE003698) CG7472 gene product [D...    18   7266
gi|18483836|ref|XP_080244.1|  (XM_080244) CG1918 [Drosophila...    18   7266
gi|249412|gb|AAB22177.1|  colonization factor antigen I, CFA...    18   7266
gi|11181880|emb|CAC16114.1|  (AL161931) bA1021O19.1 (zinc fi...    18   7266
gi|17467138|gb|AAL40101.1|  (L76577) immunoglobulin light ch...    18   7266
gi|6723184|dbj|BAA89600.1|  (AB029894) p3vc [rice grassy stu...    18   7266
gi|511635|gb|AAA35485.1|  (L24774) delta3, delta2-enoyl-CoA ...    18   7266
gi|17231439|ref|NP_487987.1|  (NC_003272) cobalt transport p...    18   7266
```

```
Query: 1   ELVIS 5
           ELVIS
Sbjct: 55 ELVIS 59
```

NCBI

Monkeys Typing

```
>dbj|BAB12211.1| (AB032549) polyketide synthase and peptide synthetase [Microcystis
          aeruginosa]
         Length = 3487

 Score = 22.2 bits (45), Expect =   294
```

```
>gb|AAF63753.1|AF242291_1 (AF242291) nuclear protein EAST [Drosophila melanogaster]
         Length = 2362

 Sco
 Ide

Query

Sbjc
```

```
>pdb|1A59|    Cold-Active Citrate Synthase
          Length = 378

 Score = 23.5 bits (48), Expect =   120
 Identities = 8/12 (66%), Positives = 9/12 (74%), Gaps = 1/12 (8%)

Query: 1   ELVISPRESLEY 12
           EL I PRE L+Y
Sbjct: 145 EL-IEPREDLDY 155
```

NCBI

Large Sequences

```
>193,787 bases
GAATTCAAGTGTCTTATTTCTTCAGGTAATACGGCAATTGTACTAGTTGGGAAATGAAATATCAAA
GGCTTCCAGTTTTAACTAACGAGTCAATTTATATGTATTAAAGCTGTCCTGGGCTTGTTTAAAAGA
```

Other advanced options: **-e 1e-34 -W 30**

```
TATCTTTGCAAACCATTTAAGAGTTTTCTGACTGACAAAATAAAGTAGGCTGGAAGTTTAAAAGAA
ATGTCTGGCTTCTGACAGGATCCTAGCTTTAGATGTTTCTATACGGTTTGTATAAATGCCAGCCCT
AAGCTATTCAATCTTTTTTCTTTTTCTTTTTTTTTCTTTTTAAAACTTTCTGACTAGATCAAGACA
```

ERROR: Bla**Megablast**mit exceeded

```
ATAAAAATTTTTTGGCTGGTAAATATATAAAGAAATCTTGAATAATTAAAAGTAATATGGGGAACAC
ATTCAAGAATGATCTCATTCTCAAACCGTTTGTTTCAAAGATATGTTAAGACATTTAAAGGACTCC
AGTGATGAATTAATTCATTAAAAATGACATGAGCAATATTAAATGATAGAATTATAATAACAGGAA
ATGTTGGTAATTCTATGGGTAGTTCATTTACTTGGAAAAGTATAGCATCCCTGGCTGGGCACAGTG
ATACCTGTAATCCCAGTGCTTTGGGAGGCTGAGTTGGGAGGATAGCTCGAGGCCAGAAGTTTGAGA
CCTGGGCAGCATAGCAAGACCGTGTCTTTACAAACGATTTTATATATATGTGTATATATATTCAAT
TATAAATATGTAAATAATTTATATATAAAATTATATAAAATATGTAAATAATTTATATACATAAAA
```

NCBI

MegaBlast

```
> 4788 gnl|UG|Os#S4788 96BS0324 Oryza sativa cDNA /clone=96BS0324
GNAATTGTAATACGACTCACTATAGGGCGAATTGGGTACCGGGCCCCCCCTCGAGTTTTT
TTTTTTTTTTTTTTGAACTGAAATCTCCGATACTAATAAGTTATAAATAGAGGGGAACTA
GCTAACATTCTCCATAACATCATCCAGTACCATAGTAAGGCTGCTGCTAGTTGCATAGCC
CGATAAGAGTCTCACACAAAGCACAGAAGGTTAAGAATGGGAAAGACCGAAAAATCACAA
GAGAAGCTAAAACAATTCTTAGAGCTAGCTAATCACGTCTTTCTTGCTCCATCCCATTCG
CCTCTTTGCCACATGCCACGGCTGCTCGGCCTCGGCAGCCTCCCTCACTTGTACTCCAGG
ATCATGTTGTTCTCGGGGGCCAGCCCAACGCAGGCCCTCATGATGTTCTCAAGCATTGCC
CTCTGCTTTGCCAGGGCGTTCACCACTGGTGTGCCAAAAAGAACAAGGGGTGCCTTGGTG
AAGTACTCAGGATGGTANCCACTGGATGGAAGGATGGACTCCCCNCCCCCCGNTTTCACT
GATCCTGGTGCTAACTCGGAAGGACACAAATCAAAAGATCGGGCGGAAGGATNATCNCCA
GGTTTNTCAAAAAAGTGCCTACCCANAAATTCGAGTTNNCTCATGCNCTGNTCCAAAATT
> 70988 gnl|UG|Os#S70988 H061C07 Oryza sativa cDNA /clone=H061C07
GGCTACCATCCTGAGCTACCTCACCAAGGCACCCCTTGTTCCTCCTGGCACACCAGTGGT
GAACGCCCATGGCAAAGCAGAGGGCAATGCTTGAGAACATCATGAGGGCCTGCGTTGGGC
TGGCACCCCGAGAACAACATGATCCTGGAGTACAAGTGAGGGAGGCTGCCGAGGCACGAG
CAGCCGTGGCATGTGGCAAAGAGGCGAATGGGATGGAGCAAGAAAGACGTGATTAGCTAG
CTCTAAGAATTGTTTTAGCTTCTCTTGTGATTTTTCGGTCTTTCCCATTCTTAACCTTCT
GTGCTTTGTGTGAGACTCTTATCGGGCTATGCAACTAGCACAGCCTTACTATGGTACTGG
ATGATGTTATGGAGAATGTTAGCTAGTTCCCCTCTATTTATAACTTATTAGTATCGGAGA
TTTCAGTTCAAAAAAA
> 69736 gnl|UG|Os#S69736 H030C04 Oryza sativa cDNA /clone=H030C04
AGCTACCTCACCAAGGCACCCCTTGTTCCTCCATGGCACACCAGTGGTGAACGCCTGGCA
AAGCAGAGGGCAATGCTTGAGAACATCATGAGGGCCTGCGTTGGGCTGGCCCCCGAGAAC
ACATGATCCTGGAGTACAAGTGAGGGAGGCTCCGAGGCCGAGCACCGTGGCATGTGGCAA
AGAGGCGAATGGGATGGAGCAAGAAAGACGTGATTAGCTAGCTCTAAGAATTGTTTTAGC
TTCTCTTGTGATTTTTCGGTCTTTCCCATTCTTAACCTTCTGTGCTTTGTGTGAGACTCT
TATCGGGCTATGCAACTAGCAGCAGCCTTACTATGGTACTGGATGATGTTATGGAGAATG
TTAGCTAGTTCCCCTCTATTTATAACTTATTAGTATCGGAGATTTCAGTTCAAAAAAAAA
AAAAAAA
> 58169 gnl|UG|Os#S58169 AU063597 Oryza sativa cDNA /clone=C63051_1A
TAGTTCTCGATNGCGGCGGCGAGGCGAACTGGGATGGAGCAAGAAAGACGTGATTAGCTA
```

NCBI

BLAST: standalone, clients, databases

Current directory is /blast

Up to higher level directory
- README
- blasturl/
- db/
- documents/
- executable
- fmerge/
- matrices/
- network/
- old/
- server/
- temp/

```
ftp> open ftp.ncbi.nih.gov
.
.
ftp> cd blast
```

ftp://ftp.ncbi.nih.gov/blast/

NCBI

BLAST Batch Client

```
C:\Netblast>blastcl3 –i input.seq –d nr –p blastn –o outfile
```

```
National Center for Biotechnology Information (NCBI)

welcome to the blast network service.
BLASTN 2.2.2 [Dec-14-2001]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= gi|19343352|gb|AAF15280.2|AF192502_1 aryl hydrocarbon receptor
[Gallus gallus]
         (535 letters)

Database: All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS,
or phase 0, 1 or 2 HTGS sequences)
           1,214,651 sequences; 1,071,392,519 total letters
```

NCBI

Genomic BLAST Pages

Genomic BLAST pages ?

- Human Genome
- Mouse Genome
- Rat Genome
- Fugu rubripes
- Zebrafish Genome
- Anopheles gambiae
- Arabidopsis thaliana
- Oryza sativa
- Other eukaryotes
- Microbial Genomes

NCBI

Microbial Genomes BLAST

BLAST with microbial genomes (126 bacterial/1

Currently available for BLAST searches are sequences fro
which GenBank has received a completed genomic sequen
for which only a partial ger
center. See [About the Dat
P - indicates the ability to s
- unfinished genomic se

Enter your query sequen
>APE0122
MVGVFGRLSRHVWVKRW
VAELEDLTMDEIVEMAK
TFKKLKPAVEEEARKEG
Select type of query and
Query: Protein Datal
You may change BLAST
Expect: 10 Filter
Select all Clear all 1

- Archaea
  - Crenarchaeota
    - Desulfurococcales
      - P Aeropyrum perr
    - Sulfolobales
      - P Sulfolobus solfat
      - P Sulfolobus tokod
  - Euryarchaeota
    - Archaeoglobales
      - P Archaeoglobus f
    - Halobacteriales
      - P Halobacterium s
    - Methanobacteriales
      - P Methanothermob
    - Methanococcales
      - P Methanococcus j
    - Methanosarcinales
      - Methanosarcina
    - Thermococcales
      - P Pyrococcus abys
      - Pyrococcus furio
- Pasteurellaceae
  - Actinobacillus actinomycetemcomitans
  - Haemophilus ducreyi
  - P Haemophilus influenzae Rd
  - P Pasteurella multocida
- Pseudomonas
  - P Pseudomonas aeruginosa
  - P Pseudomonas fluorescens
  - Pseudomonas putida KT2440
  - Pseudomonas putida PRS1
  - Pseudomonas syringae
- Salmonella
  - Salmonella enterica subsp. enterica serovar Dublin
  - Salmonella enteritidis
  - Salmonella paratyphi
  - Salmonella typhi
  - Salmonella typhimurium LT2
- Spirochaetales
  - P Borrelia burgdorferi
  - Treponema denticola
  - P Treponema pallidum
- Thermotogales
  - P Thermotoga maritima
- Thermus/Deinococcus group
  - P Deinococcus radiodurans

Hits to Unfinished Genome

Sequences producing significant

gnl|TIGR_61435|6428 Dehalococc
ref|NC_000918.1| Aquifex aeoli
gnl|DOE_2021|2351476_fasta.scre
gnl|DOE_2108|2351478_fasta.scre
gnl|TIGR_881|1531 Desulfovibri
ref|NC_002677.1| Mycobacterium
gnl|TIGR_1772|3270 Mycobacteri
emb|AL137778.1|SCL2 Streptomyc
gnl|CBCUMN_2770|paratb.fasta.sc
gnl|TIGR_35554|2947 Geobacter
gnl|TIGR_1764|3294 Mycobacteri
ref|NC_002755.1| Mycobacterium
gnl|Sanger_1765|mbovis_Contig20
gnl|TIGR_164513|mtub210_338 My
ref|NC_000962.1| Mycobacterium
emb|AX127148.1|AX127148 Sequen
gnl|Sanger_1717|Corynebacterium
ref|NC_002163.1| Campylobacter
gnl|TIGR_1097|3499 Chlorobium
ref|NC_000915.1| Helicobacter
ref|NC_000921.1| Helicobacter
gnl|SANGER_36826|Contig101 Clo
gnl|DOE_63737|Contig554 Nostoc
ref|NC_003030.1| Clostridium a
ref|NC_002570.1| Bacillus halo
gnl|SANGER_36826|Contig95 Clos
gnl|TIGR_1392|6277 Bacillus an
gnl|TIGR_1392|gba18 Bacillus a
ref|NC_003296.1| Ralstonia sol
gnl|TIGR_164513|mtub210_bmt782
gnl|TIGR_164513|mtub210_449 My
gnl|Sanger_1765|mbovis_Contig27

Entrez Genomes

Contributing Genome Centers

ASTRA
SubtiList
BSNR
ChGP
Columbia
Diversa
Genome Therapeutics
Genoscope
GTC
Heidelberg
KDRI
NITE
OU-ACGT

Thermobifida fusca, unfinished sequence

Accession: NC_002721
Number of Contigs: 64
Total Bases Sequenced: 3656222 bp
Last updated: Mar 12, 2002.

BLAST Search

Contributor: DOE-JGI

Organism: Thermobifida fusca
Genetic Code: 11
Lineage: Bacteria; Firmicutes; Actinobacteria; Actinobacteridae; Actinomycetales; Streptosporangineae; Nocardiopsaceae; Thermobifida.

COMMENT JGI Data Release Policy: The Joint Genome Institute releases data as early and often as updates are feasible. All data is available to our scientific colleagues for the purpose of searching for genes of interest for downstream biological characterization. Please recognize that these data are preliminary, and therefore may contain errors, and are subject to change (particularly with respect to contiguity) in subsequent releases. The JGI asks that you acknowledge the source of information obtained from this site in any publication benefiting from these data by including the following sentence in both the Materials and Methods and Acknowledgement sections: 'Preliminary sequence data was obtained from The DOE Joint Genome Institute (JGI)' at

NCBI

BLAST with At Genome

Arabidopsis thaliana

BLAST Home Page
BLAST overview
BLAST FAQs
BLAST news
BLAST manual

BLAST with Arabidopsis thaliana genome

Compare your query sequence to genomic DNA sequence, mRNAs or protein products.

Database: clone Program tblastn MegaBlast: ☐

Begin Search

Enter an accession, gi, or a sequence in FASTA format:

```
>gi|13878337|sp|Q13315|ATM_HUMAN Serine-protein kinase ATM (
MSLVLNDLLICCRQLEHDRATERKKEVEKFKRLIRDPETIKHLDRHSDSKQGKYLNWDAV
ETECLRIAKPNVSASTQASRQKKMQEISSLVKYFIKCANRRAPRLKCQELLNYIMDTVKD
CSNILLKDILSVRKYWCEISQQQMLELFSVYFRLYLKPSQDVHRVLVARIIHAVTKGCCS
DFFSKAIQCARQEKSSSGLNHILAALTIFLKTLAVNFRIRVCELGDEILPTLLYIWTQHR
ELFQLQIYIHHPKGAKTQEKGAYESTKWRSILYNLYDLLVNEISHIGSRGKYSSGFRNIA
```

Optional parameters

Expect	Filter	Descriptions	Alignments
0.01	default	100	100

Advanced options:

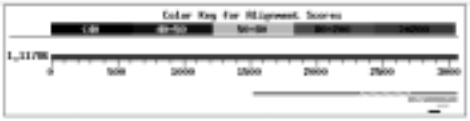
Begin Search Clear Input

NCBI

Hits to At Genome

Distribution of 10 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments




```
                                                                  Score    E
Sequences producing significant alignments:                       (bits) Value

dbj|AB011477.1|AB011477  Arabidopsis thaliana genomic DNA, c...    134   2e-47
emb|AL096856.1|AP724C20  Arabidopsis thaliana DNA chromosome...     94   7e-33
gb|AC015445.3|F2J10  Sequence of BAC F2J10 from Arabidopsis ...     66   8e-09
emb|AL022373.1|AT719K4  Arabidopsis thaliana DNA chromosome ...     44   0.088

                                   Alignments

>dbj|AB011477.1|AB011477 Arabidopsis thaliana genomic DNA, chromosome 5, P1 clone:MHK7
        Length = 78181

 Score =  134 bits (336), Expect(2) = 2e-47
 Identities = 79/213 (37%), Positives = 119/213 (55%), Gaps = 5/213 (2%)
 Frame = -1

Query: 2705  VGSDGKERRQLVKGRDDLRQDAVMQQVFQMCNTLLQRNTETRKRKLTICTYKVVPLSQRS 2764
             +G+DG E   L K +DDLR+DA M +   M N LL +  E+R+RKL I T+ V PL++
Sbjct: 19450 LGNDGIEYPFLCKPFDDLRKDAMMEFTRMINRLLEKYPESRRRKLYIRTFAVAPLTEDC 19271
```

NCBI

Genomic Context of BLAST Hits

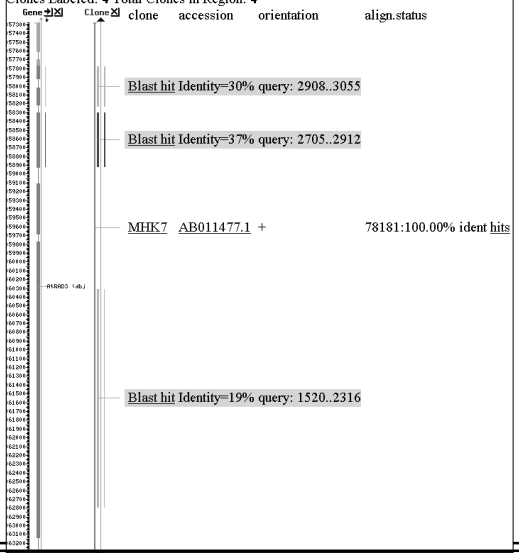
Master Map: Clone **Maps & Options**

Total Clones On Chromosome: **416**

Region Displayed: **16,057K-16,063K**

bp Download/View Sequence/Evidence

Clones Labeled: **4** Total Clones in Region: **4**



NCBI

The Rice Genome

Oryza sativa

BLAST Home Page

BLAST overview

BLAST FAQs

BLAST news

BLAST manual

BLAST against *Oryza sativa* ssp. *indica* WGS contigs

In that none of the contigs have been mapped to a rice genome, we are unable to provide a display of the location of any BLAST hits in the rice genome.

Database: contig Program tblastn MegaBlast: Begin Search

Enter an accession, gi, or a sequence in FASTA format:

```
>gi|13878337|sp|Q13315|ATM_HUMAN Serine-protein kinase ATM (
MSLVLNDLLICCRQLEHDRATERKKEVEKFKRLIRDPETIKHLDRHSDSKQGKYLNWDAV
ETECLRIAKPNVSASTQASRQKKMQEISSLVKYFIKCANRRAPRLKCQELLNYIMDTVKD
CSNILLKDILSVRKYWCEISQQQWLELFSVYFRLYLKPSQDVHRVLVARIIHAVTKGCCS
DFFSKAIQCARQEKSSSGLNHILAALTIFLKTLAVNFRIRVCELGDEILPTLLYIWTQHR
ELFQLQIYIHHPKGAKTQEKGAYESTKWRSILYNLYDLLVNEISHIGSRGKYSSGFRNIA
```

Optional parameters

Expect	Filter	Descriptions	Alignments
0.01	default	100	100

Advanced options:

Begin Search Clear Input

NCBI

Shotgun Contigs

RID: 1019619759-0

Query= gi|1387833
ATM (Ataxia telan
(3056 le

Database: Oryza s
103,04

If you have any p
please refer to t

Taxonomy reports

Mouse-over to sh

1_13834
0

Sequences produci

```
LOCUS       AAAA01003283           20559 bp    DNA     linear   PLN 04-APR-2002
DEFINITION  Oryza sativa (indica cultivar-group), whole genome shotgun
            sequence.
ACCESSION   AAAA01003283
VERSION     AAAA01003283.1  GI:19927592
KEYWORDS    .
SOURCE      Oryza sativa (indica cultivar-group).
  ORGANISM  Oryza sativa (indica cultivar-group)
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae;
            Ehrhartoideae; Oryzeae; Oryza.
REFERENCE   1  (bases 1 to 20559)
  AUTHORS   Yu,J., Hu,S., Wang,J., Wong,G.K.-S., Li,S., Liu,B., Deng,Y.,
            Dai,L., Zhou,Y., Zhang,X., Cao,M., Liu,J., Sun,J., Tang,J.,
            Chen,Y., Huang,X., Lin,W., Ye,C., Tong,W., Cong,L., Geng,J.,
            Han,Y., Li,L., Li,W., Hu,G., Huang,X., Li,W., Li,J., Liu,Z., Li,L.,
            Liu,J., Qi,Q., Liu,J., Li,L., Li,T., Wang,X., Lu,H., Wu,T., Zhu,M.,
            Ni,P., Han,H., Dong,W., Ren,X., Feng,X., Cui,P., Li,X., Wang,H.,
            Xu,X., Zhai,W., Xu,Z., Zhang,J., He,S., Zhang,J., Xu,J., Zhang,K.,
            Zheng,X., Dong,J., Zeng,W., Tao,L., Ye,J., Tan,J., Ren,X., Chen,X.,
            He,J., Liu,D., Tian,W., Tian,C., Xia,H., Bao,Q., Li,G., Gao,H.,
            Cao,T., Wang,J., Zhao,W., Li,P., Chen,W., Wang,X., Zhang,Y., Hu,J.,
            Wang,J., Liu,S., Yang,J., Zhang,G., Xiong,Y., Li,Z., Mao,L.,
            Zhou,C., Zhu,Z., Chen,R., Hao,B., Zheng,W., Chen,S., Guo,W., Li,G.,
            Liu,S., Tao,M., Wang,J., Zhu,L., Yuan,L. and Yang,H.
  TITLE     Draft Sequence of the Rice Genome (Oryza sativa L. ssp. indica)
  JOURNAL   Science 296, 79-92 (2002)
```

gb|AAAA01003283.1|	Oryza sativa (indica cultivar-group) sca...	139	4e-31
gb|AAAA01002561.1|	Oryza sativa (indica cultivar-group) sca...	124	1e-26
gb|AAAA01009014.1|	Oryza sativa (indica cultivar-group) sca...	82	4e-21
gb|AAAA01070788.1|	Oryza sativa (indica cultivar-group) sca...	102	4e-20
gb|AAAA01001778.1|	Oryza sativa (indica cultivar-group) sca...	53	5e-05

NCBI

Service Addresses

•*General Help*	`info@ncbi.nlm.nih.gov`
•*Questions about BLAST*	`blast-help@ncbi.nlm.nih.gov`

E-mail Servers

BLAST Server	`blast@ncbi.nlm.nih.gov`

NCBI