NCBI Molecular Biology Resources

A Field Guide

part 2

Feb. 25, 2003

Web Access

Text

Entrez

Sequence

BLAST

Structure

VAST
Some Web Statistics

July 2001

<table>
<thead>
<tr>
<th>Users Per Weekday</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

- Nucleotide
- Genome
- BLAST
- OMIM
- Protein
- UniGene
- Structure
- Taxonomy
- LocusLink
- GeneMap
- Books
- Genes and disease

BLAST alone: currently 80,000 searches per day

PubMed: up to 4,000,000 searches per day

Users per day

<table>
<thead>
<tr>
<th>Year</th>
<th>1997</th>
<th>1998</th>
<th>1999</th>
<th>2000</th>
<th>2001</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>250000</td>
<td>220000</td>
<td>200000</td>
<td>250000</td>
<td>300000</td>
</tr>
</tbody>
</table>

Christmas Day
The NCBI ftp site

30,000 files per day
620 Gigabytes per day

Major resources available by ftp (ftp.ncbi.nih.gov):
- BLAST Basic Local Alignment Search Tool
- GenBank
- RefSeq
- dbSNP
- Taxonomy
- UniGene
- UniSTS

Text Searching
Database Searching with Entrez

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

Entrez Nucleotides

Search for:

- glucose 6 phosphate dehydrogenase
glucose 6 phosphate dehydrogenase

1: NC_004113
   Thermosynechochoccus elongatus BPP-1, complete genome
   gi|22297544|ref|NC_004113.1|[22297544]

2: AE004956
   Pseudomonas aeruginosa PA01, section 517 of 529 of the complete genome
   gi|9951560|gb|AE004956.1|[9951560]

3: NC_001146
   Saccharomyces cerevisiae chromosome XIV, complete chromosome sequence
   gi|63239899|ref|NC_001146.1|[63239899]

4: AJ439407
   Rhodovivax opacus ORF1 (partial), ORF2, ORF3, ORF4, cleA2 gene, cleB2 gene, cleC gene, ORF9 (partial)
   gi|23094401|ref|AJ439407.1|[23094401]

5: AY117271
   Arabidopsis thaliana putative glucose-6-phosphate dehydrogenase (At1g59420) mRNA, complete cds
   gi|21436352|gb|AY117271.1|[21436352]

6: AY035632
   Arabidopsis thaliana putative glucose-6-phosphate dehydrogenase
glucose 6 phosphate dehydrogenase
Document Summaries: Limits

<table>
<thead>
<tr>
<th>Title</th>
<th>Summary</th>
<th>Gene</th>
<th>Species</th>
<th>Accession</th>
<th>Gene Bank</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana putative glucose-6-phosphate dehydrogenase (At1g09420) mRNA, complete cds</td>
<td>gi</td>
<td>214362352</td>
<td>gb</td>
<td>AY117271.1</td>
<td>gi</td>
<td>214362352</td>
</tr>
<tr>
<td>Arabidopsis thaliana putative glucose-6-phosphate dehydrogenase (At1g09420) mRNA, complete cds</td>
<td>gi</td>
<td>5810386</td>
<td>gb</td>
<td>AY016232.1</td>
<td>gi</td>
<td>15810386</td>
</tr>
<tr>
<td>Ovis aries partial mRNA for glucose-6-phosphate dehydrogenase (g6pd gene)</td>
<td>gi</td>
<td>22797197</td>
<td>emb</td>
<td>J907200.1</td>
<td>gi</td>
<td>AR597200</td>
</tr>
<tr>
<td>Rattus norvegicus Glucose-6-phosphate dehydrogenase (G6PD), mRNA</td>
<td>gi</td>
<td>E393380</td>
<td>pt</td>
<td>NM_017006.1</td>
<td>gi</td>
<td>E393380</td>
</tr>
<tr>
<td>Homo sapiens glucose-6-phosphate dehydrogenase (G6PD), nuclear gene encoding mitochondrial protein, mRNA</td>
<td>gi</td>
<td>21614519</td>
<td>ref</td>
<td>NM_000402.2</td>
<td>gi</td>
<td>21614519</td>
</tr>
<tr>
<td>AB005946</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Adding Terms: Preview/Index

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

Click AND OR NOT to add terms selected from Index to the query box.
**FASTA Format**

**FASTA Definition Line**

```plaintext
>gi|603218|gb|U18238.1|MSU18238	Medicago sativa glucose-6-phosphate dehydrogenase
cCACCAGATATAATTAAGTAGATCAGAGTAGAAGAAGATGGGAACAAATGAATGGCATGTAGAAAGAAGA
GATAGCATAGGTACTGAATCTCCTGTAGCAAGAGAGGTACTTGAAACTGGCACACTCTCTATTGTTGTGC
TTGGTGCTTCTGGTGATCTTGCCAAGAAGAAGACTTTTCCTGCACTTTTTCACTTATATAAACAGGAATT
GTTGCCACCTGATGAAGTTCACATTTTTGGCTATGCAAGGTCAAAGATCTCCGATGATGAATTGAGAAAC
AAATTGCGTAGCTATCTTGTTCCAGAGAAAGGTGCTTCTCCTAAACAGTTAGATGATGTATCAAAGTTTT
TACAATTGGTTAAATATGTAAGTGGCCCTTATGATTCTGAAGATGGATTTCGCTTGTTGGATAAAGAGAT
TTCAGAGCATGAATATTTGAAAAATAGTAAAGAGGGTTCATCTCGGAGGCTTTTCTATCTTGCACTTCCT
CCTTCAGTGTATCCATCCGTTTGCAAGATGATCAAAACTTGTTGCATGAATAAATCTGATCTTGGTGGAT
GGACACGCGTTGTTGTTGAGAAACCCTTTGGTAGGGATCTAGAATCTGCAGAAGAACTCAGTACTCAGAT
TGGAGAGTTATTTGAAGAACCACAGATTTATCGTATTGATCACTATTTAGGAAAGGAACTAGTGCAAAAC
ATGTTAGTACTTCGTTTTGCAAATCGGTTCTTCTTGCCTCTGTGGAACCACAACCACATTGACAATGTGC
AGATAGTATTTAGAGAGGATTTTGGAACTGATGGTCGTGGTGGATATTTTGACCAATATGGAATTATCCG
AGATATCATTCCAAACCATCTGTTGCAGGTTCTTTGCTTGATTGCTATGGAAAAACCCGTTTCTCTCAAG
CCTGAGCACATTCGAGATGAGAAAGTGAAGGTTCTTGAATCAGTACTCCCTATTAGAGATGATGAAGTTG
TTCTTGGACAATATGAAGGCTATACAGATGACCCAACTGTACCGGACGATTCAAACACCCCGACTTTTGC
AACTACTATTCTGCGGATACACAATGAAAGATGGGAAGGTGTTCCTTTCATTGTGAAAGCAGGGAAGGCC
CTAAATTCTAGGAAGGCAGAGATTCGGGTTCAATTCAAGGATGTTCCTGGTGACATTTTCAGGAGTAAAA
AGCAAGGGAGAAACGAGTTTGTTATCCGCCTACAACCTTCAGAAGCTATTTACATGAAGCTTACGGTCAA
GCAACCTGGACTGGAAATGTCTGCAGTTCAAAGTGAACTAGACTTGTCATATGGGCAACGATATCAAGGG
ATAACCATTCCAGAGGCTTATGAGCGTCTAATTCTCGACACAATTAGAGGTGATCAACAACATTTTGTTC
GCAGAGACGAATTAAAGGCATCATGGCAAATATTCACACCACTTTTACACAAAATTGATAGAGGGGAGTT
GAAGCCGGTTCCTTACAACCCGGGAAGTAGAGGTCCTGCAGAAGCAGATGAGTTATTAGAAAAAGCTGGAT
ATGTTCAAACACCCGGTTATATATGGATTCCTCCTACCTTATAGAGTGACCAAATTTCATAATAAAACA
AGGATTAGGATTATCAGGAGCTTATAAATAAGTCTTCAATAAGCTTGTGAAATTTTCGTTATAATCTCTC
TCATTTTGGGGTGTTAATGATATCATTCCAAACCATCTGTTGCAGGTTCTTTGCTTGATTGCTATGGAAAAAC
CCGTTTCTCTCAAGCCTGAGCACATTCGAGATGAGAAAGTGAAGGTTCTTGAATCAGTACTCCCTATTAGAG
ATGATGAAGTTGTTCTTGGACAATATGAAGGCTATACAGATGACCCAACTGTACCGGACGATTCAAACACCC
CGACTTTTGC
```

**Description**

This FASTA sequence represents the glucose-6-phosphate dehydrogenase from *Medicago sativa* (alfalfa). The sequence is accessioned as gi|603218|gb|U18238.1 in the GenBank database with the identifier MSU18238.

**Features**

- **Source**: Medicago sativa subsp. sativa
- **Organism**: *Medicago sativa* (alfalfa)
- **Reference**: Transcriptional activation of oxidative pentose phosphate pathway genes at the onset of the first loaf mold phytoplasma response
- **Authors**: Fahrendorf, R. M., Borror, S. A., and Dimm, B. A.
- **Title**: Plant Mol. Biol. 28, 51, 481-900 (1995)
- **Journal**: Submitted (06-DEC-1994) The Fahrendorf, Samuel Roberts Noble Foundation, Plant Biology Division, 2330 Sam Noble Parkway, OK 73002, USA

**Translation**

The sequence can be translated into amino acids using standard translation tables.

**Accession Number**

- **gi**: gi|603218
- **gb**: gi|603218
- **U18238**: U18238.1
- **MSU18238**: MSU18238

**Database Identifiers**

- GenBank
- EMBL
- SWISS-PROT
- dbJ
- Protein Databank
- PIR
- PRF
- RefSeq

This sequence is part of a larger dataset, and its role in the pentose phosphate pathway is highlighted through its transcriptional activation in response to certain conditions.

---

**Entrez GenBank**

- **gi|603218|gb|U18238.1|MSU18238**
- **Medicago sativa glucose-6-phosphate dehydrogenase**
- **Locus**: MSU18238
- **Description**: *Medicago sativa* glucose-6-phosphate dehydrogenase
- **Accession**: gi|603218
- **GB**: gi|603218
- **U18238**: U18238.1
- **MSU18238**: MSU18238
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 {
          name binomial {
            genus "Medicago" ,
            species "sativa" ,
            subspecies "subsp. sativa" } ,
          mod {
          }}}}}
Protein Neighbors and Links

- **Redundant Sequences**

Neighbors: Related Proteins

- **15**: AAB41552, glucone-6-phosphohydrolase [Medicago sativa subsp. sativa] g903219p[0603219]
  - **Definition**: glucone-6-phosphohydrolase.
  - **Accession**: AAB41552
  - **Version**: AAB41552.1
  - **Source**: Medicago sativa subsp. sativa
  - **Organism**: Medicago sativa subsp. sativa

- **21**: E5Y735
  - **Glucone-6-phosphohydrolase (EC 1.1.1.49) - alfalfa**

- **33**: O42819
  - **Glucone-6-phosphohydrolase, cytoplasmic isoform (G6PD)**

- **41**: BABO8897
  - **Glucone-6-phosphohydrolase, Arabidopsis thaliana**

- **52**: NP_198892
  - **Glucone-6-phosphohydrolase, Arabidopsis thaliana**

- **61**: CAB52675
  - **Glucone-6-phosphohydrolase, Arabidopsis thaliana**

- **71**: BABO2124
  - **Glucone-6-phosphohydrolase, Arabidopsis thaliana**

- **84**: NP_183966
  - **Glucone-6-phosphohydrolase, Arabidopsis thaliana**

**Links**

- Related Sequences
- Domain Relatives
- Domains
- Nucleotide
- PubMed
- Taxonomy
- LinkOut
- Medicago sativa subsp. sativa
Protein Neighbors->PopSet Links

The population genetics of the origin and divergence of the Drosophila simulans complex species. (Kliiman R.M. et al., direct submission)

Protein Neighbors->Genome Links

1: NC_004113
Thermosynechococcus elongatus BP-1, complete genome[251]

1: NC_003076
Arabidopsis thaliana chromosome 5, complete sequence[197]

gi:22328163[refNC_003076.3]

2: NC_003074
Arabidopsis thaliana chromosome 3, complete sequence[195]

gi:223331925[refNC_003074.3]

3: NC_003070
Arabidopsis thaliana chromosome 1, complete sequence[193]

gi:22330788[refNC_003070.3]

gi:2281725[refNC_003070.3]

7: NC_003919
Xanthomonas acomapodis pv. citri str. 306, complete genome[242]

gi:2424977[refNC_003919.2]
### BLink: Advanced Protein Neighbors

#### Query: gi|5335418 | glucose-6-phosphate 1-dehydrogenase, cytoplasmic isoform (G6PD)

Matching at COGm, DMPA

**COGm** assigned by Cogitor (57 best hits)

<table>
<thead>
<tr>
<th>Score</th>
<th>Accession</th>
<th>GI</th>
<th>N</th>
<th>Organism</th>
</tr>
</thead>
<tbody>
<tr>
<td>2753</td>
<td>RA058537</td>
<td>7757070</td>
<td>10</td>
<td>Arthrobacter thelii</td>
</tr>
<tr>
<td>2708</td>
<td>CAAB2349</td>
<td>471316</td>
<td>3</td>
<td>Salmontuberculosis</td>
</tr>
<tr>
<td>2765</td>
<td>CAAB0696</td>
<td>941256</td>
<td>2</td>
<td>Nicotianatabacum</td>
</tr>
<tr>
<td>2715</td>
<td>AAEL1426</td>
<td>40561144</td>
<td>3</td>
<td>Moraykarmeychamcocrystallum</td>
</tr>
<tr>
<td>2714</td>
<td>AAEN5319</td>
<td>2335933</td>
<td>3</td>
<td>Petrovilum oxican</td>
</tr>
<tr>
<td>2703</td>
<td>CAAB7160</td>
<td>8218502</td>
<td>3</td>
<td>Criticium sicilia</td>
</tr>
<tr>
<td>2708</td>
<td>CAM71374</td>
<td>2741463</td>
<td>3</td>
<td>Oryza sativa</td>
</tr>
<tr>
<td>2150</td>
<td>AAAB6236</td>
<td>8206374</td>
<td>3</td>
<td>Chlorovolviselena</td>
</tr>
<tr>
<td>2150</td>
<td>CAM8047</td>
<td>519709</td>
<td>3</td>
<td>Rattus norvegicus</td>
</tr>
<tr>
<td>2150</td>
<td>CAAB7797</td>
<td>519709</td>
<td>3</td>
<td>Mus musculus</td>
</tr>
<tr>
<td>2150</td>
<td>ACC02014</td>
<td>4882743</td>
<td>1</td>
<td>Crassostreculis graceus</td>
</tr>
<tr>
<td>2150</td>
<td>CAAB7736</td>
<td>2724836</td>
<td>2</td>
<td>Testa frugiperpensa</td>
</tr>
<tr>
<td>2150</td>
<td>AAAB7629</td>
<td>5405109</td>
<td>1</td>
<td>Mesorobus robustus</td>
</tr>
<tr>
<td>2150</td>
<td>RAAB3134</td>
<td>2724836</td>
<td>2</td>
<td>Testa frugiperpensa</td>
</tr>
<tr>
<td>2150</td>
<td>CAAB8039</td>
<td>452269</td>
<td>10</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>2150</td>
<td>AAAB7294</td>
<td>1959734</td>
<td>1</td>
<td>Sympytomma drosophila</td>
</tr>
<tr>
<td>2150</td>
<td>CAAB8039</td>
<td>452269</td>
<td>10</td>
<td>Homo sapiens</td>
</tr>
<tr>
<td>2150</td>
<td>AAAB7294</td>
<td>1959734</td>
<td>1</td>
<td>Sympytomma drosophila</td>
</tr>
<tr>
<td>2150</td>
<td>CAAB8039</td>
<td>452269</td>
<td>10</td>
<td>Homo sapiens</td>
</tr>
</tbody>
</table>

---

### BLink: CDD Search

#### NCBI Conserved Domain Summary

**RPS-BLAST 2.2.9 (Nov-16-2002)**

**Query:** gi|5335418 | glucose-6-phosphate 1-dehydrogenase

**([1513 letters])**

**Database:** cdd.v1.50

**10,013 PSSMs; 2,444,703 total columns**

| gi|CCDI0216.COG0343 | gi|CCDI0216.COG0344 | Zef | Glucose-6-phosphate dehydrogenase | S | -5296.6e-151 |

**COGs now in CDD**

**PSI-BLAST producing significant alignments**

| gi|CCDI0216.COG0343 | gi|CCDI0216.COG0344 | Zef | Glucose-6-phosphate 1-dehydrogenase | S | -5296.6e-151 |

---

**Domain Check**

- This CD alignment includes 3D structure. To display structure, download **Cn3D**

**Score E**

- (Homology value)
Human RefSeqs: Genome Reagents

1: NT_020985
   Homo sapiens chromosome X genomic contig
   g27485381[ref]NT_020985.9[ExX_2612]|g27485381

2: NM_000462
   Homo sapiens glucose-6-phosphate dehydrogenase (G6PD), nuclear
   mitochondrial protein, mRNA
   g21614519[ref]NM_000462.2|g21614519

LocusLink: Selected Higher Genomes

<table>
<thead>
<tr>
<th>LocusID</th>
<th>Org Symbol</th>
<th>Description</th>
<th>Section</th>
<th>Links</th>
</tr>
</thead>
<tbody>
<tr>
<td>2539</td>
<td>Glucose</td>
<td>glucose-6-phosphate dehydrogenase</td>
<td>Xq28</td>
<td></td>
</tr>
<tr>
<td>243</td>
<td>G6PD</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>28117</td>
<td>G6PD</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>11103</td>
<td></td>
<td></td>
<td>4q27</td>
<td></td>
</tr>
<tr>
<td>14380</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>24370</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>2540</td>
<td></td>
<td></td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>30720</td>
<td></td>
<td></td>
<td>U</td>
<td></td>
</tr>
<tr>
<td>14381</td>
<td></td>
<td></td>
<td>X:3012</td>
<td></td>
</tr>
<tr>
<td>32974</td>
<td></td>
<td></td>
<td>18D13:18F1</td>
<td></td>
</tr>
</tbody>
</table>

Links:
- GenBank
- RefSeq
- OMIM
- UniGene
- PubMed
- Protein
- SNP
- Taxonomy
- LinkOut
Microbial Genomes

Genes Database: All Genomes

Search Genes: Go: Clear

G6PD glucose-6-phosphate dehydrogenase [Homo sapiens] Gene Link

Transcripts and products:

Genomic context: chromosome: X; Maps: Xq28

Gene type: protein coding
RefSeq status: Reviewed
Organism: Homo sapiens
Lineage: Eukaryota; Metazoa; Chordata; Osteichthyes; Vertebrata; Tetrapoda; Mammalia; Primates; Hominidae; Homo;
Gene aliases: G6PD1
Summary: This gene encodes glucose-6-phosphate dehydrogenase. This protein is a cytosolic enzyme encoded by a housekeeping X-linked gene whose main function is to produce NADPH, a key electron donor in the defense against oxidizing agents and in reductive biosynthetic reactions. G6PD is remarkable for its genetic diversity. Many variants of G6PD, mostly produced from nonsense mutations, have been described with wide-ranging levels of enzyme activity and associated clinical symptoms. G6PD deficiency may cause neonatal jaundice, acute hemolysis, or severe chronic non-spherocytic hemolytic anemia.

General protein information:
Name: glucose-6-phosphate dehydrogenase

Coming soon!
Sequence Similarity Searching

**Basic Local Alignment Search Tool**

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
Molecular Evolution

Common ancestry allows us to infer similar function

3000 Myr

1000 Myr

540 Myr

Human

Fly

Worm

Yeast

Bacteria

Pancreatic carcinoma

Alzheimer's Disease

Ataxia telangiectasia

Colon cancer

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, standalone, and network clients
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)

Look Up Table (Hash Table)

Query: GTQITVEDLFYNIATRRKALKN

| Word Size | 3
| Adj | 2 or 3 for protein (3 default)
| > 7 for blastn searches (11 default) |

Neighborhood Words:
- LTV, MTV, ISV, LSV, MSV
- IAV, LAV, MAV, ITL, etc.

Make table for both query and database
An alignment that BLAST can’t find

```
1  GAATATATGAAGACCAAGATTGCAGTCCTGCTGGCCTGAACCACGCTATTCTTGCTGTTG
   || | || || || |  || || ||   ||  |  ||| |||||| | | || | ||| |
1  GAGGTGTACGATGAGCCCGAGTGTAIGCAGTGACATCTGGATTAAGAACTCTCTGGGGAGCCAGTT
   || | || || || |  || || ||   ||  |  ||| |||||| | | || | ||| |
61  GTTACGGAACCGAGAATGGTAAAGACTACTGGATCATTAAGAACTCCTGGGGAGCCAGTT
   || ||     ||  ||| ||  | |||||| || | ||||||  |||||  |    |
61  GCTATGGTGTTAAGGGTGAGAAAGTACTGGCTCGTCAAGAACAGCTGGGCTGAATCCT
```

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

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

or

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

$K =$ scale for search space

$\lambda =$ scale for scoring system

$S' =$ bitscore $= (\lambda S - \ln K)/\ln 2$

Scoring Systems

- Position Independent Matrices
  - Nucleic Acids – identity matrix
  - Proteins
    - 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

### BLOSUM62

**Common amino acids** have low weights

**Rare amino acids** have high weights

**Negative for less likely substitutions**

**Positive for more likely substitutions**
Gapped Alignments

- Gapping provides more biologically realistic alignments
- Gapped BLAST parameters must be simulated for each matrix
- Affine gap costs = -\((a+bk)\)
  \(a = \) gap open penalty \(b = \) gap extend penalty
  A gap of length 1 receives the score \(-(a+b)\)

Scores

<table>
<thead>
<tr>
<th></th>
<th>V</th>
<th>D</th>
<th>S</th>
<th>C</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOSUM62</td>
<td>+4</td>
<td>+2</td>
<td>+1</td>
<td>-12</td>
<td>+9</td>
</tr>
<tr>
<td>PAM30</td>
<td>+7</td>
<td>+2</td>
<td>0</td>
<td>-10</td>
<td>+10</td>
</tr>
</tbody>
</table>
WWW BLAST

Web BLAST

What's NEW in BLAST®

News March 5th 2002: New database links from BLAST results: Results of a BLAST search will now link sequences from the BLAST results page to the NCB LocusLink and UniGene databases. Links to additional databases soon.

Nucleotide BLAST

- Standard nucleotide-nucleotide BLAST (nblast)
- NCBILAST
- Search for short nearly exact matches

Protein BLAST

- Standard protein-protein BLAST (pblast)
- PSI-BLAST
- Search for short nearly exact matches

Translated BLAST Searches

- Nucleotide query - Protein db [Blast]
- Protein query - Translated db [Blast]
- Nucleotide query - Translated db [Blast]

Search for conserved domains

- Search the Conserved Domain Database using RPS-BLAST
- Search by domain architecture (CDART)

Pairwise BLAST

- BLAST 2 Sequences

Genomic BLAST pages

- Human
- Mouse
- Rat
- Fugu rubripes
- Zebrafish
- Flies
- Nematodes
- Plants
- Yeasts
- Other eukaryotes
- Microbial
BLAST Databases: Non-redundant protein

**nr** (non-redundant protein sequences)
- GenBank CDS translations
- NP_RefSeqs
- Outside Protein
  - PIR, Swiss-Prot, PRF
- **PDB** (sequences from structures)

BLAST Databases: Nucleic Acid

- **nr (nt)**
  - Traditional GenBank Divisions
  - NM_ and XM_RefSeqs
- **dbest**
  - EST Division
- **htgs**
  - HTG division
- **gss**
  - GSS division
- **chromosome**
  - NC_RefSeqs
Mutated in Colon Cancer

IETVIAAYLPKNTHPFLYLSLEISPQNVDVNVHPTXHEVHFLHERSILER
VQQHIESKLLGSNSSRMYFTQTLLPGLAGPSGEVLKSTTSSLTSSSTSGSS

Putative conserved domains have been detected, click on the image below for detailed results.

The guest ID is: TO00000152-0025-1-0086

The results are estimated to be ready in 36 seconds but may be done sooner.
###税BLAST: 言語学評価

<table>
<thead>
<tr>
<th>植物名</th>
<th>種類</th>
<th>植物</th>
<th>亜種</th>
<th>DNA</th>
<th>匹配</th>
<th>repair</th>
<th>位置</th>
<th>倍率</th>
<th>期待値</th>
<th>一致率</th>
<th>正点</th>
<th>ギャップ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Escherichia coli</td>
<td>大腸菌</td>
<td>602</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>602</td>
<td>44.3</td>
<td>5e-05</td>
</tr>
<tr>
<td>Salmonella typhimurium</td>
<td>間葉菌</td>
<td>602</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>602</td>
<td>44.3</td>
<td>5e-05</td>
</tr>
<tr>
<td>Helicobacter pylori</td>
<td>ガングロ菌</td>
<td>782</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>782</td>
<td>44.3</td>
<td>5e-05</td>
</tr>
<tr>
<td>Chlamydia muridarum</td>
<td>チラミデア</td>
<td>835</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>835</td>
<td>44.3</td>
<td>5e-05</td>
</tr>
<tr>
<td>Chlamydia trachomatis</td>
<td>チラミデア</td>
<td>813</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>813</td>
<td>44.3</td>
<td>5e-05</td>
</tr>
<tr>
<td>Vibrio cholerae</td>
<td>チルネラ</td>
<td>666</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>666</td>
<td>44.3</td>
<td>5e-05</td>
</tr>
<tr>
<td>Campylobacter jejuni</td>
<td>キャンピロバクター</td>
<td>155892</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>155892</td>
<td>44.3</td>
<td>5e-05</td>
</tr>
</tbody>
</table>

###BLAST Output: Alignments

| Query: | 9 | LFKTHFPLYLLEISQPQVNDVNHPKRVVF---LHE---ESILERVQQHIESKL | 59 | L + P | LEI P | VDNHPK RVVF F | +H+ | +L | +Q +E+ L |
| Sbjct: | 280 | LQADQQPAPVLFYLDHPQVNDVNHPKRVVF QHIEKVRQVHQL LQQLETLPL | 338 | | | | | | | |
**BLAST Output: Alignments**

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Identity</th>
<th>Positive</th>
<th>Score</th>
<th>Expect</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>730028</td>
<td>sp</td>
<td>P40692</td>
<td>MLH1_HUMAN</td>
<td>DNA mismatch repair protein Mhl1 1)</td>
</tr>
</tbody>
</table>

**Results from nr**

<table>
<thead>
<tr>
<th>Accession</th>
<th>Description</th>
<th>Identity</th>
<th>Positive</th>
<th>Score</th>
<th>Expect</th>
</tr>
</thead>
<tbody>
<tr>
<td>gi</td>
<td>604369</td>
<td>gb</td>
<td>AAA85687.1</td>
<td>(U17857) hMLH1 gene product [Homo sapiens]</td>
<td>233 3e-61</td>
</tr>
<tr>
<td>gi</td>
<td>4557757</td>
<td>ref</td>
<td>NP_000240.1</td>
<td>(NM_000249) mutL homolog 1; mutL (E. coli) homolog 1 (colon cancer, nonpolyposis type 2) [Homo sapiens]</td>
<td>233 4e-61</td>
</tr>
<tr>
<td>gi</td>
<td>730028</td>
<td>sp</td>
<td>P40692</td>
<td>MLH1_HUMAN</td>
<td>DNA mismatch repair protein Mhl1 (MutL protein homolog 1)</td>
</tr>
<tr>
<td>gi</td>
<td>631298</td>
<td>gb</td>
<td>AAC50285.1</td>
<td>(U07343) hMLH1 [Homo sapiens]</td>
<td>233 4e-61</td>
</tr>
<tr>
<td>gi</td>
<td>1079797</td>
<td>gb</td>
<td>AAC20793.1</td>
<td>(U04078) DNA mismatch repair protein MHL1 - human</td>
<td>233 4e-61</td>
</tr>
<tr>
<td>gi</td>
<td>139012</td>
<td>gb</td>
<td>AAH04550.1</td>
<td>AAH04550 (BC006450) mutL (E. coli) homolog 1 type 2) [Homo sapiens]</td>
<td>233 4e-61</td>
</tr>
<tr>
<td>gi</td>
<td>741682</td>
<td>prf</td>
<td>2007430A</td>
<td>DNA mismatch repair protein [Homo sapiens]</td>
<td>233 4e-61</td>
</tr>
</tbody>
</table>

**Query:** 1  IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL 60  
**Sbjct:** 276 IETVYAAYLPKNTHPFLYLSLEISPQNVDVNVHPTKHEVHFLHEESILERVQQHIESKLL 335  
**Score:** 233 bits (593), Expect = 4e-61  
**Identities:** 117/131 (89%), Positive = 117/131 (89%)
**tblastn Results Against ESTs**

query: gi:12794555 | emb: AL531062.1 | AL531062

- **cDNA clone:** CSOM005VM23
- **prime:**
- **Length:** 878
- **Score:** 167 bits (422), Expect(3) = 1e-42
- **Identities:** 81/82 (98%), Positives = 81/82 (98%)
- **Frame:** +2

Query: 1  IETVYAAYLPKNTHPFLYLSLEISPQNVDVNHPTKHEVHFLHEESILERVQQHIESKLL 60

Sbjct: 512  IETVYAAYLPKNTHPFLYLSLEISPQNVDVNHPTKHEVHFLHEESILERVQQHIESKLL 691

Score: 24.3 bits (51), Expect(3) = 1e-42

- **Identities:** 11/26 (42%), Positives = 11/26 (42%)
- **Frame:** +1

Query: 80  PSGMVXXXXXXXXXXXXXDKVYA 105

Sbjct: 748  PSGMVXXXXXXXXXXXXXDKVYA 825

**PSI-BLAST**

Confirming relationships of purine nucleotide metabolism proteins
### CDD Curated Domain

<table>
<thead>
<tr>
<th>Feature</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consensus</strong></td>
<td>110 - cDTYRKLISWHTVttt - 126</td>
</tr>
<tr>
<td><strong>gi 734878</strong></td>
<td>731 - dSaLIEQzLI3T9h - 747</td>
</tr>
<tr>
<td><strong>gi 2499756</strong></td>
<td>1296 - cDLVAVFRVQHs - 1312</td>
</tr>
<tr>
<td><strong>gi 3192128</strong></td>
<td>2163 - eELIVQFELFh - 2179</td>
</tr>
<tr>
<td><strong>gi 17506579</strong></td>
<td>186 - nLEFAALLYMNhs - 202</td>
</tr>
<tr>
<td><strong>gi 17554484</strong></td>
<td>1029 - cDEAVQLVKs - 1046</td>
</tr>
</tbody>
</table>

### 3D Domain

**CDD Descriptive Terms**

- **Name**: PTPs

**Features**

- **Sequence Alignment Viewer**

| **gi 734878** | 731 - dSaLIEQzLI3T9h - 747 |
| **gi 2499756** | 1296 - cDLVAVFRVQHs - 1312 |
| **gi 3192128** | 2163 - eELIVQFELFh - 2179 |
| **gi 17506579** | 186 - nLEFAALLYMNhs - 202 |
| **gi 17554484** | 1029 - cDEAVQLVKs - 1046 |
Genomic BLAST pages

Megablast: NCBI’s Genome Annotator

- Long alignments between similar DNA sequences
- Concatenation of query sequences
- Faster than blastn
- Discontinuous Megablast allows cross-species comparisons
Microbial Genomes BLAST

BLAST with microbial genomes (209 bacterial 18 archaeal 43 eukaryotic genomes tree)

Currently available for BLAST searches are sequences from selected completed and unfinished eukaryotic and prokaryotic genomes. Partial genomic sequences have been gracefully provided by the sequencing centers or extracted from GenBank. See [About the Database]. NCBI encourages sequencing vendors to submit partially sequenced genomes to be included in this BLAST page. You can submit your data to blastp for which you will first need to contact us at genomes@ncbi.nlm.nih.gov to set up an account.

P - indicates the ability to search against protein sequences. - completed genomic sequence. - unfinished genomic sequence, EIEI - additional reference from selection. See [Help] and [Article] for details.

Hits to Unfinished Genome

Klebsiella pneumoniae, unfinished sequence

Accession: NC_003641
Number of contigs: 1482
Total Base Sequence: 4992116 bp
Last updated: Jul 7, 1999

Contributor: WashU-GSC

Organism: Klebsiella pneumoniae
Genetic Code: 11
Lineage: Bacteria, Proteobacteria, Gammaproteobacteria, Enterobacteriaceae, Klebsiella

COMMENT: DNA sequence and predicted proteins are available for BLAST at http://www.ncbi.nlm.nih.gov/gorf/gorf.html and are not downloadable from Entrez. Currently presented here are CGC's increased stringency criteria and are a last 200 bases long after vector and low quality sequence has been stripped out. Additional contigs not meeting these criteria can be obtained directly from CGC.

Policy on Early Data Release from WashU-GSC
Finding the Human Prestin Gene

**Prestin is the motor protein of cochlear outer hair cells**

Amy Zhou, Marking Shev, David V. L. Her, Keno B. Lang, Joshua D. Madison, & Peter Dobber

Institute for Molecular, Cellular, and Developmental Biology, Department of Biology, University of Chicago, Chicago, Illinois 60637

The outer and inner hair cells of the mammalian cochlea perform different functions. In response to changes in membrane potential, the cylindrical outer hair cell rapidly alters its length and stiffness. These mechanical changes, driven by putative molecular motors, are assumed to produce amplification of vibrations in the cochlea that are translated by inner hair cells. Here we have identified an abnormally expanded RNA transcribed from a gene, designated Prestin, which is specifically expressed in outer hair cells. Regions of the encoded protein show moderate sequence similarity to perrin and related subfamily/mouse transport proteins. Voltage-dependent shape changes can be elicited in cultured human inner hair cells that express prestin. The mechanical response of outer hair cells to voltage change is accompanied by a 'gating current', which is manifested as nonlinear capacitance. We also demonstrate that prestin is the motor protein of the cochlear outer hair cell.

Cochlear hair cells are non-epithelial cells that transduce acoustic signals. Outer hair cells (OHCs) are responsible for the requisite sensitivity and frequency resolving capacity of the normal mammalian hearing organ. They provide the mechanical amplification (the 'cochlear amplifier') in the form of feedback. In contrast, inner hair cells convey auditory information to the brain. Outer hair cells have cylindrical somata of constant diameter and variable length. When their membrane potential is altered, transient shape changes of up to 9% occur, the cell shortens when depolarized and lengthens when hyperpolarized. Length changes are not driven solely by Ca(2+); instead, they are accompanied by charge movements, which are reflected in nonlinear capacitance. The nonlinear capacitance is widely used as a 'signature' of the electromechanical process. Identity is also accomplished by a change in the axial stiffness of the cell. As virtually any test, connectivity and electrically-induced stiffness changes are correlated (DLZ-2, and SL- unpublished data) and we collectively describe them as

---

Searching the Human Genome

**NCBI**

**BLAST**

Overview

FAQs

news

manual references

**BLAST the Human genome**

Compare your query sequence to the working draft sequence of the human genome or its mRNA and protein products.

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

```plaintext
>gi:12188917 | emb | AJ303372.1 | RNO303372 Rattus norvegicus
ATGGATCATGCTGAAGAAAATGAAATTCCTGCAGAGATCAGAAGTACCTCGTGGAA
GTCATCCGGTCCTCCAGGAGAGGCTGCACGTCAAGGACAAAGTCACAGACTCCATC
GCAGGCATTCACGTGCACTCCTAAAAAAGTAAGAAACATCATCTACATGTTCTTGC
TTGCCAGCATATAAATTCAAGGAGTATGTGCTGGGTGACTTGGTCTCGGGCATAAG
AGCTCCCCCAAGGCTTAGCCTTCGCGATGCTGGCAGCTGTGCCTCCGGTGTTCGGC
```

---
Genomic Context of BLAST Hits

Genome Scan Models

GenBank

Mouse EST hits

Human EST hits
Genome Resources with MLH1

Higher Genomes Access

- Genomic BLAST
- Standard
- Map Viewer
- UniGene
- LocusLink
Map View MLH1 -> Variation Data

LocusLink via analysis of contig annotation: MLH1 mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)

Gene Model (contig mRNA transcript) information from genome sequence for NM_000249

<table>
<thead>
<tr>
<th>Contig accession</th>
<th>Contig position</th>
<th>Protein position</th>
<th>Function</th>
<th>dbSNP allele</th>
<th>Protein Codon</th>
<th>Amino acid position</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT_005588</td>
<td>4022028</td>
<td>NP_000240</td>
<td>contig reference</td>
<td>A</td>
<td>Bc [I]</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>nonsynonomous change</td>
<td>G</td>
<td>Val [V]</td>
<td>1</td>
</tr>
</tbody>
</table>

Legend

rs1055095 V | LTC inherited | ☆
rs1046512 V | LTC inherited | ☆
rs3172297 V | LTC inherited | ☆
rs1800734 V | LTC inherited | ☆
rs2020872 V | LTC inherited | ☆
rs3774343 V | LTC inherited | ☆
rs3856491 V | LTC inherited | ☆
rs2289038 V | LTC inherited | ☆
Would this be functionally significant?
The Mouse Assembly

Service Addresses

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