

Sequence File Formats

Major File Formats

- GenBank
- EMBL
- FASTA / Pearson
- Phylip
- GDE
- Nexus
- ASN.1
- BSML
- PDB
- SAM

GenBank

- Developed for a single nucleic sequence
- Does not display homology information
- Can display more than one gene, but only one primary sequence
- Large amount of details
 - organism
 - references
 - features

Position	Content
-----	-----
01-05	'LOCUS'
06-12	spaces
13-28	Locus name
29-29	space
30-40	Length of sequence, right-justified
41-41	space
42-43	bp
44-44	space
45-47	spaces, ss- (single-stranded), ds- (double-stranded), or ms- (mixed-stranded)
48-53	NA, DNA, RNA, tRNA (transfer RNA), rRNA (ribosomal RNA), mRNA (messenger RNA), uRNA (small nuclear RNA), snRNA, snoRNA. Left justified.
54-55	space
56-63	'linear' followed by two spaces, or 'circular'
64-64	space
65-67	The division code (see Section 3.3)
68-68	space
69-79	Date, in the form dd-MMM-yyyy (e.g., 15-MAR-1991)


```

LOCUS      MPU28721                2283 bp    DNA     linear   ROD 15-OCT-2001
DEFINITION Mus pahari adenine phosphoribosyltransferase (APRT) gene, complete
           cds.
ACCESSION  U28721
VERSION   U28721.1  GI:881573
KEYWORDS   .
SOURCE     Mus pahari (shrew mouse)
  ORGANISM Mus pahari
           Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
           Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
           Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE  1 (bases 1 to 2283)
  AUTHORS  Fieldhouse,D., Yazdani,F. and Golding,G.P.
  TITLE    Substitution rate variation in closely related rodent species
  JOURNAL  Heredity 78 (Pt 1), 21-31 (1997)
  PUBMED  9023989
REFERENCE  2 (bases 1 to 2283)
  AUTHORS  Fieldhouse,D.
  TITLE    Direct Submission
  JOURNAL  Submitted (07-JUN-1995) Dan Fieldhouse, Biology, McMaster
           University, 1280 Main Street West, Hamilton, ON, L8S 4K1, Canada
FEATURES   Location/Qualifiers
  source    1..2283
           /organism="Mus pahari"
           /mol_type="genomic DNA"
           /db_xref="taxon:10093"
  gene     <46..>2186
           /gene="APRT"
  mRNA    join(<46..125,256..362,1509..1642,1847..1925,
           2044..>2186)
           /gene="APRT"
  CDS     /product="adenine phosphoribosyltransferase"
           join(46..125,256..362,1509..1642,1847..1925,2044..2186)
           /gene="APRT"
           /EC_number="2.4.2.7"
           /note="purine salvage enzyme"
           /codon_start=1
           /product="adenine phosphoribosyltransferase"
           /translation="M...G...G...G..."

```

Keys

Gives the biological nature of the annotated feature

source
 promoter
 mRNA
 CDS

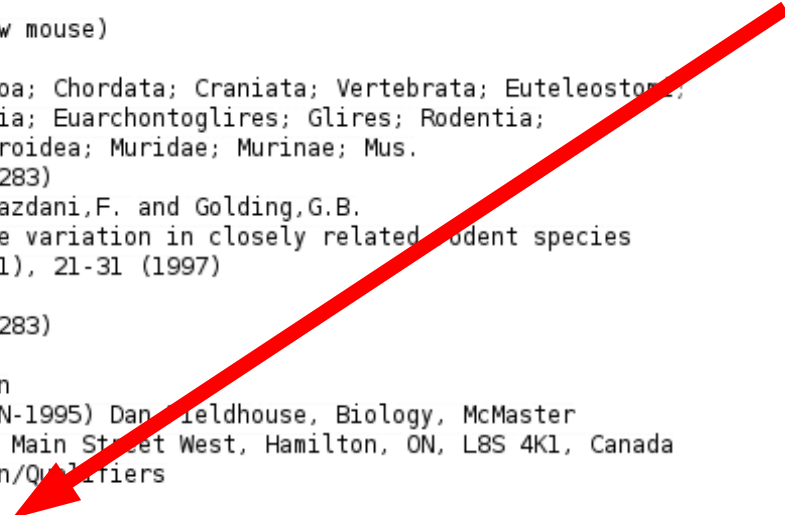

```

LOCUS      MPU28721                2283 bp    DNA     linear   ROD 15-OCT-2001
DEFINITION Mus pahari adenine phosphoribosyltransferase (APRT) gene, complete
           cds.
ACCESSION  U28721
VERSION   U28721.1  GI:881573
KEYWORDS   .
SOURCE    Mus pahari (shrew mouse)
  ORGANISM Mus pahari
           Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
           Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
           Sciurognathi; Muroidea; Muridae; Murinae; Mus.
REFERENCE  1 (bases 1 to 2283)
  AUTHORS  Fieldhouse,D., Yazdani,F. and Golding,G.B.
  TITLE    Substitution rate variation in closely related rodent species
  JOURNAL  Heredity 78 (Pt 1), 21-31 (1997)
  PUBMED   9023989
REFERENCE  2 (bases 1 to 2283)
  AUTHORS  Fieldhouse,D.
  TITLE    Direct Submission
  JOURNAL  Submitted (07-JUN-1995) Dan Fieldhouse, Biology, McMaster
           University, 1280 Main Street West, Hamilton, ON, L8S 4K1, Canada
FEATURES   Location/Qualifiers
  source   1..2283
           /organism="Mus pahari"
           /mol_type="genomic DNA"
           /db_xref="taxon:10093"
  gene     <46..>2186
           /gene="APRT"
  mRNA     join(<46..125,256..362,1509..1642,1847..1925,
           2044..>2186)
           /gene="APRT"
  CDS     join(46..125,256..362,1509..1642,1847..1925,2044..2186)
           /gene="APRT"
           /EC_number="2.4.2.7"
           /note="purine salvage enzyme"
           /codon_start=1
           /product="adenine phosphoribosyltransferase"
           /protein_id="M16087.1"

```

Qualifiers

Always in the form:
/[qualifer]=" "



```

/product="adenine phosphoribosyltransferase"
/protein_id="AAA68957.1"
/db_xref="GI:881574"
/translation="MSESELKLVARRIRSPDFPIPGVLFDRDISPLLKDPDSFRASIR
LLASHLKSTHSKIDYIAGLDSRGFLFGPSLAQELGVGCVLIRKQGLPGPTISASYA
LEYGKAELEIQKDALEPGQRVVI VDDLLATGGTMFAACDLLHQLRAEVVECVSLVELT
SLKGRERLGPPIFFSLLQYD"

```

ORIGIN

ORIGIN

```

1 cctgcgata ctcacctcct ccttgtctcc tacaagcacg cggccatgtc cgagtctgag
61 ttgaaactgg tggcgcggcg catccgcagc ttccccgact tccccatccc gggcgtgctg
121 ttcaggtgcg gtcacgagcc ggcgagggcgt tggcgccgta ctctcatccc ccggcgcagg
181 cgcgtgggca gccttgggga tcttgcgggg cctctgcccc gccacacgcg gtcactctcc
241 tgtccttgtt cccagggata tctcgcccc cttgaaagat ccggactcct tccgagcttc
301 catccgcctc ctggccagtc acctgaagtc cagcacagc ggcaagatcg actatatcgc
361 agggcaaggt ggccttgcta ggccgtactc atccccacg gtcctatccc ctatccccctt
421 tccccctggt tcaccacacag tctaccccac acccatccat tctttcttta acctctgact
481 cttcctcctt ggtttctcac tgccttggac gcttgttcac cccggatgaa ctccgtaggc
541 gtctcccttc cctgcttggg accctaaggt gccctcgggt cttgttcgta gagacgaaat
601 ctgctctgtc cttgtgtcca gaaccaagcc ttctctttt agggcacaa gctggccagc
661 atcctgacag caggctggga gacctggatc ctccagatga cggacatcct tacttagggg
721 tagcctctgg gatgaactag atattaaaaa ttaggtaac ttggggcgtg cctgggcaga
781 cctcaagtct ggtagcttca ggggctgctt cccccagga ctacaccggg gactcttct
841 cttgtcctcc ccacccccca caagcttgtg ctaaacaact gctgtatacc aggcctcatg
901 cttgagcttc agaaacaccc tagggcagct gaatgtccac caggagtgtc cagagggagg
961 gtgagcacc caagagaaca gagtggccct agtaaatgct caggaccac agaactttt
1021 cccactccac ttctatttg tacccccggc catgccccag aaatcagggc atgtttgtac
1081 cctccccacg acagctcggg ccgcttggaa ctgacctgta gacagtgtc ctgggtagat
1141 gctgcatttg aaagtgggca agagggtgg tgagatggct cagcggtag gagcactgac
1201 tgccttcca aaggtcctga gttcaaatcc cagcaaccac atggtggctc acaaccacct
1261 acagctacag tgtacacaca tataataaaa taaataaaca aatcttaaaa aaaaaaaaaa
1321 gaaagaaagg tggcaagagc caccatagtg gagaaggcag gtaggatccc caaggctaag
1381 atgctaccga gtaaccatca gtgttcttct agccatagtg ggcaagacct agtgttctta
1441 gtcaatgttg acctctccat acttgcctct cggctccatc ccacacctt cctccttac
1501 cctaacaggt ctgactcca ggggcttctt gtttggccct tccctagctc aggagctggg
1561 cgtgggctgc gtgctcatcc ggaagcaggg gaagctgccg ggccccacta tatcagctc
1621 ctatgctctg gtagtatggga aggtaaggga gctgtgggta gaggaagggc agggcttat
1681 taccacggct accagtgctt aggagtaaat gtgggtgctc agagaggttg agacattggg
1741 gtgaggttta caactctga aatgctcagc ctcaaaaatg ctccaggcta gggaggtggc
1801 cacttgttag catctagact ctcttaacgc tacttctgt ctgcaggctg agctggaaat
1861 ccagaaagat gccttagaac cggggcagag agtggctatt gtggatgacc tctggccac
1921 tggaggtaaa gaaccagccc aagacaaaaca ggcttcaaa ggcaggccc tgtctggggg
1981 gctgactaaa caaagcctt gaatacctt ctttctctg tccctcccc cccccccc
2041 caggaaacct gttgcagcc tgtgatctgc tgcaccagct acgggctgag gtgggtgag
2101 gtgtgagcct ggtggagctg acctcgtga agggcagga gaggctagga cctataccat
2161 tcttctctct cctccagat gactgagctg gctagatggt cacaccctg ctacacagcag
2221 cagtaactgc gcggtggctc agccctgggc gcctaagtga ccttgtgag ctacctgctg
2281 ccc

```

//

```
/product="adenine phosphoribosyltransferase"  
/protein_id="AAAG8957.1"  
/db_xref="GI:881574"  
/translation="MSESELKLVARRIRSFDFPIPGVLFDRDISPLLKDPDSFRASIR  
LLASHLKSTHSKIDYIAGLDSRGFLFGPSLAQELGVGCVLIRKQGLKPGPTISASYA  
LEYKAELEIQKDALEPGQRVVI VDDLLATGGTMFAACDLLHQLRAEVVECVSLVELT  
SLKGRERLGPPIFFSLLQYD"
```

ORIGIN

```
1 cctgcgata ctcacctcct ccttgtctcc tacaagcacg cggccatgct cgagtctgag  
61 ttgaaactgg tggcgcggcg catccgcagc ttccccgact tccccatccc gggcgtgctg  
121 ttcaggtgcg gtcacgagcc ggcgagggcgt tggcgcgta ctctcatccc ccggcgcagg  
181 cgcgtgggca gccttgggga tcttgcgggg cctctgcccc gccacacgcg gtcactctcc  
241 tgtccttgtt cccagggata tctcgccctt cttgaaagat ccggactcct tccgagcttc  
301 catccgcctc ctggccagtc acctgaagtc cacgcacagc ggcaagatcg actatatcgc  
361 agggcaaggt ggccttgcta ggccgtactc atccccacg gtcctatccc ctatccccct  
421 tccccctggt tcaccacacg tctaccccac acccatccat tctttcttta acctctgact  
481 cttcctcctt ggtttctcac tgccttggac gcttgttcac cccggatgaa ctccgtaggg  
541 gtctcccttc cctgcttggg accctaaggt gccctcgggt cttgttcgta gagacgact  
601 ctgctctgtc cttgtgtcca gaaccaagcc ttctctttt agggcacaaag gctgcaagc  
661 atcctgacag caggctggga gacctggatc ctccagatga cggacatcct tacctagggg  
721 tagcctctgg gatgaactag atattaaaaa ttaggtaac ttggggcgtg cctgggcaga  
781 cctcaagtct ggtagcttca gggcgtgctt cccccagga ctacaccgag gtcactctct  
841 cttgtcctcc ccacccccca caagcttgtg ctaaacaact gctgtatccc aggcctcatg  
901 cttgagcttc agaaaacccc tagggcagct gaatgtccac caggctgtgc cagagggagg  
961 gtgagcacc caagagaaca gagtggccct agtaaatgct caggaccac agaactttg  
1021 cccactccac ttctatttg taccceggc catgccccag aatcagggc atgtttgtac  
1081 cctccccacg acagctcggg ccgcctggaa ctgacctgtg gacagtgctc ctgggtagat  
1141 gctgcatttg aaagtgggca agagggtgg tgagatgct cagcggtag gagcactgac  
1201 tgccttcca aaggtcctga gttcaaatcc caggaccac atggtggctc acaaccacct  
1261 acagctacag tgtacacaca tataataaaa tataataaca aatcttaaaa aaaaaaaaaa  
1321 gaaagaaagg tggcaagagc cccatagtg agagaaggcag gtaggatccc caaggctaag  
1381 atgctaccga gtaaccatca gtgttctt agccatagtg ggcaagacct agtgttctta  
1441 gtcaatgttg acctctccat acttctctt cggctccatc ccacaccctt cctctcttac  
1501 cctaacaggt ctgactcca gggcttctt gtttggccct tccctagctc aggagctggg  
1561 cgtgggctgc gtgctcatcc caagcaggg gaagctgccg ggccccacta tatcagctc  
1621 ctatgctctg gagtatggg aggtaaggga gctgtgggta gaggaagggc agggcttat  
1681 taccacggct accagttctt aggagtaaat gtgggtgctc agagaggttg agacattggg  
1741 gtgaggttta caactctga aatgctcagc ctcaaaaatg ctccaggcta gggaggtggc  
1801 cacttgttag cacttagact ctcttaacgc tacttctctg ctgcaggctg agctggaat  
1861 ccagaaagat ccttagaac ccgggcagag agtggctatt gtggatgacc tctggccac  
1921 tggaggtat gaaccagccc aagacaaaca ggcttcaaa ggccaggccc tgtctggggt  
1981 gctgactaaa caaagcctt gaatacctt ctttctctg tccctcccc ccccccccc  
2041 cagcaccat gtttgagcc tgtgatctgc tgcaccagct acgggctgag gtgggtgagt  
2101 gctgagcct ggtggagctg acctcgtga agggcagga gaggctagga cctataccat  
2161 cttctctct cctccagat gactgagctg gctagatggt cacaccctg ctacacagcag  
2221 cagtaactgc gcggtggctc agccctgggc gcctaagtga cctttgtgag ctacctgctg  
2281 ccc
```

//

End of the entry, always //

```

/db_xref="taxon:7227"
/gene="CG11023"
/chromosome="2L"
7529..9491
/gene="CG11023"
/locus_tag="Dmel_CG11023"
/old_locus_tag="CG11023"
/note="CG11023"
/map="2LA5-2LA5"
/db_xref="FLYBASE:FBgn0031206"
/db_xref="GeneID:33155"
gene join(7529..8116,8229..8589,8668..9491)
/mRNA /gene="CG11023"
/locus_tag="Dmel_CG11023"
/old_locus_tag="CG11023"
/product="CG11023-RA"
/transcript_id="NM_175941.1"
/db_xref="GI:28573981"
/db_xref="FLYBASE:FBgn0031206"
/db_xref="GeneID:33155"
CDS join(7680..8116,8229..8589,8668..9276)
/gene="CG11023"
/locus_tag="Dmel_CG11023"
/old_locus_tag="CG11023"
/codon_start=1
/protein_id="NP_787955.1"
/db_xref="GI:28573982"
/db_xref="FLYBASE:FBgn0031206"
/gene="CG11023"
/db_xref="GeneID:33155"
gene complement(9836..21372)
/l(2)gl /gene="l(2)gl"
/locus_tag="Dmel_CG2671"
/old_locus_tag="CG2671"
/note="lethal (2) giant larvae; synonyms: CG2671, D-LGL,
L(2)GL, LGL, d[gl], gl, l(2), l(2) giant larva, l(2)giant
larvae, l-gl, l[[2]]gl, p127, p127l(2)gl"
/map="2LA5-2LA5"
/db_xref="FLYBASE:FBgn0002121"
/db_xref="GeneID:33156"
mRNA complement(join(9836..11344,11410..11518,11779..12221,
12286..12928,13520..13625,13683..14874,14933..15711,
19880..20020,21066..21200,21349..21372))
/gene="l(2)gl"
/locus_tag="Dmel_CG2671"
/old_locus_tag="CG2671"
/product="lethal (2) giant larvae CG2671-RF, transcript
variant F"

```

GenBank format can display more than one gene in a single sequence

EMBL Format

- Developed for a single nucleic sequence
- Can display more than one gene
- Large amount of detail
- Same as GenBank format but with 2 letters codes

ID U28721; SV 1; linear; genomic DNA; STD; ROD; 2283 BP.
 XX
 AC U28721;
 XX
 DT 04-JUL-1995 (Rel. 44, Created)
 DT 17-APR-2005 (Rel. 83, Last updated, Version 5)
 XX
 DE Mus pahari adenine phosphoribosyltransferase (APRT) gene, complete cds.
 XX
 KW .
 XX
 OS Mus pahari (shrew mouse)
 OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
 OC Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;
 OC Muridae; Murinae; Mus.
 XX
 RN [1]
 RP 1-2283
 RX DOI; [10.1038/sj.hdy.6881150](https://doi.org/10.1038/sj.hdy.6881150)
 RX PUBMED; [16397634](https://pubmed.ncbi.nlm.nih.gov/16397634/).
 RA Fieldhouse D., Yazdani F., Golding G.B.;
 RT "Substitution rate variation in closely related rodent species";
 RL Heredity 78(1):21-31(1997).
 XX
 RN [2]
 RP 1-2283
 RA Fieldhouse D.;
 RT ;
 RL Submitted (07-JUN-1995) to the EMBL/GenBank/DDBJ databases.
 RL Dan Fieldhouse, Biology, McMaster University, 1280 Main Street West,
 RL Hamilton, ON, L8S 4K1, Canada
 XX
 FH Key Location/Qualifiers
 FH
 FT [source](#) 1..2283
 FT /organism="Mus pahari"
 FT /mol_type="genomic DNA"
 FT /db_xref="taxon:10093"
 FT [mRNA](#) join(<46..125,256..362,1509..1642,1847..1925,2044..>2186)
 FT /gene="APRT"
 FT /product="adenine phosphoribosyltransferase"
 FT [CDS](#) join(46..125,256..362,1509..1642,1847..1925,2044..2186)

Line codes

ID - identification	RT - reference title
AC - accession number	RL - reference location
PR - project identifier	DR - database cross-reference
DT - date	CC - comments or notes
DE - description	AH - assembly header
KW - keyword	AS - assembly information
OS - organism species	FH - feature table header
OC - organism classification	FT - feature table data
OG - organelle	XX - spacer line
RN - reference number	SQ - sequence header
RC - reference comment	CO - contig/construct line
RP - reference positions	bb - (blanks) sequence data
RX - reference cross-reference	// - termination line
RG - reference group	
RA - reference author(s)	

ID U28721; SV 1; linear; genomic DNA; STD; ROD; 2283 BP.
 XX
 AC U28721;
 XX
 DT 04-JUL-1995 (Rel. 44, Created)
 DT 17-APR-2005 (Rel. 83, Last updated, Version 5)
 XX
 DE Mus pahari adenine phosphoribosyltransferase (APRT) gene, complete cds.
 XX
 KW .
 XX
 OS Mus pahari (shrew mouse)
 OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
 OC Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;
 OC Muridae; Murinae; Mus.
 XX
 RN [1]
 RP 1-2283
 RX DOI; [10.1038/sj.hdy.6881150](https://doi.org/10.1038/sj.hdy.6881150)
 RX PUBMED; [16397634](https://pubmed.ncbi.nlm.nih.gov/16397634/).
 RA Fieldhouse D., Yazdani F., Golding G.B.;
 RT "Substitution rate variation in closely related rodent species";
 RL Heredity 78(1):21-31(1997).
 XX
 RN [2]
 RP 1-2283
 RA Fieldhouse D.;
 RT ;
 RL Submitted (07-JUN-1995) to the EMBL/GenBank/DDBJ databases.
 RL Dan Fieldhouse, Biology, McMaster University, 1280 Main Street West,
 RL Hamilton, ON, L8S 4K1, Canada
 XX
 FH Key Location/Qualifiers
 FH
 FT [source](#) 1..2283
 FT /organism="Mus pahari"
 FT /mol_type="genomic DNA"
 FT /db_xref="taxon:10093"
 FT join(<46..125,256..362,1509..1642,1847..1925,2044..>2186)
 FT /gene="APRT"
 FT /product="adenine phosphoribosyltransferase"
 FT [mRNA](#)
 FT [cds](#) join(46..125,256..362,1509..1642,1847..1925,2044..2186)

DATA CLASS

How the data was generated

CON	fragments of entered sequences
ANN	fragment of entries sequences with its own annotation
PAT	Patent
EST	Expressed Sequence Tag
GSS	Genome Survey Sequence
HTC	High throughput cDNA
HTG	High throughput genomic
MGA	Mass Genome Annotation
WGS	Whole Genome Shotgun
TPA	Third Party Annotation
STS	Sequence Tag Site
STD	Standard

ID U28721; SV 1; linear; genomic DNA; STD; ROD; 2283 BP.
 XX
 AC U28721;
 XX
 DT 04-JUL-1995 (Rel. 44, Created)
 DT 17-APR-2005 (Rel. 83, Last updated, Version 5)
 XX
 DE Mus pahari adenine phosphoribosyltransferase (APRT) gene, complete cds.
 XX
 KW .
 XX
 OS Mus pahari (shrew mouse)
 OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
 OC Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;
 OC Muridae; Murinae; Mus.
 XX
 RN [1]
 RP 1-2283
 RX DOI; [10.1038/sj.hdy.6881150](https://doi.org/10.1038/sj.hdy.6881150)
 RX PUBMED; [16397634](https://pubmed.ncbi.nlm.nih.gov/16397634/).
 RA Fieldhouse D., Yazdani F., Golding G.B.;
 RT "Substitution rate variation in closely related rodent species";
 RL Heredity 78(1):21-31(1997).
 XX
 RN [2]
 RP 1-2283
 RA Fieldhouse D.;
 RT ;
 RL Submitted (07-JUN-1995) to the EMBL/GenBank/DBJ databases.
 RL Dan Fieldhouse, Biology, McMaster University, 1280 Main Street West,
 RL Hamilton, ON, L8S 4K1, Canada
 XX
 FH Key Location/Qualifiers
 FH
 FT [source](#) 1..2283
 FT /organism="Mus pahari"
 FT /mol_type="genomic DNA"
 FT /db_xref="taxon:10093"
 FT [mRNA](#) join(<46..125,256..362,1509..1642,1847..1925,2044..>2186)
 FT /gene="APRT"
 FT /product="adenine phosphoribosyltransferase"
 FT [CDS](#) join(46..125,256..362,1509..1642,1847..1925,2044..2186)

TAXONOMIC DIVISION

PHG Bacteriophage
 ENV Environmental
 FUN Fungal
 HUM *Homo sapiens*
 INV Invertebrates
 MAM Other mammals
 VRT Vertebrate
 MUS *Mus musculus*
 PLN Plant
 PRO Prokaryote
 ROD Other rodents
 SYN Synthetic
 TGN Transgenic
 UNC Unclassified
 VRL Viral

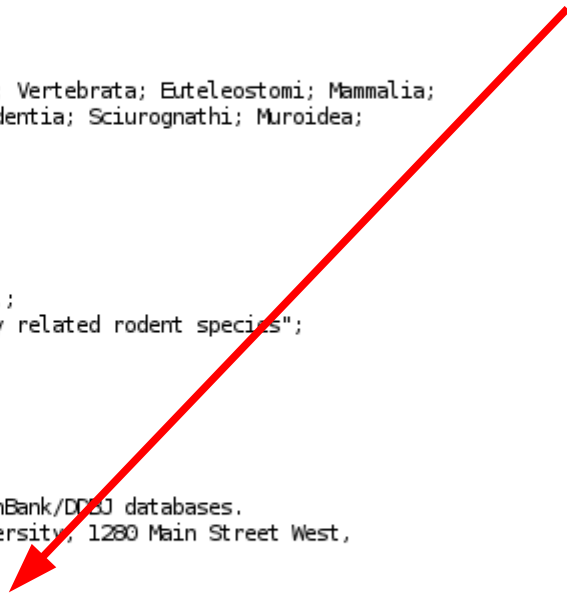
```

ID U28721; SV 1; linear; genomic DNA; STD; ROD; 2283 BP.
XX
AC U28721;
XX
DT 04-JUL-1995 (Rel. 44, Created)
DT 17-APR-2005 (Rel. 83, Last updated, Version 5)
XX
DE Mus pahari adenine phosphoribosyltransferase (APRT) gene, complete cds.
XX
KW .
XX
OS Mus pahari (shrew mouse)
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC Eutheria; Euarchontoglires; Glires; Rodentia; Sciurognathi; Muroidea;
OC Muridae; Murinae; Mus.
XX
RN [1]
RP 1-2283
RX DOI; 10.1038/sj.hdy.6881150
RX PUBMED; 16397634.
RA Fieldhouse D., Yazdani F., Golding G.B.;
RT "Substitution rate variation in closely related rodent species";
RL Heredity 78(1):21-31(1997).
XX
RN [2]
RP 1-2283
RA Fieldhouse D.;
RT ;
RL Submitted (07-JUN-1995) to the EMBL/GenBank/DBJ databases.
RL Dan Fieldhouse, Biology, McMaster University, 1280 Main Street West,
RL Hamilton, ON, L8S 4K1, Canada
XX
FH Key Location/Qualifiers
FH
FT source 1..2283
FT /organism="Mus pahari"
FT /mol_type="genomic DNA"
FT /db_xref="taxon:10093"
FT mRNA join(<46..125,256..362,1509..1642,1847..1925,2044..>2186)
FT /gene="APRT"
FT /product="adenine phosphoribosyltransferase"
FT CDS join(46..125,256..362,1509..1642,1847..1925,2044..2186)

```

FEATURE

**Qualifiers:
SAME FORMAT AS FOR GenBank**



```

FT /translation="MSESELKLVARRIRSFDFPIPGVLFDRDISPLLKDPDSFRASIRL
FT LASHLKSTHSGKIDYIAGLDSRGFLFGPSLAQELGVGCVLIRKQKGLPGPTISASYALE
FT YGKAELEIQKDALEPGQRVVIVDDLATGGTMFAACDLLHQLRAEVVECVSLVELTSLK
FT GRERLGPPIFFSLLQYD"
XX
SQ

```

Sequence 212 PP: 485 A: 696 C: 590 G: 512 T: 0 other;

```

cctgcggata ctcacctcct ccttgctccc tacaagcacg cggccatgtc cgagcttgag 60
ttgaaactgg tggcgcggcg catccgcagc tcccccgact tccccatccc gggcgtgctg 120
ttcaggtgcg gtcacgagcc ggcgaggcgt tggcgccgta ctctcatccc ccggcgcagg 180
cgcgtgggca gccttgggga tcttgccggg cctctgcccg gccacacgcg gtcactctcc 240
tgtccttggt cccagggata tctcgcccc cttgaaagat ccggactcct tccgagcttc 300
catccgcctc ctggccagtc acctgaagtc caegcacagc ggcaagatcg actatatcgc 360
agggcaaggt ggccctgcta ggccgtactc atccccacg gtcctatccc ctatccccct 420
tccccctggt tcacccacag tctaccccc acccatccat tctttcttta acctctgact 480
cttccctcctt ggtttctcac tgccttggac gcttgctcac cccggatgaa ctccgtaggc 540
gtctcccttc cctgcttggg acctaaaggt gccctcgggt cttgctcgta gagacgaact 600
ctgctctgtc cttgtgtcca gaaccaagcc ttcctctttt agggcacaaag gctggccagc 660
atcctgacag caggctggga gacctggatc ctccagatga cggacatcct tacttagggg 720
tagcctctgg gatgaactag atattaaaaa ttaggtaacc ttggggcgtg cctgggcaga 780
cctcaagctc ggtagcttca ggggctgctt ctcccagga ctacaccggg gcactcttct 840
cttgctctcc ccacccccca caagcttgtg ctaaacaact gctgtatacc aggtccatg 900
cttgagcttc agaaacacc ttagggcagct gaatgtccac caggagtgtc cagagggagg 960
gtgagcacc caagagaaca gagtggcct agtaaatgct cagggaccac agaacttttg 1020
cccactccac ttctattgg taccctggc catgccccag aaatcagggc atgtttgtac 1080
cctccccacg acagctcggg ccgcctggaa ctgacctgta gacagtgtc ctgggtagat 1140
gctgcatttg aaaggtggca agagggctgg ttagatggct cagcggttag gagcactgac 1200
tgctcttcca aaggtcctga gttcaaatcc cagcaaccac atggtggctc acaaccacct 1260
acagctacag tgtacacaca tataataaaa taaataaaca aatcttaaaa aaaaaaaaaa 1320
gaaagaaagg tggcaagagc caccatagtg gagaaggcag gtaggatccc caaggctaag 1380
atgctaccga gtaaccatca gtgttcttct agccatagtg ggcaagacct agtgttcta 1440
gtcaatgttg acctctccat acctgectct cggetccatc ccacaccctt cctccttac 1500
cctaacaggt ctagactcca ggggcttctt gtttggccct tccctagctc aggagctggg 1560
cgtgggctgc gtgctcatcc ggaagcaggg gaagctgccg ggccccacta tatcagctc 1620
ctatgctctg gtagtatggga aggtaaggga gctgtgggta gaggaagggc agggcttat 1680
taccacggct accagtgccct aggagtaaat gtgggtgctc agagaggttg agacattggg 1740
gtgaggttta caactcctga aatgctcagc ctcagaaatg ctccaggcta gggaggtggc 1800
cacttgtag catctagact ctcttaacgc tacttctctg ctgcaggctg agctggaaat 1860
ccagaaagat gccttagaac ccgggcagag agtggctatt gtggatgacc tcctggccac 1920
tggaggtaaa gaaccagccc aagacaaaaca ggcttcaaag ggccaggccc tgtctgggtg 1980
gctgactaaa caaagcctt gaataccttc tctttctctg tcccttcccc ccccccccc 2040
caggaaccat gtttgcagcc tgtgatctgc tgcaccagct acgggctgag gtggtggagt 2100
gtgtgagcct ggtggagctg acctcgctga agggcaggga gaggctagga cctataccat 2160
tcttctctct cctccagtat gactgagctg gctagatggt cacaccctg ctcacagcag 2220
cagtaactgc gcggtgctc agccctgggc gcctaagtga cctttgtgag ctacctgctg 2280
ccc 2283

```

Sequence

```

FT /translation="MSESELKLVARRIRSFDFPIPGVLFDRDISPLLKDPDSFRASIRL
FT LASHLKSTHSGKIDYIAGLDSRGFLFGPSLAQELGVGCVLIRKQKGLPGPTISASYALE
FT YGKAELEIQKDALEPGQRVVIVDDLATGGTMFAACDLLHQLRAEVVECVSLVELTSLK
FT GRERLGPPIFFSLLQYD"
XX
SQ

```

```

Sequence 2283 BP; 485 A; 696 C; 590 G; 512 T; 0 other;
cctgcggata ctcacctct cttgtctcc tacaagcagc cggccatgtc cgagtctgag 60
ttgaaactgg tggcgcggcg catccgcagc tcccccgact tccccatccc gggcgtgctg 120
ttcagggtgcg gtcacgagcc ggcgaggcgt tggcgccgta ctctcatccc ccggcgcagg 180
cgcgtgggca gccttgggga tcttgcgggg cctctgcccg gccacacgcg gtaactctcc 240
tgtccttggt cccagggata tctcgcacct cttgaaagat ccggactcct tccgagcttc 300
catccgcctc ctggccagtc acctgaagtc caegcacagc ggcaagatcg actatatcgc 360
agggcaaggt ggcccttgcta ggccgtactc atccccacag gtcctatccc ctatccccct 420
tccccctggt tcacccacag tctacccac accatccat tctttcttta acctctgact 480
cttccctcctt ggtttctcac tgccttggac gcttgttcac cccggatgaa ctccgtaggc 540
gtctcccttc cctgcttggg acctaaaggt gccctcgggt cttgttcgta gagacgaact 600
ctgctctgtc cttgtgtcca gaaccaagcc ttccctcttt agggcacaaag gctggccagc 660
atcctgacag caggctggga gacctggatc ctccagatga cggacatcct tacttgggg 720
tagcctctgg gatgaactag atattaaaaa ttaggtaacc ttggggcgtg cctgggcaga 780
cctcaagtcct ggtagcttca ggggctgctt ctcccagga ctacaccggg gtaactctct 840
cttgtctctc ccacccccca caagcttgtg ctaaacaact gctgtatacc aggtccatg 900
cttgagcttc agaaacacc tagggcagct gaatgtccac caggagcttc cagagggagg 960
gtgagcacc caagagaaca gactggccc agtaaatgct cagggtccac agaacttttg 1020
cccactccac ttctattgg taccctcggc catgccccag aaactagggc atgtttgtac 1080
cctccccacg acagctcggg ccgcttgaa ctgacctgta gacagtgctc ctgggtagat 1140
gctgcatttg aaaggtggca agagggctgg tgagatggct cagcggttag gagcactgac 1200
tgctcttcca aaggtcctga gttcaaatcc cagcaaacac atggtggctc acaaccacct 1260
acagctacag tgtacacaca tataataaaa taaataaaca aatcttaaaa aaaaaaaaaa 1320
gaaagaaagg tggcaagagc caccatagtg gagaaggcag gtaggatccc caaggctaag 1380
atgctaccga gtaaccatca gtgttcttct agccatagtg ggcaagacct agtgttctta 1440
gtcaatggtg acctctccat acctgectct cggetccatc ccacaccctt cctccttac 1500
cctaacaggt ctagactcca ggggcttact gtttggccct tccctagctc aggagctggg 1560
cgtgggctgc gtgctatcc ggaactaggg gaagctgccg ggcctcacta tatcagctc 1620
ctatgctctg gactatggga agactaaggga gctgtgggta gaggaagggc agggcttat 1680
taccacggct accagtccct tggagtaaat gtgggtgctc agagaggttg agacattggg 1740
gtgaggttta caactcctga aatgctcagc ctcagaaatg ctccaggcta gggaggtggc 1800
cacttgttag catctactct ctcttaacgc tacttctctg ctgcaggctg agctggaaat 1860
ccagaaagat gccttagaac ccgggcagag agtggctatt gtggatgacc tcctggccac 1920
tggaggtaaa gaccagccc aagacaaca ggcttcaag ggcctcaggcc tgtctgggtg 1980
gctgactaaa caaagcctt gaatacctc tctttctctg tccctcccc ccccccccc 2040
caggaacctt gtttgcagcc tgtgatctgc tgcaccagct acgggctgag gtgggtggag 2100
gtgtgacctt ggtggagctg acctcgtga agggcagggg agggctagga cctataccat 2160
tctttctct cctccagtat gactgagctg gctagatggt cacaccctg ctcacagcag 2220
cagtaactgc gcggtggctc agccctgggc gcctaagtga cctttgtgag ctacctgctg 2280
2283

```

End of the entry, always //

//

FASTA / Pearson

- Simplest format used
- No feature information is stored in this format
- Each sequence start with “>” and a sequence title
- Can be used for single sequences or multiple sequences (aligned or not)
- Information for homologous sites can be available (“-”).
- used by many software codes

```
>Mus_pahari          607 Weight: 0.75
-----CCTGCGGATACT-C
ACCTCCTCCTTGTCTCCTACAAGCACGCGGCCATGTCCGAGTCTGAGTTG
AAACTGGTGGCGCGGCATCCGCAGCTTCCCCGACTTCCCCATCCCGGG
CGTGCTGTTCAGGTGCGGTACGAGCCGGCGAGGCGTTGGCGCCGTAATC
TCATCCC-CCGGCGCAGGCGCGTGGGCAGCCTTGGGGATCTTGCGGGGCC
TCTGCCCCGCCACACGCGG-TCACTCTCCTGTCTTGTCCCAGGGATAT
CTCGCCCCTCTTGAAAGATCCGGACTCCTTCCGAGCTTCCATCCGCCTCC
TGGCCAGTCACTGAAGTCCACGCACAGCGGCAAGATCGACTATATCGCA
GGGCAAGGTGGCCTTGTAGGCCGTAATCATCCCCACGGTCTATCCCC
TATCCCCTTTCCCC-TCGTGTACCCACAGTCTACCCACACCCATCCAT
TCTTTCTTAACTCTGACTCTTCTCCTTGGTTTCTCACTGCCTTGGAC
GCTTGTTACCCCGGATGAACTCCGTAGGCGTCTCCCTTCCCTGCTTGGT
```

```
>Mus_spicilegus     632 Weight: 0.75
-----TC--GGGATTGACGTGAATTTAGCGTGCTGATACC-T
ACCTCCTCCTTGCCTCCTACACGCACGCGGCCATGTCCGAACCTGAGTTG
AAACTGGTGGCGCGGCATCCGCAGCTTCCCCGACTTCCCAATCCCGGG
CGTGCTGTTCAGGTGCGGTACGAGCCGGCGAGGCGTTGGCGCCGTACGC
TCATCCC-CCGGCGCAGGCGCGTAGGCAGCCTCGGGGATCTTGCGGGGCC
TCTGCCCCGCCACACGCGGGTCACTCTCCTGTCTTGTCCCAGGGATAT
CTCGCCCCTCTTGAAAGACCCGGACTCCTTCCGAGCTTCCATCCGCCTCT
TGGCCAGTCACTGAAGTCCACGCACAGCGGCAAGATCGACTACATCGCA
GG-CGAG-TGGCCTTGTAGGCCGTGCTCGTCCCCACGGTCTAGCCCC
TATCCCCTTTCCCCCTCGTGTACCCACAGTCTGCCCCACACCCATCCAT
TCTTTCTTCAACTCTGACTTCTCCTTGGTTCTCACTGCCTTGGAC
GCTTGTTACCCCGGATGAACTATGTAGGAGTCTCCCTTCCCTGCTAGGT
ACCCTAAGGCATCTGCCCTCGGTGCTTGTTCCTA---GAGACGAACTCTG
CTCT
```

```
>Gerbillus_campestris 615 Weight: 1.49
CCTCCGCCCTTGTTCCTGGGACAGGCTTGACCCTAGCCAGTTGACACCTC
ACCTCCGCCCTTCTC--TCACGCACGCGGCCATGGCGGAACCCGAGTTG
CAGCTGGTGGCGCGGCATCCGCAGCTTCCCCGACTTCCCCATCCCGGG
CGTGCTGTTCAGGTGCGTCCACGAGCCGCCAGGCGTTGGCGCTGCGTCC
TCAGCCCTCCGGCGCAGGCGCGTAGCTGTCTCCGGGATCTTGCGGGGCC
TCCGCCCAGCCATACCCAAGTCAACATCCTGT----GTTCCCAGGGATAT
CTCGCCCCTCTGAAAGACCCGGACTCCTTCCGAGCTTCCATCCGTCTCC
TGGCCAACCATCTGAAGTCCAAGCATGGCGGCAAAATCGACTACATCGCA
GG-CGAG-TGTTCTTGTAGGCCGTGCCCGTTCCC-ACTGTCAGGGCCGC
CATCCCGTGTTCCTT--TTTC-----GTGTCACCCACACCCACCCCT
CCTTTCTCTGACA-CTCCCAAGTTC-CCT--GTTCTCTCTGCCTTGGTC
CCATATTACCCCGGATGA-CTGCG--GAGTCTCC-----
ACCCTCTGACCTCTGCTCTCAAAGCCTGTCCCTACTAGAGAGGAACTCTG
CTCT
```

Mus pahari, Mus spicilegus and Gerbillus campestris partial APRT gene sequences

>Mus_pahari

607 Weight: 0.75

```

-----CCTGCGGATACT-C
ACCTCCTCCTTGTCTCCTACAAGCACGCGGCCATGTCCGAGTCTGAGTTG
AAACTGGTGGCGGGCGCATCCGCAGCTTCCCCGACTTCCCCATCCCGGG
CGTGCTGTTCAGGTGCGGTACGAGCCGGCGAGGCGTTGGCGCCGTACTC
TCATCCC-CCGGCGCAGGCGCGTGGGCAGCCTTGGGGATCTTGCGGGGCC
TCTGCCCCGCCACACGCGG-TCACTCTCCTGTCTTGTCCCAGGGATAT
CTCGCCCCTCTTGAAAGATCCGGACTCCTTCCGAGCTTCCATCCGCCTCC
TGCCAGTCACTGAAGTCCACGCACAGCGGCAAGATCGACTATATCGCA
GGCAAGGTGGCCTTGCTAGGCCGTACTCATCCCCACGGTCTATCCCC
TATCCCCTTTCCCC-TCGTGTACCCACAGTCTACCCACACCCATCCAT
TCTTTCTTAACTCTGACTCTTCTCCTTGGTTTCTCACTGCCTTGGAC
GCTTGTTACCCCCGGATGAACTCCGTAGGCGTCTCCCTTCCCTGCTTGGT

```

Non-essential information

Identifier of the sequence

>Mus_spicilegus 632 Weight: 0.75

```

-----C--GGGATTGACGTGAATTTAGCGTGCTGATACC-T
ACCTCCTCCTTGCCTCCTACACGCACGCGGCCATGTCCGAACCTGAGTTG
AAACTGGTGGCGGGCGCATCCGCAGCTTCCCCGACTTCCCAATCCCGGG
CGTGCTGTTCAGGTGCGGTACGAGCCGGCGAGGCGTTGGCGCCGTACGC
TCATCCC-CCGGCGCAGGCGCGTAGGCAGCCTCGGGGATCTTGCGGGGCC
TCTGCCCCGCCACACGCGGGTCACTCTCCTGTCTTGTCCCAGGGATAT
CTCGCCCCTCTTGAAAGACCCGGACTCCTTCCGAGCTTCCATCCGCCTCT
TGCCAGTCACTGAAGTCCACGCACAGCGGCAAGATCGACTACATCGCA
GG-CGAG-TGGCCTTGCTAGGCCGTGCTCGTCCCCACGGTCTAGCCCC
TATCCCCTTTCCCCCTCGTGTACCCACAGTCTGCCCCACACCCATCCAT
TCTTTCTTCAACTCTGACTTCTCCTTGGTTCTCACTGCCTTGGAC
GCTTGTTACCCCCGGATGAACTATGTAGGAGTCTCCCTTCCCTGCTAGGT
ACCCTAAGGCATCTGCCCTCGGTGCTTGTTCCTA---GAGACGAACTCTG
CTCT

```

" - " symbol for a gap

>Gerbillus campestris 615 Weight: 1.49

```

CCTCCGCCCTTGTTCCTGGGACAGGCTTGACCCTAGCCAGTTGACACCTC
ACCTCCGCCCTTCTC--TCACGCACGCGGCCATGGCGGAACCCGAGTTG
CAGCTGGTGGCGGGCGCATCCGCAGCTTCCCCGACTTCCCCATCCCGGG
CGTGCTGTTCAGGTGCGTCCACGAGCCGCCAGGCGTTGGCGCTGCGTCC
TCAGCCCTCCGGCGCAGGCGCGTGAGCTGTCTCCGGGATCTTGCGGGGCC
TCCGCCCAGCCATACCCAAGTCACCATCCTGT----GTTCCCAGGGATAT
CTCGCCCCTCCTGAAAGACCCGGACTCCTTCCGAGCTTCCATCCGTCTCC
TGCCAACCATCTGAAGTCCAAGCATGGCGGCAAAATCGACTACATCGCA
GG-CGAG-TGTTCTTGCTAGGCCGTGCCCGTTCCC-ACTGTCAGGGCCGC
CATCCCGTGTTCCCTT--TTTC-----GTGTCACCCACACCCACCCCT
CCTTTCTCTGACA-CTCCCAAGTTC-CCT--GTTCTCTCTGCCTTGGTC
CCATATTACCCCCGGATGA-CTGCG---GAGTCTCC-----
ACCCTCTGACCTCTGCTCTCAAAGCCTGTCCCTACTAGAGAGGAACTCTG
CTCT

```


FASTQ

- Format adapted for high-throughput short reads
- Only sequence and quality information is stored in this format
- Each sequence start with “@” and a sequence title (usually machine generated)
- Sequences are on one line
- Third line starts with “+”
- Fourth line stores base quality score

(Phred = ord(Q) - 33)

@HWI-EAS038:8:1:8:697#0/1
AGACTGGCTGGAGCATGTCTATGACGGACTATGATG

+

aaa'[['a'^[^U^_YPU[['ZU^VSTZVX_TB BBB

@HWI-EAS038:8:1:8:1326#0/1

AGACTACCGTGTCTGTCACGACACGGTCTGACGACCAC

+

a^a^\aa'\ZUZVPV\'SP\]aSPQSRNXWBBBBBB


@HWI-EAS038:8:1:8:1305#0/1

AGACTCGAAACGCCTTTCTGGAACACGAAAGGTCTC

+

aXa'_^'aaa_W[\ \ ^^^^VT]a_'[T^'\ 'WSW^W[

Beginning of a new sequence

@HWI-EAS038:8:1:8:697#0/1

AGACTGGCTGGAGCATGTCTATGACGGACTATGATG

+

aaa'[['a'^[^U^_YPU[['ZU^VSTZVX_TB BBB

@HWI-EAS038:8:1:8:1326#0/1

AGACTACCGTGTCTGTCACGACACGGTCTGACGACCAC

+

a^a^\aa'\ZUZVPV\'SP\]aSPQSRNXWBBBBBB

@HWI-EAS038:8:1:8:1305#0/1

AGACTCGAAACGCCTTTCTGGAACACGAAAGGTCTC

+

aXa'_^'aaa_W[\ \ ^ ^ ^ ^VT]a_ '[T^ '\ 'WSW^W[

Sequence title: often machine name, flow cell, x, y coord, etc.

@HWI-EAS038:8:1:8:697#0/1

AGACTGGCTGGAGCATGTCTATGACGGACTATGATG

+

aaa'[['a'^[^U^_YPU[['ZU^VSTZVX_TB BBB

@HWI-EAS038:8:1:8:1326#0/1

AGACTACCGTGTCTGTCACGACACGGTCTGACGACCAC

+

a^a^\aa'\ZUZVPV\'SP\[aSPQSRNXWBBBBBB

@HWI-EAS038:8:1:8:1305#0/1

AGACTCGAAACGCCTTTCTGGAACACGAAAGGTCTC

+

aXa'_^'aaa_W[\ \ ^ ^ ^ ^VT]a_'[T^'\ 'WSW^W[

Start of quality scores

@HWI-EAS038:8:1:8:697#0/1

AGACTGGCTGGAGCATGTCTATGACGGACTATGATG

+

aaa'[['a'^[^U^_YPU[['ZU^VSTZVX_TB BBB

@HWI-EAS038:8:1:8:1326#0/1

AGACTACCGTGTCTGTCACGACACGGTCTGACGACCAC

+

a^a^\aa'\ZUZVPV\'SP\]aSPQSRNXWBBBBBB

@HWI-EAS038:8:1:8:1305#0/1

AGACTCGAAACGCCTTTCTGGAACACGAAAGGTCTC

+

aXa'_^'aaa_W[\ \ ^ ^ ^ ^VT]a_ '[T^ '\ 'WSW^W[

Quality scores for each base

@HWI-EAS038:8:1:8:697#0/1

AGACTGGCTGGAGCATGTCTATGACGGACTATGATG

+

aaa'[['a'^[^U^_YPU[['ZU^VSTZVX_TB BBBB

@HWI-EAS038:8:1:8:1326#0/1

AGACTACCGTGTCGTCACGACACGGTCGACGACCAC

+

a^a^\aa'\ZUZVPV\'SP\[aSPQSRNXWBBBBBB

@HWI-EAS038:8:1:8:1305#0/1

AGACTCGAAACGCCTTTCTGGAACACGAAAGGTCTC

+

aXa'_^'aaa_W[\ \ ^ ^ ^ ^VT]a_ '[T^ '\ 'WSW^W[

@HWI-EAS038:8:1:8:697#0/1
AGACTGGCTGGAGCATGTCTATGACGGACTATGATG

+

aaa'[['a'^[^U^_YPU[['ZU^VSTZVX_TB BBB

@HWI-EAS038:8:1:8:1326#0/1

AGACTACCGTGTCTGTCACGACACGGTCTGACGACCAC

+

a^a^\aa'\ZUZVPV\'SP\]aSPQSRNXWBBBBBB

@HWI-EAS038:8:1:8:1305#0/1

AGACTCGAAACGCCTTTCTGGAACACGAAAGGTCTC

+

aXa'_^'aaa_W[\ \ ^ ^ ^ ^VT]a_'[T^'\ 'WSW^W[

ASCII Table

	30	40	50	60	70	80	90	100	110	120
0		(2	<	F	P	Z	d	n	x
1)	3	=	G	Q	[e	o	y
2		*	4	>	H	R	\	f	p	z
3	!	+	5	?	I	S]	g	q	{
4	"	,	6	@	J	T	^	h	r	
5	#	-	7	A	K	U	_	i	s	}
6	\$.	8	B	L	V	`	j	t	~
7	%	/	9	C	M	W	a	k	u	DEL
8	&	0	:	D	N	X	b	l	v	
9	'	1	;	E	O	Y	c	m	w	

Fastq-sanger

$$Q = \text{ord}(q) - 33$$

Fastq-solexa

$$Q = 10 * \log_{10}(1 + 10^{((\text{ord}(q) - 64) / 10)})$$

Fastq-illumina

$$Q = \text{ord}(q) - 64$$

As of CASAVA 1.8, the Illumina FASTQ variant use 33-offset quality encoding (ASCII '!' = 0) and have a stylized format:

```
@<instrument>:<run number>:<flowcell ID>:<lane>:<tile>:<x-pos>:<y-pos>  
<read>:<is filtered>:<control number>:<index sequence>
```

Specific example:

```
@EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG  
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA  
+  
BBBBCCCC?<A?BC?7@@????????DBBA@@@@A@@
```

Example: Casava 1.8

If $q = \text{'?'}^{\text{'}}$ then $\text{ord}(q) = 63$

$$\text{Fastq-sanger} = 63 - 33 = 30$$

$$\text{Fastq-solexa} = 10 * \log_{10}(1 + 10^{-0.1}) = 2$$

$$\text{Fastq-illumina} = 63 - 64 = \text{"undefined"}$$

Example: older Illumina

If $q = \text{'a'}$ then $\text{ord}(q) = 97$

Fastq-sanger = $97 - 33 = \text{"undefined"}$

Fastq-solexa = $10 * \log_{10}(1 + 10^{3.3}) = 33$

Fastq-illumina = $97 - 64 = 33$

GFF file format; General feature format

Example:

```
##gff-version 3
#!gff-spec-version 1.20
#!processor NCBI annotwriter
##sequence-region NC_004354.3 1 22422827
##species http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=7227
NC_004354.3 RefSeq region 1 22422827 . + .
ID=id0;Dbxref=taxon:7227;chromosome=X;gbkey=Src;genome=chromosome;genotype=y%5B1%5D%3B cn%5B1%5D bw
%5B1%5D sp%5B1%5D%3B Rh6%5B1%5D;mol_type=genomic DNA;old-lineage=Eukaryota%3B Metazoa%3B Arthropoda
%3B Hexapoda%3B Insecta%3B Pterygota%3B Neoptera%3B Endopterygota%3B Diptera%3B Brachycera%3B
Muscomorpha%3B Ephydroidea%3B Drosophilidae%3B Drosophila%3B Sophophora

NC_004354.3 RefSeq region 1 4229 . + . \
ID=id1;Dbxref=FLYBASE:FBti0102096;gbkey=mobile_element
NC_004354.3 RefSeq sequence_variant 148 148 . + . \
ID=id2;Dbxref=dbSNP:207064200;gbkey=variation

NC_004354.3 RefSeq region 10396 10433 . + . \
ID=id888;Dbxref=FLYBASE:FBti0063564;gbkey=mobile_element
```

GFF file format; General feature format

Example:

```
NC_004354.3      RefSeq   region   10396    10433    \
.                +          .        \
ID=id888;Dbxref=FLYBASE:FBti0063564;gbkey=mobile_element
```

A tab delimited file.

#1 "Name" sequence, #2 Source (program/database), #3 name of feature (gene/exon), #4 start, #5 end, #6 confidence value ('.' undefined), #7 strand, #8 frame/phase, #9 description

Tab-delimited

1. QNAME Query template/pair NAME
2. FLAG bitwise FLAG
3. RNAME Reference sequence NAME
4. POS The left most coordinate of the read using the number of the sequence in the reference genome.
5. MAPQ MAPping Quality
The map quality is Phred-scaled. A value of 255 is used for an unknown map quality.
6. CIAGR extended CIGAR string
This string describes features of the match between the read and the reference sequence. In the cases above it is '[0-9M' indicating a perfect match for the length of the read. The format is a number followed by a letter. The number indicates the number of bases and the letter designates a category; M for match, I for an insert in the read, D for a deletion in the read, N for a region skipped, etc.

7. MRNM Mate Reference sequence NaMe

In the cases above it is '*' meaning that there is no mate; these were unpaired reads.

8. MPOS Mate POSition

The bp location in the reference genome where the leftmost bp of the mate read maps.

9. TLEN inferred Template LENgth

The length of the insert between mate pairs.

10. SEQ query SEQuence

The sequence of the read.

11. QUAL query QUALity

The quality is given is $\text{ord}(\text{ASCII}) - 33$ (Sanger Phred scores).

12. OPT variable OPTional fields

GDE

- Tagged file format storing all the information about a sequence (similar to GenBank format)
- Can contain alignment information
- Text enclosed in “{ }”
- All tagged values are enclosed in “”

Offset value

```
{
name "MPU28721"
type "DNA"
longname Mus pahari
sequence-ID "U28721"
descrip "Mus pahari adenine phosphoribosyltransferase (APRT) gene, complete cds"
creator "Fieldhouse,D. and Golding,G.B."
offset 36
creation-date 1/31/98 14:18:24
direction 1
strandedness 1
comments "
NID g881573
KEYWORDS.
SOURCE shrew mouse.
Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
Vertebrata; Eutheria; Rodentia; Sciurognathi; Myomorpha; Muridae;
Murinae; Mus.
REFERENCE 1 (bases 1 to 2283)
TITLE Rates of substitution in closely related rodent species
JOURNAL Unpublished
REFERENCE 2 (bases 1 to 2283)
TITLE Direct Submission
JOURNAL Submitted (07-JUN-1995) Dan Fieldhouse, Biology, McMaster
University, 1280 Main Street West, Hamilton, ON, L8S 4K1, Canada
FEATURES Location/Qualifiers
source 1..2283
/organism=?Mus pahari?
/db_xref=?taxon:10093?
gene join(46..125,256..362,1509..1642,1847..1925,2044..2186)
/gene=?APRT?
```

```

CDS          join(46..125,256..362,1509..1642,1847..1925,2044..2186)
              /gene=?APRT?
              /EC_number=?2.4.2.7?
              /note=?purine salvage enzyme?
              /codon_start=1
              /product=?adenine phosphoribosyltransferase?
              /db_xref=?PID:g881574?
              /translation=?MSESELKLVARRIRSFDFPIPGVLFRRDISPLLKDPDSFRASIR
              LLASHLKSTHSGKIDYIAGLDSRGFLFGPSLAQELGVGCVLIRKQGKLPGPTISASYA
              LEYGKAELEIQKDALEPGQRVVIVDDLLATGGTMFAACDLLHQLRAEVECVSLVELT
              SLKGRERLGPPIFFSLLQYD?

```

```

BASE COUNT   485 a    696 c    590 g    512 t
"

```

```

sequence "CCTGCGGATACTCACCTCCTCCTT
GTCTCCTACAAGCACGCGGCCATGTCCGAGTCTGAGTTGAAACTGGTGGCGCGGCGCATC
CGCAGCTTCCCCGACTTCCCCATCCCGGGCGTGCTGTTCAGGTGCGGTCACGAGCCGGCG
AGGCGTTGGCGCCGTACTCTCATCCC-CCGGCGCAGGCGCGTGGGCAGCCTTGGGGATCT
TGCGGGGCCTCTGCCC GGCCACACGCGG-TCACTCTCCTGTCCTTGTTCCCAGGGATATC
TCGCCCCTCTTGAAAGATCCGGACTCCTTCCGAGCTTCCATCCGCCTCCTGGCCAGTCAC
CTGAAGTCCACGCACAGCGGCAAGATCGACTATATCGCAGGGCAAGGTGGCCTTGCTAGG
CCGTA CT CATCCCCACGGTCCTATCCCCTATCCCCTTTCCCC-TCGTGTCACCCACAGT
CTACCCACACCCATCCATTCTTTCTTTAACCTCTGACTCTTCTCCTTGGTTTCTCACT
GCCTTGGACGCTTGTTACCCCGGATGAACTCCGTAGGCGTCTCCCTTCCCTGCTTGGTA
CCCTAAGG----TGCCCTCGGTGCTTGTTTCGTAGAGACGAACTCTGCTCT"
}

```

Phylip

- Used by a large number of programs (PHYLIP)
- Two different formats
 - sequential format
 - interleaved : 2 numbers on the first line (number of sequences, sequence length)
- No more than 10 characters for the names of the sequences (warning)
- Can store alignment information

Interleaved format

Number of sequences / taxa

Sequence length

3

650

```
Mus_pahari -----
Mus_spicil -----
Gerbillus_ CCTCCGCCCT

TGTCTCCTAC AAGCACGCGG CCATGTCCGA GTCTGAGTTG AAAGTGGTGG CGCGGCGCAT
TGCCTCCTAC ACGCACGCGG CCATGTCCGA ACCTGAGTTG AAAGTGGTGG CGCGGCGCAT
TTCCTC--TC ACGCACGCGG CCATGGCGGA ACCCGAGTTG CAGCTGGTGG CGCGGCGCAT

CCGCAGCTTC CCCGACTTCC CCATCCCGGG CGTGCTGTTC AGGTGCGGTC ACGAGCCGGC
CCGCAGCTTC CCCGACTTCC CAATCCCGGG CGTGCTGTTC AGGTGCGGTC ACGAGCCGGC
CCGCAGCTTC CCCGACTTCC CCATCCCGGG CGTGCTGTTC AGGTGCGTCC ACGAGCCGGC

GAGGCGTTGG CGCCGTACTC TCATCCC-CC GGCAGGCG CGTGGGCAGC CTTGGGGATC
GAGGCGTTGG CGCCGTACGC TCATCCC-CC GGCAGGCG CGTAGGCAGC CTCGGGGATC
CAGGCGTTGG CGCTGCGTCC TCAGCCCTCC GGCAGGCG CGTGAGCTGT CTCCGGGATC

TTGCGGGGCC TCTGCCCGGC CACACGCGG- TCACTCTCCT GTCCTTGTTT CCAGGGATAT
TTGCGGGGCC TCTGCCCGGC CACACGCGGG TCACTCTCCT GTCCTTGTTT CCAGGGATAT
TTGCGGGGCC TCCGCCCAGC CATACCAAG TCACCATCCT GT----GTTT CCAGGGATAT

CTCGCCCCTC TTGAAAGATC CGGACTCCTT CCGAGCTTCC ATCCGCCTCC TGGCCAGTCA
CTCGCCCCTC TTGAAAGACC CGGACTCCTT CCGAGCTTCC ATCCGCCTCT TGGCCAGTCA
CTCGCCCCTC CTGAAAGACC CGGACTCCTT CCGAGCTTCC ATCCGTCTCC TGGCCAACCA

CCTGAAGTCC ACGCACAGCG GCAAGATCGA CTATATCGCA GGCAGAGGTG GCCTTGCTAG
CCTGAAGTCC ACGCACAGCG GCAAGATCGA CTACATCGCA GG-CGAG-TG GCCTTGCTAG
TCTGAAGTCC AAGCATGGCG GCAAAATCGA CTACATCGCA GG-CGAG-TG TTCTTGCTAG
```

Sequential format

3 650

Mus_pahari

-----CCTGCGGATACT-CACCTCCTCCT
TGTCTCCTACAAGCACGCGGCCATGTCCGAGTCTGAGTTGAAACTGGTGGCGCGGCCGCAT
CCGCAGCTTCCCCGACTTCCCATCCCGGGCGTGCTGTTTCCAGGTGCGGTCACGAGCCGGC
GAGGCGTTGGCGCCGTAATCTCATCCC-CCGGCGCAGGCGCGTGGGCAGCCTTGGGGATC
TTGCGGGGCTCTGCCCCGGCCACACGCGG-TCACTCTCCTGTCTTGTTCAGGGATAT
CTCGCCCCCTTTGAAAGATCCGGACTCCTTCCGAGCTTCCATCCGCCTCCTGGCCAGTCA
CCTGAAGTCCACGCACAGCGGCAAGATCGACTATATCGCAGGGCAAGGTGGCCTTGTAG
GCCGTACTCATCCCCACGGTCCTATCCCCTATCCCCTTTCCCC-TCGTGTCACCCACAG
TCTACCCACACCCATCCATTCTTTCTTAACCTCTGACTCTTCCCTCCTTGGTTTCTCAC
TGCCCTTGGACGCTTGTTCACCCCGGATGAACTCCGTAGGCGTCTCCCTTCCCTGCTTGGT
ACCCTAAGG----TGCCCTCGGTGCTTGTTCGTA---GAGACGAACTCTG

Mus_spicil

-----TC--GGGATTGACGTGAATTTAGCGTGCTGATACC-TACCTCCTCCT
TGCCCTCCTACACGCACGCGGCCATGTCCGAACCTGAGTTGAAACTGGTGGCGCGGCCGCAT
CCGCAGCTTCCCCGACTTCCCAATCCCGGGCGTGCTGTTTCCAGGTGCGGTCACGAGCCGGC
GAGGCGTTGGCGCCGTACGCTCATCCC-CCGGCGCAGGCGCGTAGGCAGCCTCGGGGATC
TTGCGGGGCTCTGCCCCGGCCACACGCGGGTCACTCTCCTGTCTTGTTCAGGGATAT
CTCGCCCCCTTTGAAAGACCCGGACTCCTTCCGAGCTTCCATCCGCCTCCTTGGCCAGTCA
CCTGAAGTCCACGCACAGCGGCAAGATCGACTACATCGCAGG-CGAG-TGGCCTTGTAG
GCCGTGCTCGTCCCCACGGTCCTAGCCCCCTATCCCCTTTCCCCCTCGTGTACCCACAG
TCTGCCCCACACCCATCCATTCTTTCTTAACCTCTGACACTTCCCTCCTTGGTTCTCAC
TGCCCTTGGACGCTTGTTCACCCCGGATGAACTATGTAGGAGTCTCCCTTCCCTGCTAGGT
ACCCTAAGGCATCTGCCCTCGGTGCTTGTTCCTA---GAGACGAACTCTG

Gerbillus_

CCTCCGCCCTTGTTCCTGGGACAGGCTTGACCCTAGCCAGTTGACACCTCACCTCCGCC
TTCTC--TCACGCACGCGGCCATGGCGGAACCCGAGTTGCAGCTGGTGGCGCGGCCGCAT
CCGCAGCTTCCCCGACTTCCCATCCCGGGCGTGCTGTTTCCAGGTGCGTCCACGAGCCGGC
CAGGCGTTGGCGCTGCGTCTCAGCCCTCCGGCGCAGGCGCGTGAGCTGTCTCCGGGATC
TTGCGGGGCTCCGCCAGCCATAACCAAGTACCATCCTGT---GTTCCAGGGATAT
CTCGCCCCCTTGAAGACCCGGACTCCTTCCGAGCTTCCATCCGTCTCCTGGCCAACCA
TCTGAAGTCCAAGCATGGCGGCAAAATCGACTACATCGCAGG-CGAG-TGTTCTTGTAG
GCCGTGCCCGTTCCC-ACTGTCAGGGCCGCCATCCCGTGTTCCTT--TTTC-----G
TGTCACCCACACCCACCCCTCCTTTCTCTGACA-CTCCCAAGTTC-CCT--GTTCTCTC
TGCCCTTGGTCCCATATTCACCCCGGATGA-CTGCG---GAGTCTCC-----
ACCCTCTGACCTCTGCTCTCAAAGCCTGTCCCTACTAGAGAGGAACTCTG

NEXUS

- Madison et al. 1997. *Syst. Biol.* 46: 590-621
- Format used with PAUP, McClade and Mr. Bayes
- Composed of different modules
 - starting with “ BEGIN XXXXX:”
 - ending with “END;”
- Standard blocks: TAXA, CHARACTERS, TREE
- Comments are enclosed within “[]”

#NEXUS

[Name: Mus_pahar. Len: 680 Check: 70E718C]
[Name: Mus_spici Len: 680 Check: D5E622FB]
[Name: Gerbillus Len: 680 Check: FBE40A58]

BEGIN TAXA;
DIMENSIONS NTAX=3;
TAXLABELS Mus_pahar Mus_spici Gerbillus;
END;

BEGIN CHARACTERS;
DIMENSIONS NCHAR=680;
FORMAT MISSING=? DATATYPE=DNA INTERLEAVE GAP=-;
MATRIX

Mus_pahari	-----	-----CCTG	CGGATACT-CACCTCCTCT	TGTCTCCTACAAGCACGCGG	CCATGTCCGAGTCTGAGTTG
Mus_spicilegus	-----TC--GGG	ATTGACGTGAATTTAGCGTG	CTGATACC-TACCTCCTCT	TGCCTCCTACACGCACGCGG	CCATGTCCGAACCTGAGTTG
Gerbillus_campestris	CCTCCGCCCTTGTTCTGGG	ACAGGCTTGACCCTAGCCAG	TTGACACCTCACCTCCGCC	TTCTCT--TCACGCACGCGG	CCATGGCGGAACCCGAGTTG
Mus_pahari	AAACTGGTGGCGGGCGCAT	CCGCAGCTTCCCCGACTTCC	CCATCCCAGGCGTGCTGTTT	AGGTGCGGTACAGAGCCGGC	GAGGCGTTGGCGCCGTAATC
Mus_spicilegus	AAACTGGTGGCGGGCGCAT	CCGCAGCTTCCCCGACTTCC	CAATCCCAGGCGTGCTGTTT	AGGTGCGGTACAGAGCCGGC	GAGGCGTTGGCGCCGTAATC
Gerbillus_campestris	CAGCTGGTGGCGGGCGCAT	CCGCAGCTTCCCCGACTTCC	CCATCCCAGGCGTGCTGTTT	AGGTGCGGTACAGAGCCGGC	GAGGCGTTGGCGCCGTAATC
Mus_pahari	TCATCCC-CCGGCGCAGGCG	CGTGGGCAGCCTTGGGGATC	TTGCGGGGCTCTGCCCGGC	CACACGCGG-TCACTCTCT	GTCCTTGTTCCAGGGATAT
Mus_spicilegus	TCATCCC-CCGGCGCAGGCG	CGTAGGCAGCCTCGGGGATC	TTGCGGGGCTCTGCCCGGC	CACACGCGGGTCACTCTCT	GTCCTTGTTCCAGGGATAT
Gerbillus_campestris	TCAGCCCTCCGGCGCAGGCG	CGTGAGCTGTCTCCGGGATC	TTGCGGGGCTCCGCCAGC	CATACCCAAGTCACCATCT	GT----GTTCCAGGGATAT
Mus_pahari	CTCGCCCCTCTTCAAAGATC	CGGACTCCTTCCGAGCTTCC	ATCCGCTCCTGGCCAGTCA	CCTGAAGTCCACGCACAGCG	GCAAGATCGACTATATCGCA
Mus_spicilegus	CTCGCCCCTCTTCAAAGACC	CGGACTCCTTCCGAGCTTCC	ATCCGCTCCTGGCCAGTCA	CCTGAAGTCCACGCACAGCG	GCAAGATCGACTACATCGCA
Gerbillus_campestris	CTCGCCCCTCTTCAAAGACC	CGGACTCCTTCCGAGCTTCC	ATCCGTCTCCTGGCCAACCA	TCTGAAGTCCAAGCATGGCG	GCAAAATCGACTACATCGCA
Mus_pahari	GGGCAAGGTGGCCTTGCTAG	GCCGTACTCATCCCCACGG	TCCTATCCCCTATCCCCTTT	CCCC-TCGTGTACCCACAG	TCTACCCACACCCATCCAT
Mus_spicilegus	GG-CGAG-TGGCCTTGCTAG	GCCGTGCTCGTCCCCACGG	TCCTAGCCCCTATCCCCTTT	CCCCCTCGTGTACCCACAG	TCTGCCCCACACCCATCCAT
Gerbillus_campestris	GG-CGAG-TGTTCTTGCTAG	GCCGTGCCCCGTTCCC-ACTG	TCAGGGCCGCATCCCCTGT	TCCCTT--TTTC-----G	TGTACCCACACCCACCCCT

File identifier

#NEXUS

```
[Name: Mus_pahar.   Len:   680  Check: 70E718C]
[Name: Mus_spici   Len:   680  Check: D5E622FB]
[Name: Gerbillus   Len:   680  Check: FBE40A58]
```

Sequences information

```
BEGIN TAXA;
  DIMENSIONS NTAX=3;
  TAX LABELS Mus_pahar Mus_spici Gerbillus;
END;
```

Block of information

```
BEGIN CHARACTERS;
  DIMENSIONS NCHAR=680;
  FORMAT MISSING=? DATATYPE=DNA INTERLEAVE GAP=-;
  MATRIX
```

```
Mus_pahari      -----
Mus_spicilegus  -----TC--GGG
Gerbillus_campestris  CCTCCGCCCTTGTTCTCTGGG

Mus_pahari      AAACTGGTGGCGCGGCGCAT  CCGCAGCTTCCCCGACTTCC  CCATCCCGGGCGTGCTGTTC  AGGTGCGGTACAGAGCCGGC  GAGGCGTTGGCGCCGTACTC
Mus_spicilegus  AAACTGGTGGCGCGGCGCAT  CCGCAGCTTCCCCGACTTCC  CAATCCCGGGCGTGCTGTTC  AGGTGCGGTACAGAGCCGGC  GAGGCGTTGGCGCCGTACGC
Gerbillus_campestris  CAGCTGGTGGCGCGGCGCAT  CCGCAGCTTCCCCGACTTCC  CCATCCCGGGCGTGCTGTTC  AGGTGCGGTACAGAGCCGGC  GAGGCGTTGGCGCCGTACGC

Mus_pahari      TCATCCC-CCGGCGCAGGGC  CGTGGGCAGCCTTGGGGATC  TTGCGGGGCTCTGCCCGGC  CACACGCGG-TCACTCTCCT  GTCCTTGTTCCCAGGGATAT
Mus_spicilegus  TCATCCC-CCGGCGCAGGGC  CGTAGGCAGCCTCGGGGATC  TTGCGGGGCTCTGCCCGGC  CACACGCGGTCACCTCTCCT  GTCCTTGTTCCCAGGGATAT
Gerbillus_campestris  TCAGCCCTCCGGCGCAGGGC  CGTGAGCTGTCTCCGGGATC  TTGCGGGGCTCCGCCAGC  CATACCCAAGTACCATCCT  GT----GTTCCCAGGGATAT

Mus_pahari      CTCGCCCCTCTTCAAAGATC  CGGACTCCTTCCGAGCTTCC  ATCCGCCTCTTGCCAGTCA  CCTGAAGTCCACGCACAGCG  GCAAGATCGACTATATCGCA
Mus_spicilegus  CTCGCCCCTCTTCAAAGACC  CGGACTCCTTCCGAGCTTCC  ATCCGCCTCTTGCCAGTCA  CCTGAAGTCCACGCACAGCG  GCAAGATCGACTACATCGCA
Gerbillus_campestris  CTCGCCCCTCTTCAAAGACC  CGGACTCCTTCCGAGCTTCC  ATCCGTCTCTTGCCAAACA  TCTGAAGTCCAAGCATGGCG  GCAAAATCGACTACATCGCA

Mus_pahari      GGGCAAGGTGGCCTTGCTAG  GCCGTAATCATCCCCACGG  TCCTATCCCCTATCCCCTTT  CCCC-TCGTGTCACCCACAG  TCTACCCACACCCATCCAT
Mus_spicilegus  GG-CGAG-TGGCCTTGCTAG  GCCGTGCTCGTCCCCACGG  TCCTAGCCCCTATCCCCTTT  CCCCCTCGTGTACCCACAG  TCTGCCCCACACCCATCCAT
Gerbillus_campestris  GG-CGAG-TGTTCTTGCTAG  GCCGTGCCCCTTCCC-ACTG  TCAGGGCCGCCATCCCCTGT  TCCCTT--TTT-----G  TGTCACCCACACCCACCCCT
```

```

Mus_pahari      GGGCAAGGTGGCCTTGCTAG  GCCGTAACTCATCCCCACGG  TCCTATCCCCTATCCCCTTT  CCCC-TCGTGTCACCCACAG  TCTACCCACACCCATCCAT
Mus_spicilegus  GG-CGAG-TGGCCTTGCTAG  GCCGTGCTCGTCCCCACGG  TCCTAGCCCCTATCCCCTTT  CCCCCTCGTGTACCCACAG  TCTGCCCCACCCATCCAT
Gerbillus_campestris  GG-CGAG-TGTTCTTGCTAG  GCCGTGCCCGTTCCC-ACTG  TCAGGGCCGCCATCCCCTGT  TCCCTT--TTTC-----G  TGTACCCACACCCACCCCT

Mus_pahari      TCTTTCTTTAACCTCTGACT  CTTCTCCTTGGTTTCTCAC  TGCCTTGGACGCTTGTTTAC  CCCGGATGAACTCCGTAGGC  GTCTCCCTTCCCTGCTTGGT
Mus_spicilegus  TCTTTCTTCAACCTCTGACA  CTTCTCCTTGGTTCTCAC  TGCCTTGGACGCTTGTTTAC  CCCGGATGAACTATGTAGGA  GTCTCCCTTCCCTGCTAGGT
Gerbillus_campestris  CCTTTCTCTGACA-CTCCCA  AGTTC-CCT--GTTCTCTC  TGCCTTGGTCCCATATTCAC  CCCGGATGA-CTGCG---GA  GTCTCC-----

Mus_pahari      ACCCTAAGG---TGCCCTC  GGTGCTTGTTCGTA---GAG  ACGAACTCTGCTCTGTCCCTT  GTGTCCAGAACCAAGCCTTC
Mus_spicilegus  ACCCTAAGGCATCTGCCCTC  GGTGCTTGTTCCTA---GAG  ACGAACTCTGCTCTGTCCCTT  GTGTCCAGAACCAGGCCTCC
Gerbillus_campestris  ACCCTCTGACCTCTGCTCTC  AAAGCCTGTCCCTACTAGAG  AGGAACTCTGCTCTGTCCAT  GTGTGCAGGGCCAGCTCTTC

```

```

;
END;
BEGIN TREES:
.   TREE tree1 = (Mus_pahar, (Mus_spici,Gerbillus));
.   TREE tree2 = (Mus_spici, (Mus_pahar,Gerbillus));
END;
BEGIN NOTES:
.   PICTURE TAXON=3 FORMAT=GIF SOURCE=FILE
.   PICTURE=a_rodent.gif
END;

```

} Tree

ASN1

- Abstract Syntax Notation
- developed to aid computer access
- made to be read/written by computers

```
Seq-entry ::= set {
  level 1 ,
  class nuc-prot ,
  descr {
    title "Mus pahari adenine phosphoribosyltransferase (APRT) gene, and
translated products" ,
    source {
      org {
        taxname "Mus pahari" ,
        common "shrew mouse" ,
        db {
          {
            db "taxon" ,
            tag
            id 10093 } } ,
        orgname {
          name
          binomial {
            genus "Mus" ,
            species "pahari" } ,
          lineage "Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
Sciurognathi; Muroidea; Muridae; Murinae; Mus" ,
          gcode 1 ,
          mgcode 2 ,
          div "ROD" } } } ,
    create-date
    std {
```

```
year 1995 ,
month 6 ,
day 28 } ,
pub {
  pub {
    article {
      title {
        name "Substitution rate variation in closely related rodent
species." } ,
      authors {
        names
        std {
          {
            name
            name {
              last "Fieldhouse" ,
              initials "D." } } ,
          {
            name
            name {
              last "Yazdani" ,
              initials "F." } } ,
          {
            name
            name {
              last "Golding" ,
              initials "G.B." } } } } ,
```

```
    affil
      str "Department of Biology, McMaster University, Hamilton,
Ontario, Canada." } ,
    from
      journal {
        title {
          iso-jta "Heredity" ,
          ml-jta "Heredity" ,
          issn "0018-067X" ,
          name "Heredity." } ,
        imp {
          date
            std {
              year 1997 ,
              month 1 } ,
          volume "78" ,
          issue " Pt 1" ,
          pages "21-31" ,
          language "eng" } } ,
        ids {
          pubmed 9023989 ,
          medline 97176443 } } ,
        pmid 9023989 ,
        muid 97176443 ,
        gen {
          serial-number 1 } } } ,
    pub {
```



```
pub {
  gen {
    serial-number 2 } ,
  sub {
    authors {
      names
      std {
        {
          name
          name {
            last "Fieldhouse" ,
            initials "D." } } } ,
      affil
      str "Dan Fieldhouse, Biology, McMaster University, 1280 Main
Street West, Hamilton, ON, L8S 4K1, Canada" } ,
      medium other ,
      date
      std {
        year 1995 ,
        month 6 ,
        day 7 } } } } ,
  update-date
  std {
    year 2001 ,
    month 10 ,
    day 15 } } ,
  seq-set {
    seq {
```

```
id {
  genbank {
    name "MPU28721" ,
    accession "U28721" ,
    version 1 } ,
  gi 881573 } ,
descr {
  title "Mus pahari adenine phosphoribosyltransferase (APRT) gene,
complete cds." ,
  molinfo {
    biomol genomic } } ,
inst {
  repr raw ,
  mol dna ,
  length 2283 ,
  seq-data
```

```
ncbi2na
'5E68C745D75FB75C4246694ED62DE2F807AE99A64D649F5587D54D5A9B9EF4AE6B4
625A629BE996C774D55A64A66EA497EA8DF9AA5DE5694466B4775ED7EF54A8CDD95
77E0235A1D7D627D4D65D7A52D1782D46449A42361CCD92A42BA5F9CA5B1D35546
B5CD57355FD576ED1512DC55115353DFDFC177877D75FAFDD1E5FA19FBD1568E075
B29B757D5E7EB15C2B95DAE7EF6C8860779DED7EED481425F5DFF2A4429E94935E1
24A7A885E8D75238684D7C7CAAC977A8E07233C00F2B05FAA6E5EA485D0B7AC9F4A
A79F755287116A91DF77ED7551554427EE701079ECC529D4E7E27D20115CA92783B5
14A2ED48A8AB8915422048BA572C0E74A851207FE54751F5CFAC55694E552034A93B
FB15D5461276A597A0785EC84B9D7AB239E4FE02BA422A7AE23A749AF2891E1E77D4
0AD78BD035490513AE9D1051712712EC444CC300C0C040DF00000008080AE908945
32E88292B28D5429C239C58B0534BBDF72532EA42172EF5CB43BE177531F97769D4D
5115F575F15C12B721D4AA7D7BFA57D5C9D289EA6EA79B9D35A092A82796A551CCD
25D739DE8B3A82B0A89EEAC8A0A92ADF3C51A714B9728B03BAB9D222BE213EAB8AF
C41D780E7497480E7529CA8AE947EF24DC8777C19C7D7B7929E27A035202397C815
A9222EB4FBA385D7A51E8AC08149508404A7D02A5295EDEAB9E1C04099F8317DDFD
DED5F5555554A053BF925EE379E45271A9E2BAE8BBB897AE89E176782A4A88A7285
CC53DF7775D4B38789E9C8EB4455E744924B0799AE9D257A9970B85FEE27179E54'H
}
```

```
annot {
  {
    data
      ftable {
        {
          data
            rna {
              type mRNA ,
              ext
```

```
    name "adenine phosphoribosyltransferase" } ,
partial TRUE ,
location
  mix {
    int {
      from 45 ,
      to 124 ,
      id
        gi 881573 ,
      fuzz-from
        lim It } ,
    int {
      from 255 ,
      to 361 ,
      id
        gi 881573 } ,
    int {
      from 1508 ,
      to 1641 ,
      id
        gi 881573 } ,
    int {
      from 1846 ,
      to 1924 ,
      id
        gi 881573 } ,
```

```

        int {
            from 2043 ,
            to 2185 ,
            id
                gi 881573 ,
            fuzz-to
            lim gt } } } ,
    {
    data
    gene {
        locus "APRT" } ,
    partial TRUE ,
    location
    int {
        from 45 ,
        to 2185 ,
        strand plus ,
        id
            gi 881573 ,
        fuzz-from
        lim lt ,
        fuzz-to
        lim gt } } } } } } ,
seq {
    id {
        genbank {
            accession "AAA68957" ,

```

```
    version 1 } ,
    gi 881574 } ,
descr {
  molinfo {
    biomol peptide ,
    tech concept-trans } ,
  title "adenine phosphoribosyltransferase [Mus pahari]" } ,
inst {
  repr raw ,
  mol aa ,
  length 180 ,
  seq-data
  ncbieaa
```

```
"MSESELKLVARRIRSFDFPIPGVLFRRDISPLLKDPDSFRASIRLLASHLKSTHSGKID
YIAGLDSRGFLFGPSLAQELGVGCVLIRKQGKLPGPTISASYALEYGKAELEIQKDALEPGQRVIVD
DLLATGGTMF
AACDLLHQLRAEVVECVSLVELTSLKGRERLGPPIFFSLLQYD" } ,
```

```
  annot {
    {
      data
      ftable {
        {
          data
          prot {
            name {
              "adenine phosphoribosyltransferase" } ,
```

```
        ec {
            "2.4.2.7" } } ,
    location
    whole
    gi 881574 } } } ,
{
db other ,
name "Annot:CDD" ,
desc {
    name "CDDSearch" ,
    create-date
    std {
        year 2007 ,
        month 6 ,
        day 18 ,
        hour 23 ,
        minute 46 ,
        second 57 } } ,
data
    ftable {
        {
            data
                region "PRK02304" ,
                comment "adenine phosphoribosyltransferase" ,
                location
                    int {
                        from 5 ,
                        to 178 ,
```

```
id
  gi 881574 } ,
ext {
  type
    str "cddScoreData" ,
  data {
    {
      label
        str "definition" ,
      data
        str "PRK02304" } ,
    {
      label
        str "short_name" ,
      data
        str "PRK02304" } ,
    {
      label
        str "score" ,
      data
        int 575 } ,
    {
      label
        str "evaluate" ,
      data
        real { 307765, 10, -64 } } ,
```



```

        {
            label
                str "bit_score" ,
            data
                real { 225315, 10, -3 } } } } ,
    dbxref {
        {
            db "CDD" ,
            tag
                id 74170 } } } } } } } } } ,
annot {
    {
        data
            ftable {
                {
                    data
                        cdregion {
                            frame one ,
                            code {
                                id 1 } } ,
                        comment "purine salvage enzyme" ,
                        product
                            whole

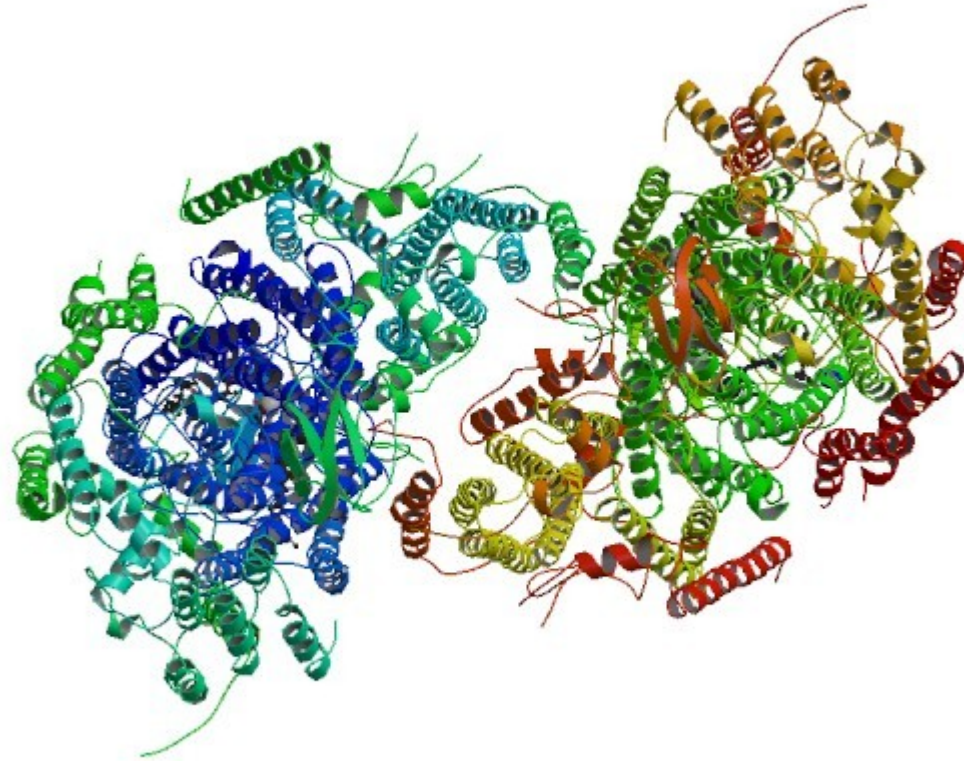
```

```
gi 881574 ,
  location
  mix {
    int {
      from 45 ,
      to 124 ,
      id
      gi 881573 } ,
    int {
      from 255 ,
      to 361 ,
      id
      gi 881573 } ,
    int {
      from 1508 ,
      to 1641 ,
      id
      gi 881573 } ,
    int {
      from 1846 ,
      to 1924 ,
      id
      gi 881573 } ,
    int {
      from 2043 ,
      to 2185 ,
      id
      gi 881573 } } } } } } }
```

PDB

- Protein file
- Store the 3D position of each amino acid
- Primary and secondary structures
- Crystallographic experiments and parameters

CYTOCHROME C OXIDASE



HEADER OXIDOREDUCTASE 26-MAY-98 2OCC
TITLE BOVINE HEART CYTOCHROME C OXIDASE AT THE FULLY OXIDIZED
TITLE 2 STATE
COMPND MOL_ID: 1;
COMPND 2 MOLECULE: CYTOCHROME C OXIDASE;
COMPND 3 CHAIN: A, N;
COMPND 4 SYNONYM: FERROCYTOCHROME C\ :OXYGEN OXIDOREDUCTASE;
COMPND 5 EC: 1.9.3.1;
COMPND 6 OTHER_DETAILS: THIS ENZYME IS A HYBRID PROTEIN COMPLEX AND
COMPND 7 IS A HOMODIMER. ONE MONOMER IS COMPOSED OF 13 DIFFERENT
COMPND 8 SUBUNITS AND SEVEN METAL CENTERS, HEME A, HEME A3, CUA,
COMPND 9 CUB, MG, NA AND ZN.;

SOURCE MOL_ID: 1;
SOURCE 2 ORGANISM_SCIENTIFIC: BOS TAURUS;
SOURCE 3 ORGANISM_COMMON: BOVINE;
SOURCE 4 ORGAN: HEART;
SOURCE 5 TISSUE: HEART MUSCLE;
SOURCE 6 ORGANELLE: MITOCHONDRION;

KEYWDS OXIDOREDUCTASE, CYTOCHROME(C)-OXYGEN, CYTOCHROME C OXIDASE
EXPDTA X-RAY DIFFRACTION
AUTHOR T.TSUKIHARA,M.YAO
REV DAT 1 13-JAN-99 2OCC 0
JRNL AUTH S.YOSHIKAWA,K.SHINZAWA-ITOH,R.NAKASHIMA,R.YAONO,
JRNL AUTH 2 E.YAMASHITA,N.INOUE,M.YAO,M.J.FEI,C.P.LIBEU,
JRNL AUTH 3 T.MIZUSHIMA,H.YAMAGUCHI,T.TOMIZAKI,T.TSUKIHARA
JRNL TITL REDOX-COUPLED CRYSTAL STRUCTURAL CHANGES IN BOVINE
JRNL TITL 2 HEART CYTOCHROME C OXIDASE.
JRNL REF SCIENCE V. 280 1723 1998
JRNL REFN ASTM SCIEAS US ISSN 0036-8075

REMARK 1
REMARK 1 REFERENCE 1
REMARK 1 AUTH T.TSUKIHARA,H.AOYAMA,E.YAMASHITA,T.TOMIZAKI,
REMARK 1 AUTH 2 H.YAMAGUCHI,K.SHINZAWA-ITOH,R.NAKASHIMA,R.YAONO,
REMARK 1 AUTH 3 S.YOSHIKAWA
REMARK 1 TITL THE WHOLE STRUCTURE OF THE 13-SUBUNIT OXIDIZED
REMARK 1 TITL 2 CYTOCHROME C OXIDASE AT 2.8 A
REMARK 1 REF SCIENCE V. 272 1136 1996
REMARK 1 REFN ASTM SCIEAS US ISSN 0036-8075
REMARK 1 REFERENCE 2

REMARK 2 RESOLUTION. 2.30 ANGSTROMS.
REMARK 3
REMARK 3 REFINEMENT.
REMARK 3 PROGRAM : X-PLOR 3.84
REMARK 3 AUTHORS : BRUNGER
REMARK 3
REMARK 3 DATA USED IN REFINEMENT.
REMARK 3 RESOLUTION RANGE HIGH (ANGSTROMS) : 2.30
REMARK 3 RESOLUTION RANGE LOW (ANGSTROMS) : 15.00
REMARK 3 DATA CUTOFF (SIGMA(F)) : 2.000
REMARK 3 DATA CUTOFF HIGH (ABS(F)) : 100000.000
REMARK 3 DATA CUTOFF LOW (ABS(F)) : 0.1000
REMARK 3 COMPLETENESS (WORKING+TEST) (%) : 88.9
REMARK 3 NUMBER OF REFLECTIONS : 278049
REMARK 3
REMARK 3 FIT TO DATA USED IN REFINEMENT.
REMARK 3 CROSS-VALIDATION METHOD : THROUGHOUT
REMARK 3 FREE R VALUE TEST SET SELECTION : RANDOM
REMARK 3 R VALUE (WORKING SET) : 0.209

REMARK 4
 REMARK 4 2OCC COMPLIES WITH FORMAT V. 3.0, 1-DEC-2006
 REMARK 4
 REMARK 4 THIS IS THE REMEDIATED VERSION OF THIS PDB ENTRY.
 REMARK 4 REMEDIATED DATA FILE REVISION 3.101 (2007-05-01)
 REMARK 200
 REMARK 200 EXPERIMENTAL DETAILS
 REMARK 200 EXPERIMENT TYPE : X-RAY DIFFRACTION
 REMARK 200 DATE OF DATA COLLECTION : MAY-1996
 REMARK 200 TEMPERATURE (KELVIN) : 283.0
 REMARK 200 PH : 6.80
 REMARK 200 NUMBER OF CRYSTALS USED : 32
 REMARK 200
 REMARK 200 SYNCHROTRON (Y/N) : Y
 REMARK 200 RADIATION SOURCE : PHOTON FACTORY
 REMARK 200 BEAMLINE : BL-6B
 REMARK 200 X-RAY GENERATOR MODEL : NULL
 REMARK 200 MONOCHROMATIC OR LAUE (M/L) : M

DBREF	2OCC	A	1	514	UNP	P00396	COX1_BOVIN	1	514
DBREF	2OCC	B	1	227	UNP	P00404	COX2_BOVIN	1	227
DBREF	2OCC	C	1	261	UNP	P00415	COX3_BOVIN	1	261
DBREF	2OCC	D	1	147	UNP	P00423	COX4_BOVIN	23	169
DBREF	2OCC	E	1	109	UNP	P00426	COXA_BOVIN	1	109
DBREF	2OCC	F	1	98	UNP	P00428	COXB_BOVIN	1	98
DBREF	2OCC	G	1	84	UNP	P07471	COXD_BOVIN	13	96
DBREF	2OCC	H	1	85	UNP	P00429	COXG_BOVIN	1	85
DBREF	2OCC	I	1	73	UNP	P04038	COXH_BOVIN	1	73
DBREF	2OCC	J	1	59	UNP	P07470	COXK_BOVIN	22	80
DBREF	2OCC	K	1	56	UNP	P13183	COXM_BOVIN	33	88
DBREF	2OCC	L	1	47	UNP	P00430	COXO_BOVIN	17	63

DBREF	2OCC	M	1	46	UNP	P10175	COXQ_BOVIN	25	70
DBREF	2OCC	N	1	514	UNP	P00396	COX1_BOVIN	1	514
DBREF	2OCC	O	1	227	UNP	P00404	COX2_BOVIN	1	227
DBREF	2OCC	P	1	261	UNP	P00415	COX3_BOVIN	1	261
DBREF	2OCC	Q	1	147	UNP	P00423	COX4_BOVIN	23	169
DBREF	2OCC	R	1	109	UNP	P00426	COXA_BOVIN	1	109
DBREF	2OCC	S	1	98	UNP	P00428	COXB_BOVIN	1	98
DBREF	2OCC	T	1	84	UNP	P07471	COXD_BOVIN	13	96
DBREF	2OCC	U	1	85	UNP	P00429	COXG_BOVIN	1	85
DBREF	2OCC	V	1	73	UNP	P04038	COXH_BOVIN	1	73
DBREF	2OCC	W	1	59	UNP	P07470	COXK_BOVIN	22	80
DBREF	2OCC	X	1	56	UNP	P13183	COXM_BOVIN	33	88
DBREF	2OCC	Y	1	47	UNP	P00430	COXO_BOVIN	17	63
DBREF	2OCC	Z	1	46	UNP	P10175	COXQ_BOVIN	25	70

SEQRES	1	A	514	MET	PHE	ILE	ASN	ARG	TRP	LEU	PHE	SER	THR	ASN	HIS	LYS
SEQRES	2	A	514	ASP	ILE	GLY	THR	LEU	TYR	LEU	LEU	PHE	GLY	ALA	TRP	ALA
SEQRES	3	A	514	GLY	MET	VAL	GLY	THR	ALA	LEU	SER	LEU	LEU	ILE	ARG	ALA
SEQRES	4	A	514	GLU	LEU	GLY	GLN	PRO	GLY	THR	LEU	LEU	GLY	ASP	ASP	GLN
SEQRES	5	A	514	ILE	TYR	ASN	VAL	VAL	VAL	THR	ALA	HIS	ALA	PHE	VAL	MET
SEQRES	6	A	514	ILE	PHE	PHE	MET	VAL	MET	PRO	ILE	MET	ILE	GLY	GLY	PHE
SEQRES	7	A	514	GLY	ASN	TRP	LEU	VAL	PRO	LEU	MET	ILE	GLY	ALA	PRO	ASP
SEQRES	8	A	514	MET	ALA	PHE	PRO	ARG	MET	ASN	ASN	MET	SER	PHE	TRP	LEU

SEQRES	1	B	227	MET	ALA	TYR	PRO	MET	GLN	LEU	GLY	PHE	GLN	ASP	ALA	THR
SEQRES	2	B	227	SER	PRO	ILE	MET	GLU	GLU	LEU	LEU	HIS	PHE	HIS	ASP	HIS
SEQRES	3	B	227	THR	LEU	MET	ILE	VAL	PHE	LEU	ILE	SER	SER	LEU	VAL	LEU
SEQRES	4	B	227	TYR	ILE	ILE	SER	LEU	MET	LEU	THR	THR	LYS	LEU	THR	HIS
SEQRES	5	B	227	THR	SER	THR	MET	ASP	ALA	GLN	GLU	VAL	GLU	THR	ILE	TRP
SEQRES	6	B	227	THR	ILE	LEU	PRO	ALA	ILE	ILE	LEU	ILE	LEU	ILE	ALA	LEU
SEQRES	7	B	227	PRO	SER	LEU	ARG	ILE	LEU	TYR	MET	MET	ASP	GLU	ILE	ASN

HET HEA A 515 60
 HET HEA A 516 60
 HET PER A 520 2
 HET HEA N 515 60
 HET HEA N 516 60
 HET PER N 520 2
 HETNAM CU COPPER (II) ION
 HETNAM MG MAGNESIUM ION
 HETNAM NA SODIUM ION
 HETNAM ZN ZINC ION
 HETNAM HEA HEME-A
 HETNAM PER PEROXIDE ION
 FORMUL 27 CU 6(CU 2+)
 FORMUL 28 MG 2(MG 2+)
 FORMUL 29 NA 2(NA 1+)
 FORMUL 32 ZN 2(ZN 2+)
 FORMUL 39 HEA 4(C49 H56 FE N4 O6)
 FORMUL 41 PER 2(O2 2-)

HELIX	1	1	PHE A	2	TRP A	6	1	
HELIX	2	2	HIS A	12	LEU A	41	1	
HELIX	3	3	ASP A	51	ILE A	87	1	
HELIX	4	4	PRO A	95	MET A	117	1	
HELIX	5	5	ALA A	141	ASN A	170	1	
HELIX	6	6	LEU A	183	ASN A	214	1	

5
 30
 37
 23
 30
 32

SHEET	1	A	5	LEU B	116	SER B	120	0		
SHEET	2	A	5	TYR B	105	TYR B	110	-1	N	TYR B 110 O LEU B 116
SHEET	3	A	5	LEU B	95	HIS B	102	-1	N	HIS B 102 O TYR B 105
SHEET	4	A	5	ILE B	150	SER B	156	1	N	ARG B 151 O LEU B 95
SHEET	5	A	5	ASN B	180	LEU B	184	-1	N	LEU B 184 O ILE B 150
SHEET	1	B	3	VAL B	142	PRO B	145	0		
SHEET	2	B	3	ILE B	209	VAL B	214	1	N	GLU B 212 O VAL B 142
SHEET	3	B	3	GLY B	190	GLY B	194	-1	N	GLY B 194 O ILE B 209
SHEET	1	C	2	HIS B	161	VAL B	165	0		
SHEET	2	C	2	LEU B	170	ALA B	174	-1	N	ALA B 174 O HIS B 161
SHEET	1	D	3	ASN F	47	SER F	51	0		

ATOM	5	CB	MET	A	1	53.723	339.248	225.450	1.00	79.30	C
ATOM	6	CG	MET	A	1	54.110	339.452	226.915	1.00	87.90	C
ATOM	7	SD	MET	A	1	55.300	338.229	227.515	1.00	97.07	S
ATOM	8	CE	MET	A	1	54.166	336.799	228.014	1.00	96.59	C
ATOM	9	N	PHE	A	2	53.673	338.579	222.494	1.00	61.89	N
ATOM	10	CA	PHE	A	2	53.412	337.739	221.322	1.00	56.50	C
ATOM	11	C	PHE	A	2	54.569	337.917	220.303	1.00	53.31	C
ATOM	12	O	PHE	A	2	55.076	336.947	219.739	1.00	53.84	O
ATOM	13	CB	PHE	A	2	52.077	338.127	220.683	1.00	55.21	C
ATOM	14	CG	PHE	A	2	51.737	337.334	219.459	1.00	54.54	C
ATOM	15	CD1	PHE	A	2	51.050	336.138	219.565	1.00	55.24	C
ATOM	16	CD2	PHE	A	2	52.126	337.775	218.200	1.00	55.62	C
ATOM	17	CE1	PHE	A	2	50.756	335.388	218.432	1.00	58.99	C
ATOM	18	CE2	PHE	A	2	51.839	337.035	217.059	1.00	57.84	C
ATOM	19	CZ	PHE	A	2	51.155	335.840	217.171	1.00	58.36	C
ATOM	20	N	ILE	A	3	55.010	339.158	220.116	1.00	47.37	N

HETATM28635	CU	CU	A	517	67.173	310.978	190.358	1.00	16.27	CU
HETATM28636	MG	MG	A	518	62.605	315.176	179.115	1.00	19.26	MG
HETATM28637	NA	NA	A	519	42.250	318.661	179.405	1.00	26.18	NA
HETATM28638	CU	CU	B	228	57.527	320.742	171.423	1.00	21.99	CU
HETATM28639	CU	CU	B	229	56.638	319.970	173.568	1.00	24.27	CU
HETATM28640	ZN	ZN	F	99	71.521	300.480	232.843	1.00	33.09	ZN

CONNECT 35128637
CONNECT 47428647
CONNECT 183628635
CONNECT 223928635
CONNECT 224928635
CONNECT 283428636
CONNECT 284228636
CONNECT 290228707
CONNECT 292328647
CONNECT 343128637

CONNECT2888828880
CONNECT288892882928890
CONNECT288902864128889
MASTER 425 0 18 98 30 0 2 928864 26 308 292
END