

Protein sequence databases and sequence annotation at UniProtKB

WELCOME !

14 October 2022

SIB Swiss Institute of Bioinformatics



Friday 14 October 2022

- 09:00 - 13:00 **Protein sequence databases and sequence annotation in UniProtKB** (theory and practicals)
13:00 - 14:00 LUNCH
14:00 - 17:00 **Proteomes & Automated annotation in UniProtKB** (theory and practicals)
17:00 End & exam (deadline: 22h, email MCB) (0.25 ECTS)



Do not hesitate to ask questions !

Chat, Google doc and orally...



Several Coffee Breaks

Marie-Claude.Blatter@sib.swiss

&

Elisabeth.Gasteiger@sib.swiss

&

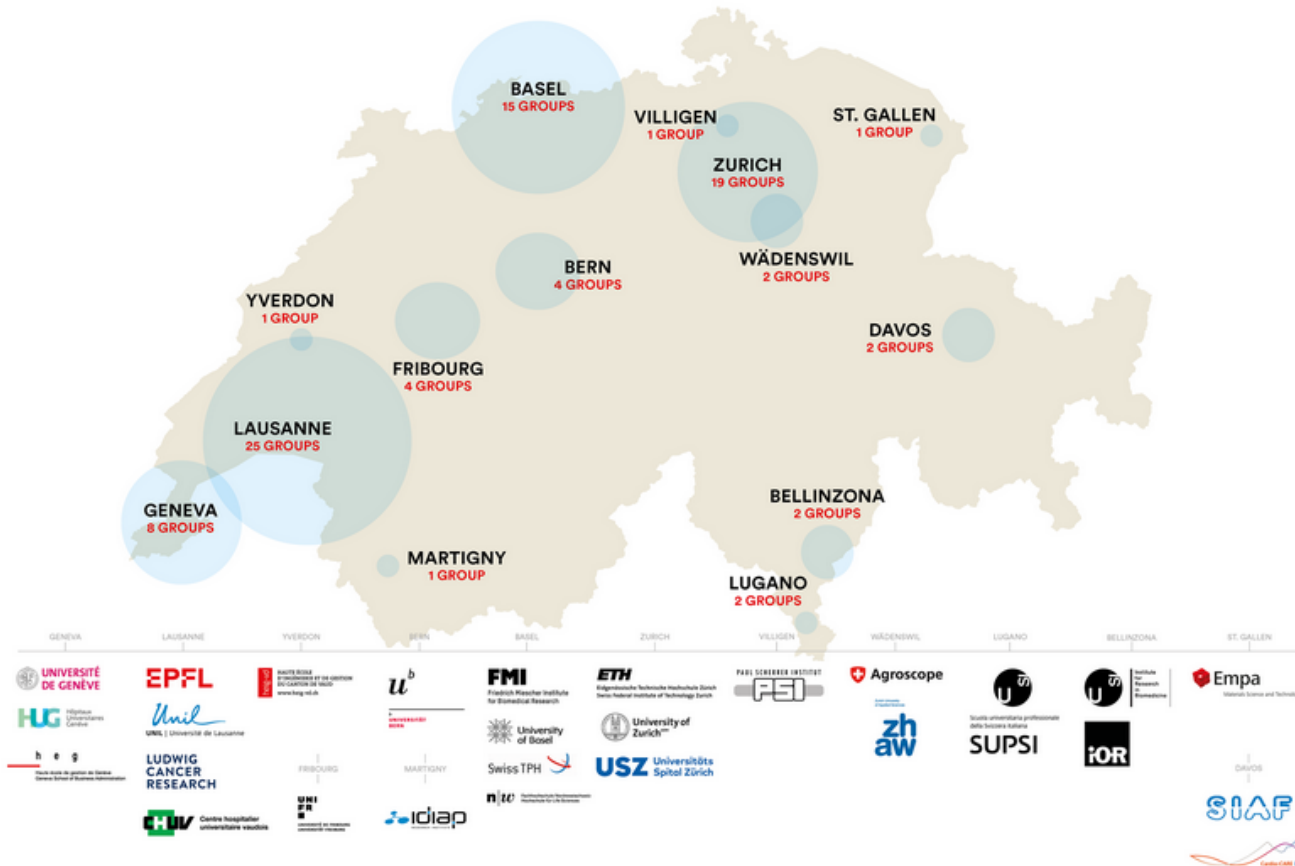
Ivo.Pedruzzi@sib.swiss

Swiss-Prot group, Geneva

SIB Swiss Institute of Bioinformatics

SIB Swiss Institute of Bioinformatics

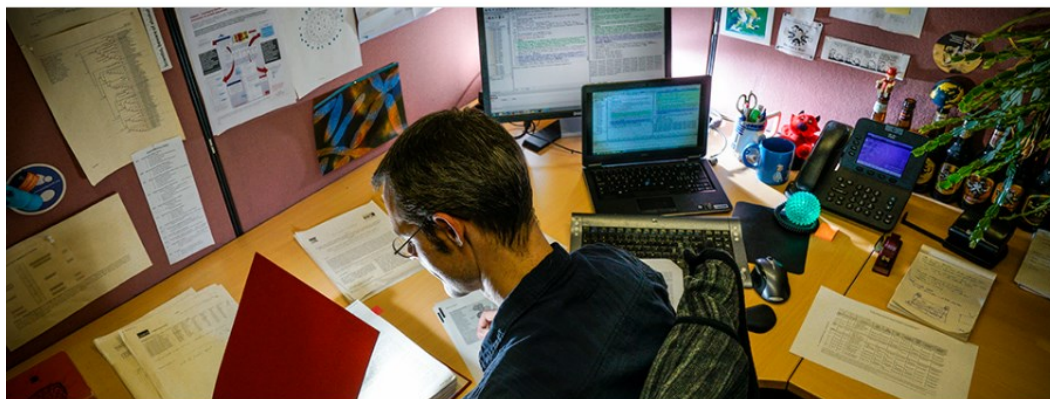
SIB is a federation of bioinformatics research and service groups from the major Swiss schools of higher education and renowned Swiss research institutes:



80 groups, 800 collaborators

www.sib.swiss





Home

Swiss-Prot

The Swiss-Prot team excels in the art of generating machine-readable knowledge of biology from the ever growing body of scientific publications...

Focus on the group's mission

The Swiss-Prot team excels in the art of generating machine-readable knowledge of biology from the ever growing body of scientific publications. It is harnessing the power of deep learning to accelerate literature triage and information extraction, thus delivering the most accurate and informative evidence to users in a timely manner.



“As a competence centre for biocuration and knowledge management, we develop, annotate and maintain internationally renowned knowledge resources such as UniProtKB/Swiss-Prot. Our resources provide an essential framework for biological data science.”

Alan Bridge, Director

Related content

- Happy Evolution Day! And a new website to celebrate...
- Cellosaurus and Rhea join the portfolio of ELIXR Core Data Resources
- Frédérique Lisacek's group
- Amos Bairoch - Lydie Lane's group
- Enhanced enzyme annotation in UniProtKB using Rhea

Domain(s) of activity:

- **Proteins** and proteomes



<https://www.sib.swiss/alan-bridge-group>



What is your current position?


by a guest · less than a minute ago · 




Choose one answer:

- Students (master, PhD)
- Working at university (postdoc, professor)
- Working in industry
- Other

Vote

 Results

 Share

<https://strawpoll.com/polls/6QnM7KOR1Ze>

- Google doc (Q&A):

<https://docs.google.com/document/d/1EGornxLPILbg6GgTVrkQfrny--1gWyR74PZTfkvLqLY/edit#heading=h.erifm2ftu8c>

In one sentence:

Why are your expectations for today ?

Which database(s) are you familiar with ?

Choose one or more answers:

- NCBIInr
- Swiss-Prot
- UniProtKB
- neXtProt
- RefSeq
- PDB
- GenPept
- PIR

<https://strawpoll.com/polls/kogjvakq1g6>

How often do you use UniProt?





by a guest · just now

Make a choice:

- daily
- once a week
- once a month
- less
- never

Vote

 Results

 Share

<https://strawpoll.com/polls/GPgV3K6pAZa>

UniProtKB, protein sequence databases and sequence annotation

Protein sequence and annotation: overview

Nucleic acid sequence databases

INSDC, Ensembl, RefSeq

UniProtKB

UniProtKB/Swiss-Prot

UniProtKB/TrEMBL

Biochemical data (Rhea & ChEBI)

Gene Ontology

UniProt web sites and tools

NCBI Protein sequence databases

RefSeq



UniProtKB, protein sequence databases and sequence annotation

Protein sequence and annotation: overview

Nucleic acid sequence databases

INSDC, Ensembl, RefSeq

UniProtKB

UniProtKB/Swiss-Prot

UniProtKB/TrEMBL

Biochemical data (Rhea & ChEBI)

Gene Ontology

UniProt web sites and tools

NCBI Protein sequence databases

RefSeq



Protein sequences

- > 200 billion 'different' proteins on earth ($\sum N$ species x M genes)
- ~ 227 million 'known and public' protein sequences today
 - 50 % more by next year !

Protein sequences

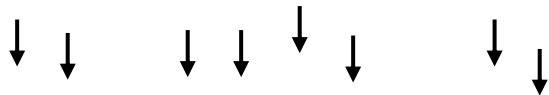
- > 200 billion 'different' proteins on earth ($\sum N$ species x M genes)
- ~ 227 million 'known and public' protein sequences today
 - 50 % more by next year !
- ~ 99% of the protein sequences are derived from the **translation of nucleotide sequences** (mRNA or DNA/genome)
- ~ 1 % come from direct protein sequencing (Edman, MS/MS...)

The life of a protein sequence ...

```

1 agcttctggg cttccagacc cagctacttt ggggaactca gaaacccagg catctctgag
61 tctccgccca agaccgggat gcccccacag aggtgtccgg gagcccagcc ttccocagat
121 aqacagctccg ccagtcoccaa ggttgcccaa ccgctctgac tccctcccg ccaccocagg
181 tctgggca gcccccata cccacagca cctctgcagc agcccctca gcccccagg
241 tctgggca gcccccata cccacagca cctctgcagc agcccctca gcccccagg
301 tctgggca gcccccata cccacagca cctctgcagc agcccctca gcccccagg
361 ccacccccc ccagctctgc agactccctg ggcaccccg gccctctgct gctgtgcgcc
421 gcaaccgctc gctctccgg agccggaccg gggccaccgc gccctctctg ctcocgaccc
    
```

mRNA, genes, genomes, ...

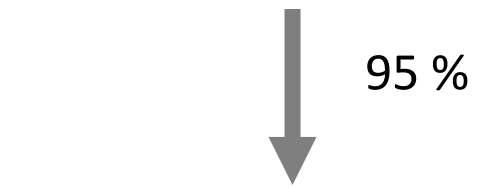


ENA, GenBank, DDBJ

Nucleic acid databases (INSDC)

...if the submitters provide an annotated CoDing Sequence (CDS)

no CDS
Ensembl, RefSeq
gene prediction
4 %



Direct protein sequencing
1 %

```

MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVSASSPPLAVASQSDSKQRR
LLVDPKCGSNMAGQDDISKANSLSMGLVMGRETETKIMGNDLGRDQCCQISLSSGRTDLK
LLECSIANLNRSTSVPENPKSSASTAVSAAPEKEFPKTHSDVSSEQHRLRGQTGTNGN
VKLVTITDITDILQILEPSSSPGKETNESPWRSLLIDENCLLSPLAGEDDSFLLEGN
SNECKRLLPLFLLGKSSVSSSSVSSSSVSSSSVSSSSVSSSSVSSSSVSSSSVSSSSV
YCOASFFGANIIGNKMSAIVHGVTSGGQMYHYDINIASLSQQDQKRFVNIPIPIVVG
SENINRCQSGDDNLTSLGTLNFPGRVTFVSNYSSPSMRPDVSSPPSSSSTATTPPEPKL
CLVCSDEASGCHYGVLTCGCKVVEKRAVGOHNYLCAGNDCTIIDKIRRNKCPACRFKK
CLOAGMNEARKTKKIKGSGVSTGKQVSTGKQVSTGKQVSTGKQVSTGKQVSTGKQVSTG
PEVLYAGYDSSVPDSTWRIMTILNMLGGRVIAAVRARAIFGFRNLHLDQMTLLQYSW
MFTLAPALGPRSVKSSANILGAPDILITNEQVTLGQVQDQTHMNYSSSEIPLQVSY
EEYLCMKRTLLLLSSVPRDGLRSQELPDEIRMTYIKELGKAIKREGNSSQNWRQRFYQLTK
LLDSMHEVVENLLNYCFQTFLDKMTSIEFPEMLAEIITNQIPKYSNGNIKKLLFHQR
    
```

Protein sequence databases

UniProtKB
NCBI protein



GenBank
Nucleotide sequence

```

541 cttcccgga tgaggcccc cgggtgtggtc acccggcgcc ccaggtoctt gagggaaccc
601 ggccaggcgc ggagatgggg gtgcacgggt agtactcggc ggctgggggc tcccggccgc
661 cggggtccct gtttgagcgg ggatttagcg ccccggtcat tggccaggag gtggctgggt
721 tcaaggaccg gcgacttgtc aaggaccccc gaagggggag ggggggtggg cagcctccac
781 gtgccagcgg ggacttgggg gactccttgg ggatggcaaa aacctgacct gtgaaggggg
841 cacagtttgg ggggttaggg gaagaaggtt tggggggttc tgctgtgcca gtggagagga
901 agctgataag ctgataacct gggcgctgga gccaccactt atctgccaga ggggaagcct
961 ctgtcacacc aggattgaag tttggccgga gaagtggatg ctgttagacct ggggggtggg
1021 tgtgcacacg gcagcaggat tgaatgaagg ccaggggaggc agcaccttag tgcttgcatt
1081 gtgggggaca ggaaggacga gctggggcag agacgtgggg atgaaggaag ctgtccttcc
1141 acagccaccc ttctccctcc ccgcctgact ctacagctgg ctatctgttc tagaatgtcc
1201 tgcttggtg tggcttctcc tgcctctgct gtcgctccct ctgggctccc cagtctggg
1261 cggccacca ccgctcatct gtgacagccg agtccctcag aggtacctct tggaggccaa
1321 ggaggccgag aatatcacgg tgagaccctt tcccagcac attccacaga actcacgctc
1381 agggcttcag ggaactcctc ccagatccag gaacctggca ctgtggttgg ggtggagttg
1441 ggaagctaga cactgcccc ctacataaga ataagcttgg tggccccaaa ccatacctgg
1501 aaactaggca aggagcaaa ccagcagatc ctacgcctgt ggccaggggc agagccttca
1561 gggacccttg actccccggg ctgtgtgcat ttacagcggg ctgtgctgaa cactgcagct
1621 tgaatgagaa tatcactgtc ccagacacca aagttaattt ctatgcctgg aagaggatgg
1681 aggtgagttc cttttttttt ttttttctt tcttttggg aatctcattt gcgagcctga
1741 ttttgatga aagggagaa gtcgagggg aaggtaaaaa agatgagcag agatgagcct
1801 gcctgggggc agaggctcac gtctataatc ccaggtgag atggccgaga tgggagaatt
1861 gcttgagccc tggagtttca gaccaaccta ggagcagatg tgagatcccc catctctaca
1921 aacatttaaa aaaattagtc aggtgaagtg gtgcatgggt gtagtcccag atatttggaa
1981 ggctgaggcg ggaggatcgc ttgagcccag gaatttgagg ctgacgtgag ctgtgatcac
2041 accactgcac tccagcctca gtgacagagt gaggccctgt ctcaaaaaaa ggtcactcct
2101 agaaaaataa tgagggtctg atggaatacg ttcattatct attcactcac tcaactcctc
2161 attcattcat tcattcattc aacaagtctt attgcatacc ttctgtttgc tcagcttggg
2221 gcttggggct gctgaggggc aggagggaga gggtgacatc cctcagctga ctcccagagt
2281 ccactccctg taggtcgggc agcaggccgt agaagctctg cagggcctgg cctgtctgtc
2341 ggaagctgtc ctgcccgggc aggcctgtt ggtcaactct tcccagccgt gggagccctc
2401 gcagctgcat gtggataaa cctgctgagg ccttcgcagc ctaccactc tgcttggggc
2461 tctgggagcc caggtgagta ggagcggaca ctctgcttgg ccctttctgt aagaagggga
2521 gaagggtctt gctaaggagt acaggaactg tccgtattcc ttccctttct gtggcactgc
2581 agcgaacctc tgttttctcc ttggcagaag gaagccatct cccctccaga tgggctccta
2641 gctgctccac tccgaacaat cactgctgac actttccgca aactcctccg agtctactcc
2701 aatttctccc ggggaaagct gaagctgtac acaggggagg cctgcaggac aggggacaga
2761 tgaccagggt gtccacctg gccatatcca ccacctccct caccacaatt gcttgtgcca
2821 caccctcccc cgcactcctt gaaccccgtc gaggggctct cagctcagcg ccagcctgtc
2881 ccattggcac tccagtgcga ccaatgacat ctacggggcc agaggaactg tccagagagc
2941 aactctgaga tctaaggatg tcacagggcc aacttgaggg cccagagcag gaagcattca
3001 gagagcagct ttaaactcag ggacagacc atgctgggaa gacgcctgag ctactcggc
3061 accctgcaaa attgatgcca ggacacgctt tggaggcgat ttacctgttt tcgcaacctac
3121 catcagggac aggatgacct ggagaactta ggtggcaagc tgtgacttct ccaggcttca
3181 cgggcatggg cactcctctg gtggcaagag ccccttgac accgggttgg tgggaacctc
3241 gaagacagga tgggggctgg cctctggctc tcattggggtc caacttttgt gtattcttca
3301 acctcattga caagaactga aaccaccaat atgactctgt gcttttctgt tttctgggaa
3361 cctccaaatc cctggtctct gtcccactcc tggcagca

```


Question

<https://www.ncbi.nlm.nih.gov/nucore/x02158>

- Look for this GenBank entry (Human EPO gene)
- Click on the annotated 'CDS' link
- Bonus: click on the link to the UniProtKB entry

Gene EPO (DNA)

GenBank X02158 Nucleotide (DNA) sequence

Annotated CDS submitted by the submitter

CDS

```
join(615..627,1194..1339,1596..1682,2294..2473,2608..2763)
/codon_start=1
/product="erythropoietin"
/protein_id="CAA26095.1"
/db_xref="GDB:119110"
/db_xref="GOA:P01588"
/db_xref="HGNC:HGNC:3415"
/db_xref="InterPro:IPR001323"
/db_xref="InterPro:IPR003013"
/db_xref="InterPro:IPR009079"
/db_xref="InterPro:IPR012351"
/db_xref="InterPro:IPR019767"
/db_xref="PDB:1BUY"
/db_xref="PDB:1CN4"
/db_xref="PDB:1EER"
/db_xref="UniProtKB/Swiss-Prot:P01588"
/translation="MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLQRYLL
EAKEAENITTGCAEHCSSLNENITVPTDKVNFYAWKRMEVQQQAVEVWQGLALLSEAVL
RQQALLVNSSQPWEPLQLHVDKAVSGLRSL/TTLLRALGAQKEAISPPDAASAAPLRTI
TADTFRKLLFRVYSNFLRGLKLLKYTGACRTGDR"
```

Automated translation
-> protein sequence

```
541 cttccccggga tgagggcccc cgggtgtggtc acccggcgcc ccaggtcgct gagggacccc
601 ggccaggcgc ggagatgggg gtgcacggtg agtactcgcg ggctggggcg tcccggccgc
661 ccgggtccct gtttgagcgg ggatttagcg ccccggtat tggccaggag gtggctgggt
721 tcaaggaccg gcgacttgtc aaggaccocg gaagggggag gggggtgggg cagcctccac
781 gtgccagcgg ggacttgagg gagtccttgg ggatggcaaa aacctgacct gtgaagggga
841 cacagtttgg gggttgagg gaagaaggtt tggggggttc tgctgtgcca gtggagagga
901 agctgataag ctgataacct ggagcctgga gccaccactt atctgcaga ggggaagcct
961 ctgtcacacc aggattgaag tttggccgga gaagtggatg ctggtagcct gggggtgggg
1021 tgtgcacacg gcagcaggat tgaatgaagg ccaggggagg agcacctgag tgcctgcatg
1081 gttggggaca ggaaggacga gctggggcag agacgtgggg atgaaggaa ctgtccttcc
1141 acagccacc tctccctcc ccgcctgact ctacgctgg ctatctgttc tagaatgtcc
1201 tggcctggctg tggcttctcc tgtccctgct gtcgctccct ctgggctccc cagtcctggg
1261 cgccccaaca cgctcatct gtgacagcgg agtcctgcag agtacctct tggaggccaa
1321 ggaggccgag aatatcacgg tgagaccct tccccagcac attccacaga actcacgctc
1381 agggcttcag ggaactcct ccagatccag gaacctggca ctgggtttgg ggtggagtgg
1441 ggaagctaga cactgcccc ctacataaga ataagtctgg tggccocaaa ccatacctgg
1501 aactaggca aggagcaag ccagcagatc ctacgctgt ggccagggcc agagccttca
1561 gggacccttg actccccggg ctgtgtgcat ttcagacggg ctgtgtgtaa cactgagctt
1621 tgaatgagaa tatcactgtc ccagacacca aagttaatt ctatgctctg aagaggtagg
1681 aggtgagttc ctttttttt ttttttctt tcttttgag aactcattt gcgagcctga
1741 ttttgatga aagggagaat gatcgagga aaggtaaat ggagcagcag agatgaggct
1801 gcctggggcg agaggctcac gtctataat ccaggctgag atggccgaga tgggagaatt
1861 gcttgagccc tggagtttca gaccaaccta ggcagcatag tgagatcccc catctctaca
1921 aacatttaa aaaattagtc aggtgaagtg gtgcatggtg gtagtcccag atatttggaa
1981 ggcctgaggg ggaggtatcg ttgagccag gaatttgagg ctgctgatcc
2041 accactgac tccagcctca gtgacagagt gaggccctgt ctcaaaaaa aaaagaaaa
2101 agaaaaataa tgagggctgt atggaatac ttcattatc attcactcac tcaactctc
2161 attcattcat tcattcattc aacaagtctt attgcatac ttctgtttg tcagcttgg
2221 gcttggggct gctgaggggc aggagggaga ggtgacatc cctcagctga ctccagagt
2281 ccactcctg taggtcgggc agcaggccgt agaagtctgg cagggctgg cctgctgtc
2341 ggaagctgtc ctgcggggcc aggcctgtt ggtcaactct tcccagccgt gggagccct
2401 gcagctgcat gtggataaag ccgtcagtg ccttcgacg ctccaccact tgcctcgggc
2461 tctggggagc caggtgagta ggagcggaca cttctgctt cctttctgt aagaagggga
2521 gaaggtctct gctaaggagt acaggaactg tccgtattcc tccccttctc gtggcactgc
2581 agcagcctcc tgttttctcc ttggcagaag gaagccatct cccctccaga tggggcctca
2641 gctgctccac tccgaacaat cactgctgac actttccgca aactcttccg agtctactcc
2701 aatttctccc ggggaaagct gaagctgtac acaggggagg cctgcaggac aggggacaga
2761 tgaaccagtg tgtccactg ggcatacca ccactcctc caccacact gcttgtgcca
2821 caccctccc ccactcct gaaccccgtc gaggggctct cagctcagc ccagcctgtc
2881 ccattggcac tccagtgcga ccaatgacat ctacggggcc agaggaactg tccagagagc
2941 aactctgaga tctaaggatg tcacagggcc aacttgagg ccagagcag gaagcattc
3001 gagagcagct ttaaaactcag ggacagacc atgctgggaa gacgcctgag ctcaactggc
3061 accctgcaaa attgatgcca ggacacgctt tggaggcgat ttacctgtt tcgcacctac
3121 catcagggac aggatgacct ggagaactta ggtgcaaac tgtgacttct ccaggtctca
3181 cgggcatggg cactcccttg gtggcaagag ccccttgac accggggtgg tgggaacct
3241 gaagacagga tggggctgg cctctggctc tcatggggtc caactttgt gtattctca
3301 acctcattga caagaactga aaccaccaat atgactctg gcttttctgt tttctgggaa
3361 cctccaaatc ccctggctct gtccactcc tggcagca
```



This CDS is derived from experimental data ?

by a guest · just now

Make a choice:

- yes
- no
- I don't know

Vote

Results

Share

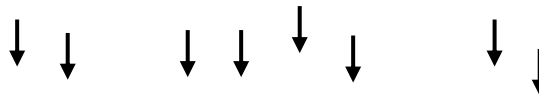
<https://strawpoll.com/polls/mpnbaApVvy5>

The life of a protein sequence ...

```

1 agcttctggg cttccagacc cagctacttt ggggaactca gaaacccagg catctctgag
61 tctccgcca agaccgggat gcccccagg aggtctccg gagcccaacc ttccocagat
121 aqagctccg ccagtcocaa ggtgcccga ccgctcgcac tccctcccg ccaccocagg
181 tccgggca gcccccata cccacagca cctctgcagc agcccctca gcccccagg
241 tccgggca gcccccata cccacagca cctctgcagc agcccctca gcccccagg
301 tccgggca gcccccata cccacagca cctctgcagc agcccctca gcccccagg
361 ccacccccc ccagctcgc agactccctg ggcaccccg gccctctcgt cgtctgcgac
421 gcaaccgctt gctctccg agccggacc gggccaccg gccctctctg ctcocagacc
    
```

mRNA, genes, genomes, ...



ENA, GenBank, DDBJ

Nucleic acid databases (INSDC)

...if the submitters provide an

annotated CoDing Sequence (CDS)
(gene prediction or experimental data)

Not so well documented...

no CDS
Ensembl, RefSeq
gene prediction

4 %

95 %

Direct protein sequencing
1 %

```

MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVSASSPSLAVASQSDSKQRR
LLVDPKCSVSNMDDPDIKSNLSI SMLVYGRTEKIMGNDI GEPDCCQI SLSSGRTDLK
LLECSIANLNRSTSVPENPKSSASTAVSAAPTEKEFPKTHSDVSSEQHRLRGQTGINGN
VKLTTTSTEDPILQLEPSSSPGKETNESPWRSLLIDENCLLSPLAGEDDSFLLEGN
SNECKRLLPLFLEKSSSSTSSSSTSSSSTSSSSTSSSSTSSSSTSSSSTSSSSTSSS
YCOASFFGANIIGNKMSAIVHGVTSGGQMYHYDINIASLSQDDQKRFVNIPIPIVVG
SENINRCQSGDDNLTSLGTLNFPGRITVFSNGYSSPSMRPDVSSPPSSSSTATTGPPK
CLVCSDEASGCHYGLVTCGCKVVEKRAVGOHNYLCAGBNDCTIIDKIRRNCPACRFK
CLOAGMNEARKTKKIKGKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQKQK
PEVLYAGYDSSVPDSTWRIMTILNMLGGRVIAAVRRAKIPFERNLHDDQMTLQYSW
MFTLAKALGKPSVQSSANILGAPDILITNEQVITLQVYDQCTHMVYSSSEIHLQYS
EEYLCMKRTLLLLSSVPKGLKSQLPDEIRMTYIKELGKAIKREGNSSQNWRQRYQLTK
LLDSMHEVVENLLNYCFQTFLDKIMSIEFPEMLAEIITNQIPKYSNGNIKKLLFHQR
    
```

Protein sequence databases

UniProtKB
NCBI protein

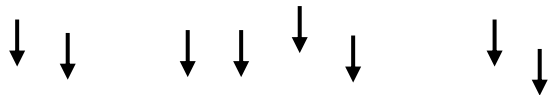


The life of a protein sequence ...

mRNA, genes, genomes, ...

```

1 agattctggg ctccagacc cagctacttt gggaaactca gaaacccagg catctctgag
61 tctccgcca agaccgggat gcccaccagg aggtctccg gagcccaagc ttccocagat
121 aqagctccg ccagtcocaa ggtgcccga ccgctcgcac tccctcccg ccaccocagg
181 tccgggca gccccatga cccagccga cctctgcagc agcccatga gcccccagg
241 tccagag gcaatgag gcaatgag gcaatgag gcaatgag gcaatgag gcaatgag
301 tcccaaca gcccaccagg aggtctccg gagcccaagc ttccocagat
361 ccacccccc ccagctcgc agactccctg ggcaccccg gccctctcgt gctgtgcgcc
421 gcaaccgct gtctcccg agccgaccg gggccaccg gccctctcgt ctccagacc
    
```



Nucleic acid databases (INSDC)

ENA, GenBank, DDBJ

...if the submitters provide an annotated CoDing Sequence (CDS) (gene prediction or experimental data)

no CDS
Ensembl, RefSeq
gene prediction
4 %

95 %

Protein sequence databases

```

MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVSASSPPLAVASQSDSKQRR
LLVDPKCSNSNDQQDDLSKANSLSMGLVYCFETETKIMGNDLGRPQQQQLSLSSGFDLK
LLEESIANLNRSTSVPENPKSSASTAVSAAPTEKEFPKTHSDVSSEQHRLKQGTGNGN
VKLTTTSTPDILOLEPSSSPGKETNESPWRSDLLIDENCLLSPLAGEDDSFLLEGN
SNECKRLLPLPHEPHEPHEPHEPHEPHEPHEPHEPHEPHEPHEPHEPHEPHEPHEPHE
YCOASFFGANIIGNKMSAIVHGVTSGGQMYHYDINTASLSQQDQKRFVNIPIPIVVG
SENINRCQSGDDNLTSLGTLNFPGRITVFSNGYSSPSMRPDVSSPPSSSSTATTPPEKL
CLVCSDEASGCHYGLVTCGCKVPEKRAVGOHNYLCAGBNDCTIDKIRRNKCPACRFK
CLOAGMNEARKTKKIKGKGLGKGLGKGLGKGLGKGLGKGLGKGLGKGLGKGLGKGLGK
PEVLYAGYDSSVPDSTWRMTLNLMLGGRVIAAVRARAIFPFRNLHLDQMTLLQYSW
MFTLAPALGKPSVYDSSANILGAPDILINRQVTLRQVYDQCTHMVYSSFLRDLQYSY
EEYLCMKRTLLLLSSVPKGLKSQLPDEIRMTYIKELGKAVKREGNSSQNWRQRFYQLTK
LLDSMHEVVENLLNYCFQTFLDKTMSEIEPPEMLAEIITNQIPKYSNGNIKKLLFHQR
    
```

UniProtKB
NCBI protein

Direct protein sequencing
1 %



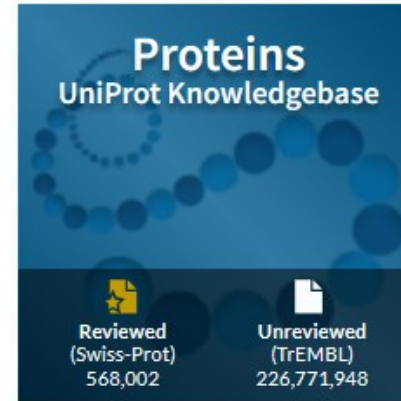


Protein sequence databases organization

PIR, PDB, PRF,
Ensembl, RefSeq
integration + cross-links

UniProtKB: Swiss-Prot + TrEMBL

www.uniprot.org



NCBI protein: Swiss-Prot + GenPept + RefSeq + PIR + PDB + PRF

Protein

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

<https://www.ncbi.nlm.nih.gov/protein/>

The definition of 'redundancy' varies among databases

UniProtKB/Swiss-Prot

one record – one gene – one or several protein sequences

UniProtKB/TrEMBL


one record – one protein sequence


RefSeq





















one record – one mRNA sequence – one protein sequence

Due to these different concepts,
it is not possible to draw conclusions
on the quality and completeness of a database
according to the number of entries.

(gene:GYPA) AND (taxonomy_id:9606)

 Reviewed (Swiss-Prot) (1)

 Unreviewed (TrEMBL) (49)

<input type="checkbox"/>	A0A0C4DFT7		A0A0C4DFT7_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	150 AA
<input type="checkbox"/>	A0A7G1PFV2		A0A7G1PFV2_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	150 AA
<input type="checkbox"/>	B8Q183		B8Q183_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	150 AA
<input type="checkbox"/>	P02724		GLPA_HUMAN	Glycophorin-A[...]	GYPA, GPA	Homo sapiens (Human)	150 AA
<input type="checkbox"/>	Q58HE7		Q58HE7_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	150 AA
<input type="checkbox"/>	X5M4Z9		X5M4Z9_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	150 AA
<input type="checkbox"/>	A0A087WU29		A0A087WU29_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	148 AA
<input type="checkbox"/>	A0A2R8Y7F9		A0A2R8Y7F9_HUMAN	Glycophorin-A	GYPA	Homo sapiens (Human)	145 AA
<input type="checkbox"/>	E9PD10		E9PD10_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	137 AA
<input type="checkbox"/>	Q8WWP1		Q8WWP1_HUMAN	Glycophorin A	GYPA	Homo sapiens (Human)	132 AA
<input type="checkbox"/>	Q8WWP2		Q8WWP2_HUMAN	Glycophorin A	GYPA	Homo sapiens (Human)	132 AA
<input type="checkbox"/>	Q8WWP3		Q8WWP3_HUMAN	Glycophorin A	GYPA	Homo sapiens (Human)	132 AA
<input type="checkbox"/>	Q8WWP4		Q8WWP4_HUMAN	Glycophorin A	GYPA	Homo sapiens (Human)	132 AA
<input type="checkbox"/>	E7EQF3		E7EQF3_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	118 AA
<input type="checkbox"/>	E9PH25		E9PH25_HUMAN	Glycophorin-A	GYPA	Homo sapiens (Human)	105 AA
<input type="checkbox"/>	K9JI14		K9JI14_HUMAN	Glycophorin-A	GYPA	Homo sapiens (Human)	104 AA
<input type="checkbox"/>	Q13030		Q13030_HUMAN	Glycophorin Erik I-IV[...]	GYPA, GPErik, hCG_2026259	Homo sapiens (Human)	85 AA
<input type="checkbox"/>	A0A8E8D7U9		A0A8E8D7U9_HUMAN	GPA	GYPA	Homo sapiens (Human)	77 AA
<input type="checkbox"/>	A0A8E8D9B7		A0A8E8D9B7_HUMAN	GPA	GYPA	Homo sapiens (Human)	77 AA
<input type="checkbox"/>	G8CW02		G8CW02_HUMAN	Glycophorin A	GYPA	Homo sapiens (Human)	77 AA

(...)

Human GYPA gene and its protein sequences
Biology and/or redundancy



Protein

Protein

(gypa[Gene]) AND "Homo sapiens"[porgn]

Create alert Advanced

Species

Animals (537)

Customize ...

Source databases

RefSeq (13)

UniProtKB / Swiss-Prot (1)

Customize ...

Sequence length

Custom range...

Molecular weight

Custom range...

Release date

Custom range...

Revision date

Custom range...

[Clear all](#)

[Show additional filters](#)

Summary 20 per page Sort by Default order

Send to:

See the [results of this search \(2 items\)](#) in our new [Identical Protein Groups](#) database.

Items: 1 to 20 of 537

<< First < Prev Page 1 of 27 Next > Last >>

- [RecName: Full=Glycophorin-A; AltName: Full=MN sialoglycoprotein; AltName: Full=PAS-2; AltName: Full=Sialoglycoprotein alpha; AltName: CD_antigen=CD235a; Flags: Precursor](#)

150 aa protein

Accession: P02724.2 GI: 259016238

[PubMed](#) [Taxonomy](#)

[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

Swiss-Prot

- [glycophorin-B isoform 1 precursor \[Homo sapiens\]](#)

2. 91 aa protein

Accession: NP_002091.4 GI: 1736863041

[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)

[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

RefSeq

(...)

NCBI protein: Swiss-Prot + GenPept + RefSeq + PIR + PDB + PRF



04

```
/translation="MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLQRYLL  
EAKEENITTGCAEHCSLNENITVPDTRKVNIFYANKRMEVGGQAVEVWQGLALLSEAVL  
RGQALLVNSSQPWEPLQLHVDRVSGLRSLTTLRLALGAQKEAISPPDAASAAPLRTI  
TADTFRKLFRVYSNFLRGKLLKLYTGEACRTGDR'LLRALGAQKEAISPPDAAS/
```

Biocuration

From Wikipedia, the free encyclopedia

Biocuration is the field of [life sciences](#) research dedicated to translating and integrating biomedical knowledge from scientific articles to interoperable databases.^{[1][2]} The biocuration of biomedical knowledge is made possible by the cooperative work of biocurators, [software developers](#) and [bioinformaticians](#).^[1]

= adding biological information (annotation) mainly to a (reviewed or not) protein sequence.

- **expert biocurators** (reviewed) - source: publication & prediction and validation
- **automated** (unreviewed) - source: prediction

- **Free text**
- **Controlled vocabulary (CV)**, i.e. Keywords, in-house CV...
- **Ontology**, i.e. Gene Ontology, ChEBI, ...

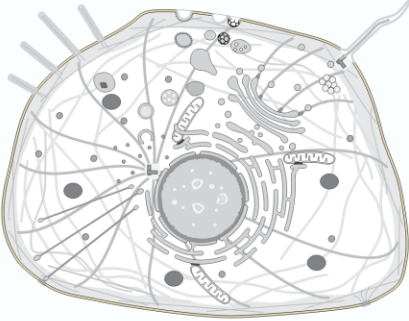
Example of biocuration/annotation (subcellular location) (same gene, same protein sequence)

P02724

UniProtKB/Swiss-Prot annotation

Subcellular Location

UniProt Annotation GO Annotation



Cell membrane 1 Publication; Single-pass type I membrane protein 1 Publication
Appears to be colocalized with SLC4A1.

The diagram shows a cell with various organelles. A red box highlights the 'UniProt Annotation' tab. The 'Cell membrane' is selected, and a legend indicates it is a single-pass type I membrane protein with one publication. A note states it appears to be colocalized with SLC4A1.

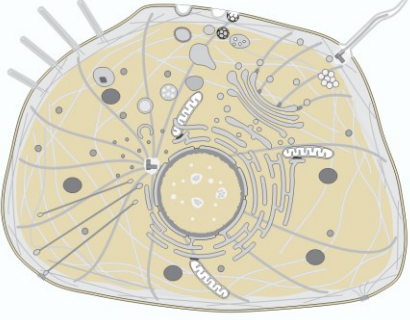
- **expert biocurators** (reviewed) - source: **publication** & prediction
- **automated** (unreviewed) - source: prediction
- **Free text**
- **Controlled vocabulary (CV)**, i.e. **in-house CV**, Keywords, ...
- **Ontology**, i.e. Gene Ontology, ChEBI, ...

A0A0C4DFT7

UniProtKB/TrEMBL annotation

Subcellular Location

UniProt Annotation GO Annotation



cytosol
integral component of plasma membrane
membrane
nucleoplasm
plasma membrane
Complete GO annotation on QuickGO

The diagram shows a cell with various organelles. A red box highlights the 'GO Annotation' tab. The legend lists several subcellular locations: cytosol, integral component of plasma membrane, membrane, nucleoplasm, and plasma membrane, all of which are currently unselected.

- **expert biocurators** (reviewed) - source: publication & prediction
- **automated** (unreviewed) - source: prediction
- **Free text**
- **Controlled vocabulary (CV)**, i.e. in-house CV, Keywords, ...
- **Ontology**, i.e. **Gene Ontology**, ChEBI, ...

Human gene: GYPA @ UniProtKB

NP_002090.4

RefSeq annotation

Summary: Glycophorins A (GYPA) and B (GYPB) are major sialoglycoproteins of the human erythrocyte membrane which bear the antigenic determinants for the MN and Ss blood groups. In addition to the M or N and S or s antigens that commonly occur in all populations, about 40 related variant phenotypes have been identified. These variants include all the variants of the Miltenberger complex and several isoforms of Sta, as well as Dantu, Sat, He, Mg, and deletion variants Ena, S-s-U- and Mk. Most of the variants are the result of gene recombinations between GYPA and GYPB. [provided by RefSeq, Jul 2008].

- **expert biocurators** (reviewed) - **source:** publication & prediction -> provided by RefSeq
- **automated** (unreviewed) - source: prediction
- **Free text**
- **Controlled vocabulary**, i.e. in-house CV, Keywords, ...
- **Ontology**, i.e. Gene Ontology, ChEBI, ...

Human gene: GYPA @ RefSeq

We will discuss questions such as (theory and practices):

- Where do the protein sequences come from?
- What are the differences between the major protein sequence databases?
- What are the manual and automated gene / protein annotation pipelines?
- What are the Gene Ontology (GO) annotation pipelines?
- How to assess protein sequence accuracy and annotation quality?
- How to extract biological knowledge from a Blast result or gene list?

https://www.sib.swiss/training/course/20221014_PRODB

UniProtKB, protein sequence databases and sequence annotation

Protein sequence and annotation: overview

Nucleic acid sequence databases

INSDC, Ensembl, RefSeq

UniProtKB

UniProtKB/Swiss-Prot

UniProtKB/TrEMBL

Biochemical data (Rhea & ChEBI)

Gene Ontology

UniProt web sites and tools

NCBI Protein sequence databases

RefSeq

The life of a protein sequence ...

```

1 agcttctggg cttccagacc cagctacttt ggggaactca gaaacccagg catctctgag
61 tctccgcca agaccgggat gcccccagg aggtctccg gagcccaacc ttccocagat
121 aqagctccg ccagtcccaa ggtgcccga ccgctcgcac tccctcccg ccaccocagg
181 cccgggca gcccccata ccccaagca cctctgcagc agcccctca gcccccagg
241 cccagg gctcccaag gctcccaag gctcccaag gctcccaag gctcccaag
301 ccccaag gctcccaag gctcccaag gctcccaag gctcccaag gctcccaag
361 ccaccccagg ccagctcgc agactccctg ggcaccccg gccctctcgt gctgtgcgc
421 gaaaccgct gctccccg agccggacc gggccaccg gccctctctg ctcagacacc
    
```

mRNA, genes, genomes, ...

Nucleic acid
databases (INSDC)

ENA, GenBank, DDBJ

...if the submitters provide an
annotated CoDing Sequence (CDS)
(gene prediction or experimental data)

no CDS
Ensembl, RefSeq
gene prediction

4 %

95 %

Direct protein sequencing
1 %

Protein sequence
databases

```

MDSKESLTPGREENPSSVLAQERGDVMDYFKTLRGGATVKVSASSPSLAVASQSDSKQRR
LLVDPKCSVSNMQDDIISKANSISMGIVMGTETKIMGNDIQRDCCQIISLSSGTDLK
LLECSIANLNRSTSVPENPKSSASTAVSAAPEKEFPKTHSDVSSEQHRLRGQTGNGN
VKLTTTSTEDILQLEPSSSPGKETNESPWRSLLIDENCLLSPLAGEDDSFLLEGN
SNECKRLLPLFHEGNSVPSSEKQREKIEGKIEGKIEGKIEGKIEGKIEGKIEGKIEGK
YCOASFFGANIIGNKMSAIVHGVTSGGQMYHYDINTASLSQQDQKRFVNIPIPIVVG
SENINRCQSGDDNLTSLGTLNFPGRVTFVSNYSSPSMRPDVSSPPSSSSTATTPPKL
CLVCSDEASGCHYGLVTCGCKVPEKRAVGOHNYLCAGNDCTIIDKIRRNCPACRFK
CLOAGMNEARKTKKIKGKIEGKIEGKIEGKIEGKIEGKIEGKIEGKIEGKIEGKIEGK
PEVLYAGYDSSVPDSTWRIMTILNMLGGRVIAAVKRAKIFGRNLHLDQMTLLQYSW
MFTLAKALGKPSVQSSANILGKADLIITNEQVTLKQVQDQKRFVNIPIPIVVG
EEYLCMKRTLLLLSSVPRDGLKSQELPDEIRMTYIKELGKAIKREGNSSQNWRQRFYQLTK
LLDSMHEVVENLLNYCFQTFDLKTMSEIEPPEMLAEIITNQIPKYSNGNIKKLLFHQR
    
```

UniProtKB
NCBI protein



ENA/GenBank/DDBJ

INSDC

About Policy Advisors Submitting Standards News & Announcements Documents



The International Nucleotide Sequence Database Collaboration (INSDC) is a long-standing foundational initiative that operates between [DDBJ](#), [EMBL-EBI](#) and [NCBI](#).

INSDC covers the spectrum of data raw reads, through alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations



NCBI



Databases

Data type	DDBJ	EMBL-EBI	NCBI
Next Generation reads	Sequence Read Archive		Sequence Read Archive
Assembled Sequences	DDBJ	European Nucleotide Archive	GenBank
Samples	BioSample		BioSample
Studies	BioProject		BioProject

DNA Data Bank of Japan, Mishima, Japan.

EMBL-EBI, European Nucleotide Archive, Cambridge, UK.

GenBank, NCBI, Bethesda, MD, USA.

<http://www.insdc.org/>

ENA/GenBank/DDBJ

Serve as nucleic sequence archives (primary databases):

- Contain all public sequences derived from:

- Genome projects (> 80 % of entries)
- Sequencing centers (cDNAs, ESTs, RNAs...)
- Individual scientists (15 % of entries)
- Patent offices (i.e. European Patent Office, EPO)

! Sequence quality !
Protein sequences in UniProtKB
map to genome sequences

- Sequences from > 1'250'000 different species;
- It contains about 3'000 millions sequence entries

- An example...

Complete nucleotide sequence of the fast-twitch isoform of chicken skeletal muscle α -tropomyosin

C.Gooding*, F.C.Reinach⁺ and A.R.Macleod

Individual scientists

Ludwig Institute for Cancer Research, MRC Centre, Hills Road, Cambridge CB2 2QH, UK

Submitted September 10, 1987

Accession no.Y00456

The complete nucleotide sequence of the mRNA encoding the fast-twitch isoform of chicken skeletal muscle α -tropomyosin has been determined. The sequence comprises 136 nucleotides of the 5' untranslated region, an open reading frame of 852 nucleotides encoding a 284 amino-acid tropomyosin and the complete 3' untranslated sequence of 205 nucleotides. This sequence differs from the partial sequence reported previously (1) at eight positions. These are 315 (T to C), 450 (G to C), 630 (T to C), 670 (A to G), 995 (deletion of T), 1085 (insertion of T), 1093 (insertion of C) and 1104 (deletion of A). Of the four coding sequences changes only one (630) results in an amino-acid substitution (V to A).

```

AATTCGCGCGTCCCGCACTCGTTGGCCCCAGCGCTCCCCGGGGCCGCGGCTCGGATTTCGTATCGGGGCTCTCCGCGCCTTTCTGCTCTG   90
                                     M D A I K K K M Q M L K L D K
AATTCGGCTGTACTTCTCGCGGGAACGGCCCTAACCCACCGCCGCCATGGATGCCATCAAGAAAAAGATGCAGATGCTGAAACTGGACAA   180
   E N A L D R A E Q A E A D K K A A E E R S K Q L E D E L V A
GGAGAATGCCTTGGACAGAGCCGAGCAAGCCGAAGCGGACAAGAAGGCAGCGGAGGAGAGCAAGCAGCTGGAGGACGAGCTGGTGGC   270
   L Q K K L K G T E D E L D K Y S E S L K D A Q E K L E L A D
TCTGCAAAAAGAAGCTGAAGGGCACTGAGGATGAGCTGGACAAATACTCCGAGTCCCTTAAAGATGCACAGGAAAAGTTGGAACCTGGCTGA   360
   K K A T D A E S E V A S L N R R I Q L V E E E L D R A Q E R
CAAAAAGGCCACAGATGCTGAGAGTGAAGTAGCTTCCCTGAACAGACGCATCCAACCTGGTTGAGGAAGAGTTGGATCGGGCTCAGGAGCG   450
   L A T A L Q K L E R A E K A A D E S E R G M K V I E N R A Q
CTTGGCTACTGCCCTGCAGAAGCTGGAGGAGGCTGAGAAGGCTGCAGATGAGAGTGAAAGAGGAATGAAGGTCATTGAAAAATAGAGCCCA   540

```

GenBank M32441

LOCUS CHKATRO 1194 bp mRNA linear VRT 28-APR-1993
DEFINITION Chicken fast-twitch alpha-tropomyosin mRNA, complete cds.
ACCESSION M32441
VERSION M32441.1 GI:211225
KEYWORDS alpha-tropomyosin.
SOURCE Gallus gallus (chicken)
ORGANISM [Gallus gallus](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria;
Aves; Neognathae; Galliformes; Phasianidae; Phasianinae; Gallus.
REFERENCE 1 (bases 1 to 1194)
AUTHORS Gooding,C., Reinach,F.C. and Macleod,A.R.
TITLE Complete nucleotide sequence of the fast-twitch isoform of chicken skeletal muscle alpha-tropomyosin
JOURNAL Nucleic Acids Res. 15 (19), 8105 (1987)
PUBMED [3671073](#)
COMMENT Original source text: Chicken skeletal muscle, cDNA to mRNA.
FEATURES
Location/Qualifiers
source 1..1194
/organism="Gallus gallus"
/mol_type="mRNA"
/db_xref="taxon:9031"
mRNA <1..1194
/product="alpha-tropomyosin mRNA"
CDS 137..991
/note="alpha-tropomyosin"
/codon_start=1
/protein_id="AAA48610.1"
/db_xref="GI:211226"
/translation="MDAIKKKMQMLKLDKENALDRAEQAEADKKAEEERSKQLEDELV
ALQKRLKGTDELDKYSESLKDAQEKLELADKKATDAESEVASLNRRIQLVEEELDRA
QERLATALQKLEEAekaADESERGMKVIENRAQKDEEKMEIQEIQKKAKHIAEEADR
KYEEAARKLVIIIEGDLERAEEERAELSESKCAELEELKTVTNMLKSLEAQAEKYSQKE
DKYEEEIKVLTDKLKEAETRAEFAERSVTKLEKSIDDLEDELYAQKLKYKAISEELDH
ALNDMTSI"
ORIGIN
1 aattccgccc tcccgcactc gttggcccca gcgctccccg gggcccgggc tcggattcgt
61 atcggggctc tccgcccgtt tctgtctctg attcggcctg tacttctcgc gggaacggcc
121 ctaaccacc gccgccatgg atgccatcaa gaaaaagatg cagatgctga aactggacaa
181 ggagaatgcc ttggacagag ccgagcaagc cgaagcggac aagaaggcag cggaggagag
241 gagcaagcag cttgaggcag agctggtggc tctgcaaaaag aagctgaagg gcactgagga
301 tgagctggac aaatactccg agtcccttaa agatgcacag gaaaagtgg aactggctga
361 caaaaaggcc acagatgctg agagtgaagt agcttccctg aacagacgca tccaactggt
421 tgaggaagag ttggtatcgg ctcaggagcg cttgggtact gccctgcaga agctggagga
481 ggctgagaag gctgcagatg agagtgaag agaatgaa gtcattgaaa atagagccca
541 gaagatgaa gagaagatgg aaatccaaga gatccagctt aaagaagcta agcattgctc
601 tgaagaggct gaccgcaagt atgaagagcc ggctcgtgta ctcgtgatca ttgagggtga

taxonomy

publication

CDS annotation

protein or gene name

CDS translation
(automated)

Nucleic acid sequence
(mRNA or DNA)

ENA M32441

```
ID M32441; SV 1; linear; mRNA; STD; VRT; 1194 BP.
XX
AC M32441;
XX
DT 26-JUL-1991 (Rel. 28, Created)
DT 27-JUN-2018 (Rel. 137, Last updated, Version 3)
XX
DE Chicken fast-twitch alpha-tropomyosin mRNA, complete cds.
XX
KW alpha-tropomyosin.
XX
OS Gallus gallus (chicken)
OC Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC Archelosauria; Archosauria; Dinosauria; Saurischia; Theropoda;
OC Coelurosauria; Aves; Neognathae; Galloanserae; Galliformes; Phasianidae;
OC Phasianinae; Gallus.
XX
RN [1]
RP 1-1194
RX DOI; 10.1093/nar/15.19.8105.
RX PUBMED; 3671073.
RA Gooding C., Reinach F.C., Macleod A.R.;
RT "Complete nucleotide sequence of the fast-twitch isoform of chicken
RT skeletal muscle alpha-tropomyosin";
RL Nucleic Acids Res. 15(19):8105-8105(1987).
XX
DR MD5; ec744967e13e772fd6b1bc2b3ea14588.
XX
CC Original source text: Chicken skeletal muscle, cDNA to mRNA.
FH Key Location/Qualifiers
FH
FT source 1..1194
FT /organism="Gallus gallus"
FT /mol_type="mRNA"
FT /db_xref="taxon:9031"
FT mRNA <1..1194
FT /product="alpha-tropomyosin mRNA"
FT CDS 137..991
FT /codon_start=1
FT /note="alpha-tropomyosin"
FT /db_xref="GOA:P04268"
FT /db_xref="InterPro:IPR000533"
FT /db_xref="PDB:1IC2"
FT /db_xref="PDB:3MTU"
FT /db_xref="PDB:3MUD"
FT /db_xref="PDB:3U1A"
FT /db_xref="PDB:3U1C"
FT /db_xref="UniProtKB/Swiss-Prot:P04268"
FT /protein_id="AAA48610.1"
FT /translation="MDAIKKKMQLKLDKENALDRAEQAEADKKAAEERSKQLEDELVA
FT LQKKLKGTEDELDKYSESLKDAQEKLELADKKATDAESEVASLNRRIQLVEEELDRAQE
FT RLATALQKLEEAKEAADESERGMKVIENRAQKDEEKMEIQEIQLKEAKHIAEEADRKYE
FT EAARKLVIIEGDLERAEEERAELSEKCAELEELKTVTNNLKSLEAQAEKYSQKEDKYE
FT EEIKVLTDKLEAETRAEFAERSVTKLEKSIDDLEDELYAQKLYKAISELDHALNDM
FT TSI"
XX
X1
```

taxonomy

publication

protein or gene name

cross-references added by ENA

CDS translation
(automated)

<https://www.ebi.ac.uk/ena/browser/api/embl/M32441>

```
3Q Sequence 1194 BP; 347 A; 261 C; 333 G; 233 T; 0 other;
aattccgcgcg tcccgcactc gttggccccc gcgctcccgc gggccgcggc tcggattcgt 60
atcggggctc tccgcgcgtt tctgctctg attcggctg tactctcgc gggaaacggcc 120
ctaaccacc cgcgcattg atgccatcaa gaaaaagatg cagatgctga aactggacaa 180
ggagaatgcc ttggacagac ccgagcaagc cgaagcggac aagaaggcag cggaggagag 240
gagcaagcag ctggaggacg agctggtggc tctgcaaaag aagctgaagg gcactgagga 300
tgagctggac aaatactcgc agtcccctaa agatgcacag gaaaagtgg aactgctga 360
```



ENA/GenBank/DDBJ

“Beyond limited editorial control and some internal integrity checks (for example, proper use of INSD formats and translation of coding regions specified in CDS entries are verified), **the quality and accuracy of the record are the responsibility of the submitting author, not of the database.**”

<http://www.insdc.org/policy>

ENA/GenBank/DDBJ

- many scientists assume that GenBank annotation is kept up to date, and they are surprised to hear that it is not
- the **annotation has remained static**: a gene labeled 'hypothetical protein' a few years ago might now have a known function.
- **erroneous and inconsistent naming of genes.**
- scientists should fix errors that they find. But this would quickly destroy the **archival function of GenBank**, as original entries would be erased over time.

From
ENA/GenBank/DDBJ
to
protein sequence databases

What is transferred from nucleotide sequence databases such as UniProtKB ?

- The protein sequence (translated CDS)
- Publication provided by the submitting author
- Gene and protein names
- EC number (enzyme classification)
- Origin of the sequence (tissues)
- Taxonomy

What is transferred from nucleotide sequence databases such as UniProtKB ?

- The protein sequence (translated CDS)
- Publication provided by the submitting author
- Gene and protein names
- EC number (enzyme classification)
- Origin of the sequence (tissues)
- Taxonomy

What is transferred to UniProtKB ('imported'):

DR EMBL; DQ339047; ABC68418.1; -; mRNA.

FT source 1..1397

FT /organism="Rattus norvegicus"

FT /strain="Sprague-Dawley"

FT /mol_type="mRNA"

FT /sex="female"

FT /tissue_type="ovary"

FT /db_xref="taxon:10116"

FT CDS 70..1329

FT /codon_start=1

FT /product="testis derived transcript"

FT /note="TES"

FT /db_xref="GOA:Q2LAP6"

Publications for Q2LAP6

No title available.

Seo Y.M., Jang S.J., Chun S.Y.

Cited for NUCLEOTIDE SEQUENCE [MRNA]

Strain Sprague-Dawley

Tissue Ovary

Categories Sequences


Source UniProtKB reviewed (Swiss-Prot)



Protein namesⁱ

Recommended name | Testin

Gene namesⁱ

 Name | Tes

<https://www.ncbi.nlm.nih.gov/nucleotide/DQ339047>

<https://www.uniprot.org/uniprotkb/Q2LAP6/entry>

What is transferred from nucleotide sequence databases such as UniProtKB ?

- The protein sequence (translated CDS)
- Publication provided by the submitting author
- Gene and protein names
- **EC number (enzyme classification)**
- Origin of the sequence (tissues)
- Taxonomy

ref NP_884595.1	4-aminobutyrate aminotransferase [Bordetella pa...	186	1e-45
ref NP_926795.1	4-aminobutyrate aminotransferase [Gloeobacter v...	183	8e-45
ref NP_126958.1	4-aminobutyrate qui se dilate aminotransferase ...	182	1e-44
gb AAH43680.2	Alanine glyoxylate aminotransferase 2 like 1 [Mus...	182	2e-44
dhl RAC33993.1	unnamed protein product [Mus musculus] >gil35193...	182	2e-44

```

FT      CDS                complement(45959..47332)
FT                                     /db_xref="SPTREMBL:Q9UZ71"
FT                                     /note="PAB2386"
FT                                     /transl_table=11
FT                                     /product="4-AMINOBTYRATE qui se dilate AMINOTRANSFERASE
FT                                     (EC 2.6.1.19)"
FT                                     /protein_id="CAB50188.1"
FT                                     /translation="MDYPRIVVNPPGPKAKELIEREKRVLSTGIGVKLFPLVPKRGFGP
FT                                     FIEDVDGNVFIDFLAGAAAASSTGYSHPKLVKAVKEQVELIQHSMIGYTHSERAIRVAEK
FT                                     LVKISPIKNSKVLFGLSGSDAVDMAIKVSKFSTRRPWILAFIGAYHGQTLGATSVASFQ
FT                                     VSQKRGYSPLMPNVFVWPYPNPYRNPWGINGYEEPQELVNRVVEYLEDYVFSHVPPDE
FT                                     VAAFFAEPIQGDAGIVVPPENFFKELKLLDEHGILLVMDEVQTGIGRTGKWFASEWFE
FT                                     VKPDMIIIFGKGVASGMGLSGVIGREDIMDITSGSALLTPAANPVISAAADATLEIIIEE
FT                                     NLLKNAIEVGSFIMKRLNELKEQFDIIGDVRGKGLMIGVEIVKENGRPDPEMTGKICWR
FT                                     AFELGLLILPSYGMFGNVIRITPPLVLTKEVAEKGLEIIEKAIKDAIAGKVERKVVVTH"

```

Challenge: making the link between enzyme activity and protein sequence(s)...

Challenge: making the link between enzyme activity and protein sequence(s)...

CDS

```
complement(47269..48642)
/locus_tag="PAB2386"
/codon_start=1
/transl_table=11
/product="Pyridoxal phosphate-dependent aminotransferase"
/protein_id="CAB50188.1"
/db_xref="GOA:Q9UZ71"
/db_xref="InterPro:IPR005814"
/db_xref="InterPro:IPR015421"
/db_xref="InterPro:IPR015422"
/db_xref="InterPro:IPR015424"
/db_xref="UniProtKB/TrEMBL:Q9UZ71"
/translation="MDYPRIVVNPPGPKAKELIEREKRVLSTGIGVKLFPPLVPKRGFG
PFIEDVDGNVFDIDFLAGAAAASTGYSHPKLVKAVKEQVELIQHSMIGYTHSERAIRVA
EKLVKISPIKNSKVLFGLSGSDAVIDMAIKVSKFSTRRPWILAFIGAYHGQTLGATVA
SFQVSQKRGYSPLMPNVFWVPYPNPNRNPWINGYEEPQELVNRVVEYLEDYVFSHVV
PPDEVAFFFAEPIQGDAGIVVPPENFFKELKLLDEHGILLVMDEVQGTGIGRTGKWFA
SEWFEVKPDMIIFGKGVASGMGLSGVIGREDIMDITSGSALLTPAANPVISAAADATL
EIIIEENLLKNAIEVGSFIMKRLNELKEQFDIIGDVRGKGLMIGVEIVKENGRPDPEM
TGKICWRAFELGLILPSYGMFGNVIRITPPLVLTKEVAEKGLEIIEKAIKDAIAGKVE
RKVVTWH"
```

<https://www.ncbi.nlm.nih.gov/nucore/AJ248287.2>

Q9UZ71 · Q9UZ71_PYRAB

4-aminobutyrate aminotransferase · *Pyrococcus abyssi* (strain GE5 / Orsay) · Gene: PAB2386 · 457 amino acids · Inferred from homology

Names & Taxonomyⁱ

Protein namesⁱ

Submitted names

4-aminobutyrate aminotransferase	Imported
Pyridoxal phosphate-dependent aminotransferase	Imported

Gene namesⁱ

Ordered locus names

PAB2386	Imported
---------	----------

Organism namesⁱ

Organism | *Pyrococcus abyssi* (strain GE5 / Orsay) Imported

Taxonomic identifierⁱ | 272844 NCBI [↗](#)

Taxonomic lineageⁱ | cellular organisms > Archaea > Euryarchaeota > Thermococci > Thermococcales > Thermococcaceae > Pyrococcus > Pyrococcus abyssi

<https://www.uniprot.org/uniprotkb/Q9UZ71/entry>

What is transferred from nucleotide sequence databases such as UniProtKB ?

- The protein sequence (translated CDS)
- Publication provided by the submitting author
- Gene and protein names
- EC number (enzyme classification)
- Origin of the sequence (tissues)
- **Taxonomy**

Taxonomy biocuration: examples

Collaboration with NCBI taxonomy

UniProtKB taxonomy data is manually curated: next to manually verified [organism names](#), we provide a selection of external links, [organism strains](#) and [viral host](#) information.

Taxonomy 2,417,253 results

Taxonomy biocuration: examples

Collaboration with NCBI taxonomy

Drosophila	fruit fly <genus>	7215	fruit fly <subgenus>	32281	basidiomycete fungus <genus>	2081351
-------------------	-------------------	------	----------------------	-------	------------------------------	---------

Bacteria	prokaryotes <superkingdom>	2	<i>Insect</i> Bacteria Latreille et al. 1825 [synonym: Bacteria] <genus> [06.21.2017] renamed Bacteria Latreille et al. 1825	629395
-----------------	----------------------------	---	--	--------

o [Bacteria Latreille et al. 1825](#) Click on organism name to get more information.

- [Bacteria abnormis](#)
- [Bacteria aborigena](#)
- [Bacteria acuminatocercata](#)
- [Bacteria aetolus](#)
- [Bacteria amazonica](#)
- [Bacteria ambigua](#)
- [Bacteria annulicornis](#)
- [Bacteria apolinari](#)
- [Bacteria baculus](#)
- [Bacteria bahiensis](#)
- [Bacteria ferula](#)
- [Bacteria horni](#)
- [Bacteria ploiaria](#)
- [Bacteria versiniana](#)
- [Bacteria sp. 1 JAR-2018](#)

Stick insects (phasmes)

Venturia	fungus <genus>	5024	ichneumonid wasp <genus>	92443
-----------------	----------------	------	--------------------------	-------



Asterina	ascomycetes <genus>	859380	starfish <genus>	7593
-----------------	---------------------	--------	------------------	------



Question

For fun:

Look at this GenBank entry:

<https://www.ncbi.nlm.nih.gov/nucore/Z71230>

Where does the sequence come from ?

- Which organism ? Which isolate ?

What is the length of the protein sequence?

Taxonomy biocuration: examples

Collaboration with NCBI taxonomy

<https://www.ncbi.nlm.nih.gov/nuccore/Z71230>

```
source      1..124
            /organism="Nicotiana tabacum"
            /organelle="plastid:chloroplast"
            /mol_type="genomic DNA"
            /isolate="Cuban cahibo cigar, gift from President Fidel
            Castro"
            /db_xref="taxon:4097"
```



<https://www.pinterest.ch/abidabdellatif/castro/>

<https://www.ncbi.nlm.nih.gov/nuccore/AAFZ00000000.1>

```
source      1..29934
            /organism="whale fall metagenome"
            /mol_type="genomic DNA"
            /isolation_source="microbial mat from gray whale carcass
            in the Pacific Ocean (depth=1674 meters), Santa Cruz Basin
            (N33.30 W119.22)"
            /db_xref="taxon:412756"
            /environmental_sample
            /country="USA: Pacific Ocean, Santa Cruz Basin"
            /note="metagenomic"
```



What is transferred from nucleotide sequence databases such as UniProtKB ?

- The protein sequence (translated CDS)
- Publication provided by the submitting author
- Gene and protein names
- EC number (enzyme classification)
- Origin of the sequence (tissues)
- Taxonomy

GenBank Nucleotide sequence Annotated CDS CoDing Sequence

[CDS](#)

```

join(615..627,1194..1339,1596..1682,2294..2473,2608..2763)
/codon_start=1
/product="erythropoietin"
/protein_id="CAA26095.1"
/db_xref="GDB:119110"
/db_xref="GOA:P01588"
/db_xref="HGNC:HGNC:3415"
/db_xref="InterPro:IPR001323"
/db_xref="InterPro:IPR003013"
/db_xref="InterPro:IPR009079"
/db_xref="InterPro:IPR012351"
/db_xref="InterPro:IPR019767"
/db_xref="PDB:1BUY"
/db_xref="PDB:1CN4"
/db_xref="PDB:1EER"
/db_xref="UniProtKB/Swiss-Prot:P01588"
/translation="MGVHECPAFLWLLLSLLSLPLGLPVLGAPPRLICDSRVLQRYLL
EAKAEENITTCGAEHCSSLNENITVPTDKVNFYAWKRMEVQQQAVEVWQGLALLSEAVL
RGQALLVNSSQPWPELQLHVDKAVSGLRSL/TTLLRALGAQKEAISPDAASAAPLRTI
TADTFRKLLFRVYSNFLRGLKLLKYTGACRTGDR"
    
```

Protein sequence:
which accuracy ?

```

541 cttccccggga tgagggcccc cgggtgtggtc acccggcgcc ccaggtcgct gagggacccc
601 ggccaggcgc ggagatgggg gtgcacggtg agtactcgcg ggctggggcg tcccggccgc
661 ccgggtccct gtttgagcgg ggatttagcg ccccggtat tggccaggag gtggctgggt
721 tcaaggaccg gcgacttgtc aaggaccccc gaagggggag gggggtgggg cagcctccac
781 gtgccagcgg ggacttgggg gagtccttgg ggatggcaaa aacctgacct gtgaagggga
841 cacagtttgg gggttgaggg gaagaagggt tgggggggtc tgctgtgcca gtggagagga
901 agctgataag ctgataacct gggcgctgga gccaccactt ctctgcaga ggggaagcct
961 ctgtcacacc aggattgaag tttggccgga gaagtggatg ctggtagcct gggggtgggg
1021 tgtgcacacg gcagcaggat tgaatgaagg ccaggggagg agcacctgag tgcctgcatg
1081 gttggggaca ggaaggacga gctggggcag agacgtgggg atgaaggaag ctgtccttcc
1141 acagccacc cttccctccc ccgcctgact ctacgctgg ctatctgttc tagaatgtcc
1201 tggcctggctg tggcttctcc tgtccctgct gtcgctccct ctgggctccc cagtcctggg
1261 cggcccacca cgcctcatct gtgacagccg agtcctgcag aggtacctct tggaggccaa
1321 ggaggccgag aatatcacgg tgagaccctt tccccagcac attccacaga actcacgctc
1381 agggcttcag ggaactcctc ccagatccag gaacctggca ctgggtttgg ggtggagttg
1441 ggaagctaga cactgcccc ctacataaga ataagtctgg tggccocaaa ccatacctgg
1501 aaactaggca aggagcaaa ccagcagatc ctacgctgt ggccagggcc agagccttca
1561 gggacccttg actccccggg ctgtgtgcat ttcagacggg ctgtgtgtaa cactgcagct
1621 tgaatgagaa tatcactgtc ccagacacca aagttaatt ctatgctgg aagaggtgg
1681 aggtgagttc cttttttttt ttttttccct tcttttgag aactcattt gcgagcctga
1741 ttttggatga aagggagaat gatcgagga aaggtaaat ggagcagcag agatgaggct
1801 gcctggggcg agaggctcac gtctataatc ccaggctgag atggccgaga tgggagaatt
1861 gcttgagccc tggagtttca gaccaacctc ggcagcatag tgagatcccc catctctaca
1921 aacatttaa aaaattagtc aggtgaagtg gtgcatggtg gtagtcccag atatttggaa
1981 ggcctgaggg ggagatcgc ttgagccacg gaatttgag ctgtgtgac cctgtgatcc
2041 accactgcac tccagcctca gtgacagagt gaggccctgt ctcaaaaaa aaaagaaaa
2101 agaaaaataa tgagggctgt atggaatacg ttcattatc attcactcac tcaactctc
2161 attcattcat tcattcattc aacaagtctt attgcatac ttctgtttgc tcagcttgg
2221 gcttggggct gctgaggggc aggagggaga gggtgacat cctcagctga ctccagagt
2281 ccactcctg taggtcgggc agcagggcct agaagtcctg cagggcctgg cctgtctgtc
2341 ggaagctgtc ctgctggggc aggcctgtt ggtcaactct tcccagccgt gggagccct
2401 gcagctgcat gtggataaag ccgtcagtg ccttcgacg ctccaccact tgcctcgggc
2461 tctggggagc caggtgagta ggagcggaca cttctgctt cctttctgt aagaagggga
2521 gaaggtctct gctaaggagt acaggaactg tccgtattcc tcccctttct gtggcactgc
2581 agcagcctcc tgttttctcc ttggcagaag gaagccatct cccctccaga tggggcctca
2641 gctgctccac tccgaacaat cactgctgac actttccgca aactcttccg agtctactcc
2701 aatttctccc ggggaaagct gaagctgtac acaggggagg cctgcaggac aggggacaga
2761 tgaaccagtg tgtccactg ggcatacca ccactcctc caccacact gcttgtgcca
2821 caccctccc ccactcctt gaaccccgtc gaggggctct cagctcagc ccagcctgtc
2881 ccattggcac tccagtggca ccaatgacat ctacggggcc agaggaactg tccagagagc
2941 aactctgaga tctaaggatg tcacagggcc aacttgaggg ccagagcag gaagcattca
3001 gagagcagc ttaaactcag ggacagacc atgctgggaa gacgctgag ctcaactcgc
3061 acctgcaaa attgatcca ggacacgctt tggaggcgat ttacctgtt tcgcactcac
3121 catcagggac aggatgacct ggagaactta ggtggcaagc tgtgacttct ccaggtctca
3181 cgggcatggg cactcccttg gtggcaagag ccccctgac accggggtgg tgggaacct
3241 gaagacagga tggggctgg cctctggctc tcatggggtc caactttgt gtattcttca
3301 acctcattga caagaactga aaccaccaat atgactctg gcttttctgt tttctgggaa
3361 cctccaaatc ccctggctct gtcccactcc tggcagca
    
```

GenBank Nucleotide sequence

GenBank ▾

Human gene for erythropoietin

GenBank: X02158.1

[FASTA](#) [Graphics](#)

[Go to:](#)

LOCUS X02158 3398 bp DNA linear PRI 14-NOV-2006

DEFINITION Human gene for erythropoietin.

ACCESSION X02158

VERSION X02158.1

KEYWORDS erythropoietin; glycoprotein hormone; hormone; signal peptide.

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 3398)

AUTHORS Jacobs,K., Shoemaker,C., Rudersdorf,R., Neill,S.D., Kaufman,R.J.,
Mufson,A., Seehra,J., Jones,S.S., Hewick,R., Fritsch,E.F.,
Kawakita,M., Shimizu,T. and Miyake,T.

TITLE Isolation and characterization of genomic and cDNA clones of human
erythropoietin

JOURNAL Nature 313 (6005), 806-810 (1985)

PUBMED 3036566

COMMENT Data kindly reviewed (24-FEB-1986) by K. Jacobs.

```

1 agcttctggg cttccagacc cagctacttt gcggaactca gcaaccagg catctctgag
61 tctccgccca agaccgggat gccccccagg aggtgtccgg gaccgccacc tttccccagat
121 agcagctccg ccagtcocaa ggggtgcgca ccggtgtcac tcccctcccg cgaccgccgg
181 cccgggagca gcccccatga cccacacgca cgtctgcagc agccccgtea gccccggagc
241 ctcaaccacg gcgtcctgcc cctgctctga ccccgggtgg cccctacccc tggcgacccc
301 tcacgcacac agcctctccc ccacccccac ccgcgcacgc acacatgcag ataacagccc
361 cgacccccgg ccagagccgc agagtcccct ggcccccccg gcccctcgtc gcgctgcgcc
421 gcaccgcgct gtctctcccg agccggaccg gggccaccgc gcccctctct cctccgacacc
481 gcgccccctg gacagccgcc ctctcctcca gcccctgggg gctggcccctg caaccgcgag
541 cttccccgga tgagggcccc cgggtgtggtc acccggcgcc ccaggtoctg gaggggacccc
601 gtcacggcgc gggagtgggg gtgcacgggt agtactcgcg ccctggggcc tcccgcgccg
661 cccgggtccc gtttgagcgg ggtattagcg ccccggctat tggccaggag gtggctgggt
721 tcaaggaccg gcgacttgtc aaggaccccc gaagggggag ggggggtggg cagcctccac
781 gtgccagcgg ggacttgggg gagtccctgg gtagtggcaa aacctgacct gtgaagggga
841 cacagtctgg gggttgaggg gaagaagggt tgggggggtc tgctgtgcca gtcgagagga
901 agctgataag ctgataacct gggcgcctga gccaccactt atctgccaga ggggaagcct
961 ctgtcacacc aggattgaag tttggcccga gaagtggatg ctggtgacct ggggggtggg
1021 tgtgcacacg gcagcaggat tgaatgaagg ccaggggagg agcacctgag tgctgtcatg
1081 gttggggaca ggaaggaaga gctggggcag agacgtgggg atgaaggaaag cttcccctcc
1141 acagccaccc tctctcccct cccgctgact ctccagctgg ctatctgttc tagaatgtcc
1201 tgccctggctg tggcttctcc tgtcccctgt gtccctcctc ctgggcccctc cagtcctggg
1261 cgcgccacca cgcctcatct gtgacagccg agtccctgca agttaacct tggaggccaa
1321 ggaggccgag aatatcacgg tgagaccctc tcccacgca acatcacgctc
1381 agggctccag ggaactcctc ccagatccag gaacctggca cttggtttgg ggtggagttg
1441 ggaagctaga cactgcccc ctcacataaga ataagctctg tggccccaaa ccataacctg
1501 aaactaggca aggagcaaa ccagcagatc ctaccgctgt ggcaccgggc agagccttca
1561 gggacacctg actccccggg ctgtgtgcat ttcagacggg ctgtgtgtaa cactgcagct
1621 tgaatgagaa taccactgtc ccagacacca aagttaattt ctatgctctg aagaggatgg
1681 aggtgagttc cttttttttt ttttttccct tcttttggag aactcaattt gcgagcctga
1741 ttttggatga aagggagaat gatcgaggga aaggtaaaa gtagcagcag agatgagggt
1801 gctggggcgc agagggctcc gctctaatc ccaggctgag atggccgaga tgggagaatt
1861 gcttgagccc tggagtcca gaccacacta ggcagcatag tgagatcccc catctctaca
1921 aacattttaa aaaattagtc aggtgaaagt gtgcatggtg gtagtcccag atatttggaa
1981 ggtgaggggc gggagatcgc ttgagcccag gaatttgagg ctccagttgag ctgtgatcac
2041 accactgcac tccagctcca gtgacagagt gaggccctgt ctcacaaaaa aaaagaaaa
2101 agaaaaataa tgagggtctg atggaatacg ttcattatcc atctactcac tcactcactc
2161 atctattcat tcaattcatt caacaagctt attgcatacc tctgttttgc tcactctggt
2221 gcttggggct gctgaggggc aggagggaga ggggtgacac cctcagctga ctcccagagt
2281 ccaactccctg taggtcggcc agcagggcgt agaagtctgg ccaggccctg cctctgcttc
2341 ggaagctgtc ctgcccggcc aggcctctgt ggtcaactct tcccagcctg gggagcccct
2401 gcagctgcat gtggataaag ccgtcagttg ccttcgcagc ctcaaccactc tgcttcgggg
2461 tctggggacc caggtgagta ggagccgaca ctctgctgtt cctcttctgt aagaagggga
2521 gaaggtctct gctaaaggat acaggaactg tccgtattcc tctcccctct gtgacctgct
2581 agcagacctc tgttttctcc ttggcagaag gaagccatct cccctccaga tgcggcctca
2641 gctgtccacc tccgaacaat caactgctac actttcccga aactctcccg agtctactcc
2701 aatttctccc ggggaaaagt gaagctgtac acaggggagg cctgcaggac aggggacaga
2761 tgaccaggtg tgtcccctg ggcataacca ccaactcctc ccccaacatt gctgtgcca
2821 cccctcccc cgcactcctc gaaccccctc gagggtctct cagctcagcg ccagcctgtc
2881 ccatggacac tccagtcca ccaatgacat ctccagggcc agaggaactg tccagagagc
2941 aactctgaga tctaaggatg tccagggccc aacttgaggg cccagagcag gaagcattca
3001 gagagcagct ttaaaactca ggacagaccc atgctgggaa gacgcctgag ctcaactcgc
3061 accctgcaaa attgatgcca ggacacgctt tggaggcgat ttacctgttt tccgacctac
3121 catcagggac aggatgacct ggaagaacta ggtggcaagc tgtgactctc ccaggtctca
3181 cgggcatggg cactcccctg gtggcaagag cccccttgac accgggggtg tgggaacctca
3241 gaagacagga tggggctcgg cctctggctc tcaatgggtc caactttttg gtattcttca
3301 acctcattga caagaactga aaccaccaat atgactcttg gcttttctgt tttctgggaa
3361 cctccaaatc ccttggtctc gtcccactcc tggcagca
    
```

<https://www.ncbi.nlm.nih.gov/nucleotide/x02158>



GenBank Nucleotide sequence Annotated CDS

*gene prediction or
experimentally proven (cDNA)*

[CDS](#)

```

join(615..627,1194..1339,1596..1682,2294..2473,2608..2763)
/codon_start=1
/product="erythropoietin"
/protein_id="CAA26095.1"
/db_xref="GDB:119110"
/db_xref="GOA:P01588"
/db_xref="HGNC:HGNC:3415"
/db_xref="InterPro:IPR001323"
/db_xref="InterPro:IPR003013"
/db_xref="InterPro:IPR009079"
/db_xref="InterPro:IPR012351"
/db_xref="InterPro:IPR019767"
/db_xref="PDB:1BUY"
/db_xref="PDB:1CN4"
/db_xref="PDB:1EER"

```

```

/translation="MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLQRYLL
EAKEAENITGCAEHCSLNENITVPDPKVNIFYAWKRMEVQQAVEVWQGLALLSEAVL
RQQALLVNSSQPWEPLQLHVDKAVSGLRSL/TTLLRALGAQKEAISPPDAASAAPLRTI
TADTFRKLLFRVYSNFLRGLKLLKYTGACRTGDR"

```

Protein sequence

```

541 cttccccggga tgagggcccc cgggtgtggtc acccggcgcc ccaggtcgct gagggacccc
601 ggccaggcgc ggagatgggg gtgcacggtg agtactcgcg ggctggggcg tcccggccgc
661 ccgggtccct gtttgagcgg ggatttagcg ccccggtat tggccaggag gtggctgggt
721 tcaaggaccg gcgacttgtc aaggaccccc gaagggggag gggggtgggg cagcctccac
781 gtgccagcgg ggacttgggg gagtccttgg ggatggcaaa aacctgacct gtgaagggga
841 cacagtttgg gggttgaggg gaagaagggt tggggggttc tgctgtgcca gtgggaagga
901 agctgataag ctgataacct gggcgctgga gccaccactt ctctgcaga ggggaagcct
961 ctgtcacacc aggattgaag tttggccgga gaagtggatg ctggtagcct gggggtgggg
1021 tgtgcacacg gcagcaggat tgaatgaagg ccaggggagg agcacctgag tgcctgcatg
1081 gttggggaca ggaaggacga gctggggcag agacgtgggg atgaaggaag ctgtccttcc
1141 acagccaccc ttctccctcc ccgcctgact ctacgctgg ctatctgttc tagaatgtcc
1201 tgcctggctg tggcttctcc tgtcctgct gtctctcct ctgggctcc cagtctggg
1261 cgcccaacca cgctcatct gtgacagccg agtcctgcag agtaactct tggaggccaa
1321 ggagggccag aatatcacg tgagaccct tccccagcac attccacaga actcacgctc
1381 agggcttcag gaaactcct ccagatccag gaacctggca ctgggtttgg ggtggagtgt
1441 ggaagctaga cactgcccc ctacataaga ataagtctgg tggcccaaaa ccataactgg
1501 aactaggca aggagcaaa ccagcagatc ctacgctgt ggccagggcc agagccttca
1561 gggacccttg actccccggg ctgtgtgcat ttcagacggg ctgtgtgtaa cactgagct
1621 tgaatgaaa tatcactgtc ccagacacca aagttaatt ctatgctgg aagagatga
1681 aggtgagttc ctttttttt ttttttctt tcttttgag aactcattt gcgagcctga
1741 ttttggatga aagggagaat gatcgagga aaggtaaat ggagcagcag agatgaggct
1801 gcctggggcg agaggctcac gtctataat ccaggctgag atggccgaga tgggagaatt
1861 gcttgagccc tggagtttca gaccaacct ggcagcatag tgagatcccc catctctaca
1921 aacatttaa aaaattagc aggtgaagtg gtgcattggt gtagtcccag atatttggaa
1981 ggtcgaggcg ggagatcgc ttgagccag gaatttggg ctgctgacac ctgtgatcac
2041 accactgac tccagcctca gtgacagagt gaggccctgt ctcaaaaaa aaaagaaaa
2101 agaaaaataa tgagggctgt atggaatac ttcattatc attcactcac tcaactactc
2161 attcattcat tcattcattc aacaagtctt attgcatac ttctgtttg tcagcttgg
2221 gcttggggct gctgaggggc aggagggaga gggtgacat cctcagctga ctccagagt
2281 ccactcctg taggtcgggc agcagggcgt agaagtctgg cagggcctgg cctgctgtc
2341 ggaagctgtc ctgctggggc aggcctgtt ggtcaactct tcccagccgt gggagccct
2401 gcagctgcat gtggataaag ccgtcagtg ccttcgcatg ctccaccact tgcctcgggg
2461 tctggggagc caggtgagta ggagcggaca cttctgctt cctttctgt aagaagggga
2521 gagggtctt gctaaggat acaggaactg tccgtattcc tccctttct gtggcactgc
2581 agcagctcc tgttttctcc ttggcagaag gaagccatct cccctccaga tggggcctca
2641 gctgctccac tccgaacaat cactgctgac actttccgca aactcttccg agtctactcc
2701 aatttctccc ggggaaagct gaagctgtac acaggggagg cctgcaggac aggggacaga
2761 tgaaccagtg tgtccactg ggcatacca ccactcctc caccacact gcttgtgcca
2821 caccctccc ccgcaactct gaacccctc gaggggctc tagctcagc ccagcctgtc
2881 ccattggcac tccagtggca ccaatgacat ctacggggcc agaggaactg tccagagagc
2941 aactctgaga tctaaggatg tcacagggcc aacttgaggg ccagagcag gaagcattca
3001 gagagcagc ttaaactcag ggacagacc atgctgggaa gacgctgag ctcaactcgc
3061 acctgcaaa attgatcca ggacacgct tggaggcgat ttacctgtt tcgacctac
3121 catcagggac aggatgacct ggagaactta ggtggcaagc tgtgacttct ccaggtctca
3181 cgggcatggg cactccttgg gtggcaagag ccccttgac accggggtgg tgggaacct
3241 gaagacagga tggggctgg cctctggctc tcatggggtc caactttgt gtattcttca
3301 acctcattga caagaactga aaccaccaat atgactctg gcttttctgt tttctgggaa
3361 cctccaaatc cctcggctct gtcccactcc tggcagca

```


Question

Haemophilus influenzae genome in GenBank

<https://www.ncbi.nlm.nih.gov/nuccore/JMQP01000002>

- (1) What are the different CDS 'inferences' ?
- (2) Look at the information available for the gene ProS
 - Look at the Enzyme classification number (EC number)
- (3) Follow the link to 'Protein' (column on the right; Related information)
How many proteins are predicted to be encoded by this genome ?
- (4) How many proteins in UniProtKB? (query with JMQP01000002)
- (5) Look at the entry corresponding to the gene ProS in UniProtKB (A0A0D0IH15)

GenBank JMQP01000002

Haemophilus influenzae strain 1209 contig000002, whole genome shotgun sequence

GenBank: JMQP01000002.1

[FASTA](#) [Graphics](#)

LOCUS JMQP01000002 1799554 bp DNA linear BCT 20-FEB-2015
DEFINITION Haemophilus influenzae strain 1209 contig000002, whole genome
shotgun sequence.

```
/inference="ab initio prediction:Prodigal:2.60"  
/inference="similar to AA sequence:UniProtKB:P0A8H6"
```

```
/inference="ab initio prediction:Prodigal:2.60"  
/inference="protein motif:Pfam:PF10675.3"
```

```
/inference="ab initio prediction:Prodigal:2.60"  
/inference="protein motif:Cdd:COG5567"
```

Gene prediction 'validated' by similarity (homology)

Gene prediction 'validated' with the presence of a known protein domain.

This type of information is not mandatory for the submitters (not always present)...

Gene ProS

Gene name

CDS



8186..9904

/gene="proS"

EC number



/locus_tag="NTHI1209_00160"

/EC_number="6.1.1.15"

Protein name



/inference="ab initio prediction:Prodigal:2.60"

/inference="similar to AA sequence:UniProtKB:P43830"

/codon_start=1

/transl_table=11

/product="Proline--tRNA ligase"

/protein_id="KIS34560.1"

/translation="MRTSQYLFSTLKETPNDAQVVS HQMLLRAGMIRPMASGLYNWLP

TGIRVLKKVEKVVREEMNKGAIEVLMPPVQPAELWEE SGRWDQYGPPELLRFEDRGNR

NFVLGPTHEEVITDLVRREVSSYKQLPLNLYQIQTKFRDEVRPRFGVMRSREFIMKDA

YSFHTTQESLQATYDVMYQVYSNIFNRLGLDFRAVQADTGSIGGSASHEFQVVLASSGE

DDVVFSTESDFAANIELAEAI AIGERQAPTAEMCLVDTPNAKTIAELVEQFNLP IEKT

VKTLIVKGADENQPLVALI IRGDHELNEIKAQKHPLVADPLEFADETEIKAKIGAGVG

SLGFPVNLNIPAIIDRTVALMSDFSCGANIDGKHYFNVNWERDVAMPEVFDLRNVVEGD

PSPDGKGT LQIKRGIEVGHIFQLGKKYSEAMKATVQGEDGKPLVMTMGCYIGVTRRVV

ASAIEQHHD ERGI IWP SDEIAPFTVAIVPMNMHKSEAVQKYAEELYRTLQSQGVDVIF

DDRKERPGVMFADMELIGVPHMVVIG EKNLDNGEIEYKNRRTGEKEMISKDKLLSVLN

EKLGNL"

The corresponding protein in UniProtKB: <https://www.uniprot.org/uniprotkb/A0A0D0IH15/entry>

A0A0D0IH15 · A0A0D0IH15_HAEIF

Proline--tRNA ligase · [Haemophilus influenzae](#) · [EC:6.1.1.15](#) · [Gene: proS](#) · 572 amino acids · Inferred from homology · **Annotation score:** [2/5](#)

[Entry](#) [Feature viewer](#) [Publications](#) [External links](#) [History](#)

[BLAST](#) [Align](#) [Download](#) [Add](#) [Add a publication](#) [Entry feedback](#)

Functionⁱ

Catalyzes the attachment of proline to tRNA(Pro) in a two-step reaction: proline is first activated by ATP to form Pro-AMP and then transferred to the acceptor end of tRNA(Pro). As ProRS can inadvertently accommodate and process non-cognate amino acids such as alanine and cysteine, to avoid such errors it has two additional distinct editing activities against alanine. One activity is designated as 'pretransfer' editing and involves the tRNA(Pro)-independent hydrolysis of activated Ala-AMP. The other activity is designated 'posttransfer' editing and involves deacylation of mischarged Ala-tRNA(Pro). The misacylated Cys-tRNA(Pro) is not edited by ProRS. [1 Automatic Annotation](#)

Caution

The sequence shown here is derived from an EMBL/GenBank/DBJ whole genome shotgun (WGS) entry which is preliminary data. [Imported](#)

Catalytic Activity

ATP + L-proline + tRNA(Pro) = AMP + diphosphate + L-prolyl-tRNA(Pro) [2 Automatic Annotations](#)

EC:6.1.1.15 ([UniProtKB](#) | [ENZYME](#) | [Rhea](#))

Source: [Rhea 14305](#)

[^ Hide Rhea reaction](#)

Sequence databases

[RefSeq](#) | [WP_005665064.1](#) [NZ_UUEWZ01000029.1](#)

SEQUENCE	PROTEIN	MOLECULE TYPE	STATUS
JMQP01000002 (EMBL GenBank DDBJ)	KIS34560.1 (EMBL GenBank DDBJ)	Genomic DNA	

The corresponding protein in UniProtKB:

<https://www.uniprot.org/uniprotkb/A0A0D0IH15/entry>

Names & Taxonomyⁱ

Protein namesⁱ

Recommended name	Proline--tRNA ligase 1 Automatic Annotation	
EC number	EC:6.1.1.15 1 Automatic Annotation	EC: validated by HAMAP rules
<p>Automatic assertion according to rules (Automatically inferred from sequence model)ⁱ UniRule HAMAP-Rule: MF_01569</p> <p>(UniProtKB ENZYME ↗ Rhea ↗)</p>		
Alternative names	Prolyl-tRNA synthetase 1 Automatic Annotation (ProRS 1 Automatic Annotation)	

Gene namesⁱ

Name	proS 1 Automatic Annotation Imported
ORF names	NTHI1209_00160 Imported

Number of entries @ NCBI protein

Items: 1 to 20 of **2037**

<< First < Prev Page **1** of 102 Next >

[Ribosomal RNA large subunit methyltransferase L \[Haemophilus influenzae\]](#)

1. 711 aa protein

Accession: KIS36588.1 GI: 757821690

[BioProject](#) [Nucleotide](#) [Taxonomy](#)

[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

[Membrane-bound lytic murein transglycosylase A precursor \[Haemophilus influenzae\]](#)

2. 369 aa protein

Accession: KIS36587.1 GI: 757821689

[BioProject](#) [Nucleotide](#) [Taxonomy](#)

[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

[Sulfur carrier protein ThiS adenylyltransferase \[Haemophilus influenzae\]](#)

3. 261 aa protein

Accession: KIS36586.1 GI: 757821688

[BioProject](#) [Nucleotide](#) [Taxonomy](#)

[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

[hypothetical protein NTH1209_02245 \[Haemophilus influenzae\]](#)

4. 43 aa protein

Accession: KIS36585.1 GI: 757821687

[BioProject](#) [Nucleotide](#) [Taxonomy](#)

[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

Number of entries @ UniProtKB

UniProtKB 2,016 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
<input type="checkbox"/> A0A0D0IE41	A0A0D0IE41_HAEIF	dITP/XTP pyrophosphatase[...]	NTHI1209_02050	Haemophilus influenzae	195 AA
<input type="checkbox"/> A0A158SZC3	A0A158SZC3_HAEIF	Ribosomal RNA large subunit methyltransferase J[...]	rlmJ, NTHI1209_01858	Haemophilus influenzae	299 AA
<input type="checkbox"/> A0A158SVF9	A0A158SVF9_HAEIF	Coenzyme A biosynthesis bifunctional protein CoaBC[...]	coaBC, NTHI1209_00456	Haemophilus influenzae	400 AA
<input type="checkbox"/> A0A158T077	A0A158T077_HAEIF	Na(+)-translocating NADH-quinone reductase subunit C[...]	nqrC, NTHI1209_02176	Haemophilus influenzae	257 AA
<input type="checkbox"/> A0A158T084	A0A158T084_HAEIF	Phosphatidylserine decarboxylase proenzyme[...]	psd, NTHI1209_02183	Haemophilus influenzae	290 AA
<input type="checkbox"/> A0A158SZX4	A0A158SZX4_HAEIF	Queuine tRNA-ribosyltransferase[...]	tgt, NTHI1209_02068	Haemophilus influenzae	396 AA
<input type="checkbox"/> A0A158T071	A0A158T071_HAEIF	tRNA-specific 2-thiouridylase MnmA[...]	mnmA, NTHI1209_02170	Haemophilus influenzae	413 AA
<input type="checkbox"/> A0A0D0ILW3	A0A0D0ILW3_HAEIF	CTP synthase[...]	pyrG, NTHI1209_00584	Haemophilus influenzae	545 AA

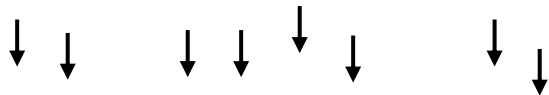
100 % identical protein sequences are merged.
Do not hesitate to contact us !

The life of a protein sequence ...

mRNA, genes, genomes, ...

```

1 agttcttggg cttccagacc cagctacttt ggggaactca gaaacccagg catctctgag
61 tctccgcca agaccgggat gcccccaggg aggtctccgg gagcccaagc ttccocagat
121 aqagctccg ccagtcocaa ggttgcccaa ccgctcgcac tccctcccg ccaccaggg
211 cctgggca gcccccata cccacagca cctctgcagc agcccctca gcccccagg
311 cctgggca gcccccata cccacagca cctctgcagc agcccctca gcccccagg
361 cgaccccagg ccagctccgc agagtccttg ggcaccccg gccctctcgt cgtctgcgac
421 gcaaccgctt gctctccgg agccggaccg gggccaccgc gccctctctg ctcocagacc
    
```



ENA, GenBank, DDBJ

Nucleic acid databases (INSDC)

...if the submitters provide an **annotated CoDing Sequence (CDS)** (gene prediction or experimental)

no CDS
Ensembl, RefSeq
gene prediction

4 %

95 %

Protein sequence databases

```

MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVSASSPSLAVASQSDSKQRR
LIVDPKCGSVSNQQRDIKANSI SMGI YVCETETKIMGNDI GEPQQQI SLSSGRTDLK
LLECSIANLNRSTVSPENPKSSASTAVSAAPEKEFPKTHSDVSSEQHRLRGQTGINGN
VKLVTITDITDILQILEPSSSPGKETNESPWRSLLIDENCLLSPLAGEDDSFLLEGN
SNECKRLLPLFLEKSSPSSSPPSSSPPSSSPPSSSPPSSSPPSSSPPSSSPPSSSPPSSS
YCOASFFGANIIGNKMSAIVHGVTSGGQMYHYDINIASLSQQDQKRFVNIPIPIVVG
SENINRCQSGDDNLTSLGTLNFPGRITVFSNGYSSPSMRPDVSSPPSSSSTATTPPKL
CLVCSDEASGCHYGLVTCGCKVVEKRAVGOHNYLCAGNDCTIIDKIRRNCPACRFK
CLOAGMNEARKTKKIKGKQVSPKQVSPKQVSPKQVSPKQVSPKQVSPKQVSPKQVSPKQV
PEVLYAGYDSSVPDSTWRIMTILNMLGGRVIAAVRARAIFGRNLHLDDQMTLQYSW
MFTLAPALGKPSVQSSANILGAPDILITNEQVITLQVQDQTHMNYSSSEIHLQVSY
EEYLCMKRTLLLLSSVPKGLRSQELFDEIRMTYIKELGKAIKREGNSSQNWRQRFYQLTK
LLDSMHEVVENLLNYCFQTFDLKTMSEIEPPEMLAEIITNQIPKYSNGNIKKLLFHQR
    
```

UniProtKB
NCBI protein

Direct protein sequencing
1 %



Question

Look at this record

<https://www.ncbi.nlm.nih.gov/nucore/AADB02000196>

Any annotated CDS ?

20850, USA

COMMENT This is the November 2001 combined whole genome shotgun assembly applied to the 27 million reads of Celera's whole genome shotgun data and 16 million reads of shredded GenBank data from other human genome projects (Nature 2001. 409:860-921). It relied on Celera's paired reads and BAC end reads from TIGR for long range order and orientation. Its scaffolds were mapped to chromosomes using STS maps. For more detailed information about whole genome sequencing and Celera's assembly process, please refer to Venter, J.C. et al. Science 2001. 291:1304-1351.

FEATURES Location/Qualifiers

source 1..129668

/organism="Homo sapiens"

/mol_type="genomic DNA"

/db_xref="taxon:9606"

/chromosome="1"

No submitted annotated CDS

ORIGIN

1 cagcgtgagc cactgcacct ggccagttct catgatttaa aagagagga actgcatttt

61 tcttttttct ttttttttat ggagtctcgc tctgtcgccc aggctggagt gcagtggcat

121 gatctcgget cactgcaacc tccaactccc gagttcaagc aattctcctg cctcagcctc

181 ccgagtagct gggattacag gcatgggtgg gcccggtctaa tttttgtatt ttttttagta

241 cagacagagt tttaccatgt tggccaggct ggtctcaaac tcttgacctc aggtgatccg

301 cctgctttgg cctcccaagg tgctgggatt acaggcgtaa gccactgcac ccagctacca

361 aaacaatttt gaaaaagaat gttgaggggg ctgggggtga cttgttctgc ccataccaaa

421 atgctgaata aagtgaaaat tttaaatgga ggaccagcac aggaatatat aataggccaa

481 caagaagtag tttttaaagg ggcattattac atcacagatc agccaggcat agtggcatgc

541 atttgtagtt ccagctaccc agcaggctga ggtgagagga ttaggaagag ctggaccatt

601 aattaacatt aataaatacg agggctgggt gcogtggctc atgcctgtaa tccaacact

661 ttgggaggcc aaggcaagca gatcacctga ggtctggagt ttgagaccag cctggccaac

721 acggtgaaac tggatcctta cctcatgcca tatataaagc tggatcctta cctcatgaca

781 tatataaaga tctaaacata agggaacctg gccgggcatg accgctcacc cctgtaatcc

841 tagcattttg ggaggccaag gcaggaggat cccttgagcc caggagtctg agagcagcct

901 gggcaacata gggagacctt gtctctatca aaaataaaaa taaaataaaa taaataaaaa

961 cataagggga ccaaaaacca gcagtgaaaa acaaaaaaag atctaaacct aaaaaccaa

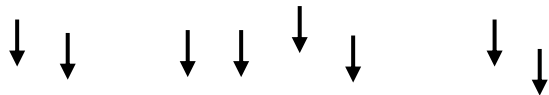
1021 atcaaattgt cctaaaacta taaaagaaaa aaggcacttt tgtggccaac cgtggtggct

The life of a protein sequence ...

```

1 agctttctggg cttccagacc cagctacttt ggggaactca gaaacccagg catctctgag
61 tctccgcca agaccgggat gcccaccagg aggtctccg gagcccaagc ttccocagat
121 aqagctccg ccagtcocaa ggttgccaa ccgctgcac tccctcccg ccaccocagg
181 tccgggaa gccccatga cccacagca cctctgcagc agcccatga gcccccagg
241 tccagag gctctccg gcccacagc aggttgccg ggggtggc ggggtggc
301 tcccaaca gcccacagc aggttgccg ggggtggc ggggtggc ggggtggc
361 ccacccccc ccagctccg agactccctg ggcaccccg gccctctcgt gctgtgcgc
421 gcaaccgct gctctccg agccggacc gggccaccg gccctctcgt ctcagacac
    
```

mRNA, genes, genomes, ...



ENA, GenBank, DDBJ

Nucleic acid databases (INSDC)

...if the submitters provide an

annotated CoDing Sequence (CDS)
(gene prediction or experimental)

no CDS
Ensembl, RefSeq
gene prediction

4 %

95 %

Direct protein sequencing
1 %

```

MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVSASSPPLAVASQSDSKQRR
LIVDPKCSVSNMQDDI SKANSI SMGI YVGRTEKIMGNDI GRDQCCQI SLSSGRTDLK
LLECSIANLNRSTSVPENPKSSASTAVSAAPTEKEFPKTHSDVSSEQHRLRGQTGINGN
VKLVTITDITDILQLEPSSSPGKETNESPWRSLLIDENCLLSPLAGEDDSFLLEGN
SNECKRLLPLFHLGKSSVSSSSVSSSSVSSSSVSSSSVSSSSVSSSSVSSSSVSSSSV
YCOASFFGANIIGNKMSAIVHGVTSGGQMYHYDINTASLSQQDQKRFVNIPIPIVVG
SENINRCQSGDDNLTSLGTLNFPGRVTFVSNYSSPSMRPDVSSPPSSSSTATTPPKL
CLVCSDEASGCHYGLVTCGCKVVEKRAVGOHNYLCAGBNDCTIIDKIRRNCPACRFK
CLOAGMNEARKTKKIKGSGVSPVSPVSPVSPVSPVSPVSPVSPVSPVSPVSPVSPVSPV
PEVLYAGYDSSVPDSTWRIMTILNMLGGRVIAAVRARAIFPFRNLHLDQMTLLQYSW
MFTLAPALGKPSVQSSANILGAPDILITNEQVTLGQVQDQTHMNYSSSEIHLQYSY
EEYLCMKRTLLLLSSVPKGLRSQELPDEIRMTYIKELGKAIKREGNSSQNWRQRFYQLTK
LLDSMHEVVENLLNYCFQTFDLKTMSEIEPPEMLAEIITNQIPKYSNGNIKKLLFHQR
    
```

Protein sequence databases

UniProtKB
NCBI protein



UniProtKB and Ensembl, RefSeq

Complementary pipelines for import of protein sequences have been developed in collaboration with

- **Ensembl** for vertebrate species,
- Ensembl Genomes for non-vertebrate species,
- WormBase, ParaSite for parasitic nematodes
- VectorBase for pathogen vector genomes.

In addition, a new pipeline imports selected non-redundant genomes annotated by NCBI **RefSeq**.

These sources provide proteome sequences for a number of key genomes of special interest where the INSDC submission is lacking gene model annotation (CDS annotation).

UniProtKB, protein sequence databases and sequence annotation

Protein sequence and annotation: overview

Nucleic acid sequence databases

INSDC, Ensembl, RefSeq

UniProtKB

UniProtKB/Swiss-Prot

UniProtKB/TrEMBL

Biochemical data (Rhea & ChEBI)

Gene Ontology

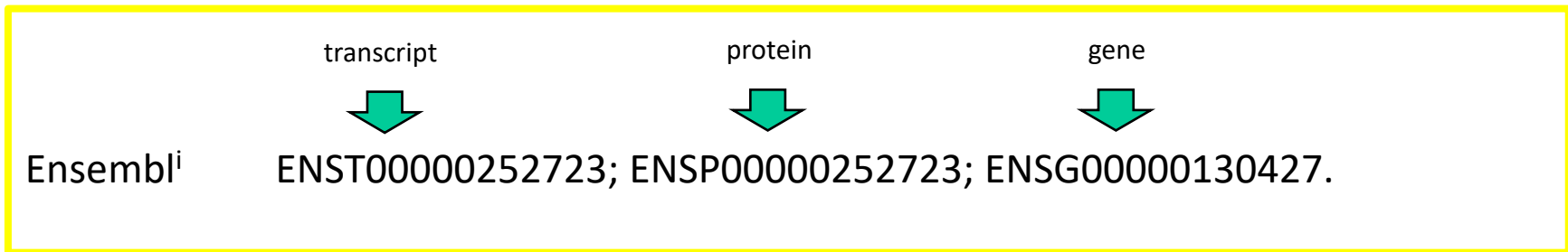
UniProt web sites and tools

NCBI Protein sequence databases

RefSeq



- Creates, integrates and distributes reference datasets
- Joint project between EBI and the Sanger Centre



- UniProtKB contains protein sequences derived from Ensembl gene prediction pipelines are not submitted to ENA/GenBank/DDBJ



Gene prediction

<https://www.ensembl.org/info/genome/index.html>

Ensembl annotation

Protein-coding genes are automatically annotated using Ensembl's genebuild pipeline. All transcripts are based on mRNA and proteins in public scientific databases.

- [Genome assemblies](#)

Pax6 INS
FOXP2
BRCA2
DMD ssh

<https://www.ensembl.org/info/genome/genebuild/index.html>

UniProtKB, protein sequence databases and sequence annotation

Protein sequence and annotation: overview

Nucleic acid sequence databases

INSDC, Ensembl, RefSeq

UniProtKB

UniProtKB/Swiss-Prot

UniProtKB/TrEMBL

Biochemical data (Rhea & ChEBI)

Gene Ontology

UniProt web sites and tools

NCBI Protein sequence databases

RefSeq

RefSeq: NCBI Reference Sequence Database

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

- Creates, integrates and distributes reference datasets, constructed from INSDC sequences.

[EPO – erythropoietin](#)

[Homo sapiens \(human\)](#)

Also known as: DBAL, ECYT5, EP, MVCD2

Gene ID: 2056

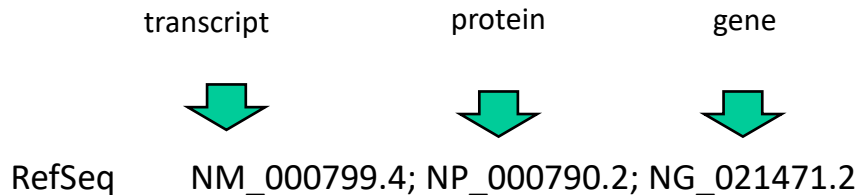
[RefSeq transcripts \(1\)](#) [RefSeq proteins \(1\)](#) [RefSeqGene \(1\)](#) [PubMed \(418\)](#)

Official Symbol EPO provided by [HGNC](#)

Official Full Name erythropoietin provided by [HGNC](#)

Primary source [HGNC:HGNC:3415](#)

See related [Ensembl:ENSG00000130427](#) [MIM:133170](#)



- UniProtKB contains protein sequences derived from RefSeq gene prediction pipelines which are not submitted to ENA/GenBank/DDBJ

RefSeq: NCBI Reference Sequence Database

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

The NCBI Eukaryotic Genome Annotation Pipeline

The NCBI Eukaryotic Genome Annotation Pipeline provides content for various NCBI resources including [Nucleotide](#), [Protein](#), [BLAST](#), [Gene](#) and the [Genome Data Viewer](#) genome browser.

This page provides an overview of the annotation process. Please refer to [the Eukaryotic Genome Annotation chapter of the NCBI Handbook](#) for algorithmic details.

The pipeline uses a modular framework for the execution of all annotation tasks from the fetching of raw and curated data from public repositories (sequence and [Assembly](#) databases) to the alignment of sequences and the prediction of genes, to the submission of the accessioned annotation products to public databases. Core components of the pipeline are alignment programs ([Splign](#) and [ProSplign](#)) and an HMM-based gene prediction program ([Gnomon](#)) developed at NCBI.

https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/

PGAP is now available as a [stand-alone software package](#). You can annotate your genomes on your own machine, local cluster or the Cloud! Get started by watching a [short video](#)!

NCBI Prokaryotic Genome Annotation Pipeline

The NCBI Prokaryotic Genome Annotation Pipeline (PGAP) is designed to annotate bacterial and archaeal genomes (chromosomes and plasmids).

Genome annotation is a multi-level process that includes prediction of protein-coding genes, as well as other functional genome units such as structural RNAs, tRNAs, small RNAs, pseudogenes, control regions, direct and inverted repeats, insertion sequences, transposons and other mobile elements.

NCBI has developed an automatic prokaryotic genome annotation pipeline that combines *ab initio* gene prediction algorithms with homology based methods. The first version of NCBI Prokaryotic Genome Pipeline was developed in 2001 and is regularly upgraded to improve structural and functional annotation quality ([Haft DH et al 2018](#), [Tatusova T et al 2016](#)). Recent improvements utilize curated protein profile hidden Markov models (HMMs), including [TIGRFAMS](#) and new HMMs for antimicrobial resistance proteins, and curated complex domain architectures for functional annotation of proteins.

https://www.ncbi.nlm.nih.gov/genome/annotation_prok/

RefSeq: NCBI Reference Sequence Database

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

Announcements

July 15, 2022

RefSeq Release 213 is available for FTP

This release includes:

Proteins: 234,520,053

Transcripts: 45,781,716

Organisms: 121,461

Available at: <ftp://ftp.ncbi.nlm.nih.gov/refseq/release/>

Documentation: [Release Notes](#)

See [previous announcements](#), follow [NCBI on Twitter](#), or subscribe to [NCBI's refseq-announce mail list](#) to receive announcements.

Organisms:

UniProtKB/Swis-Prot: ~12,000

UniProtKB/TrEMBL: ~1,290,000



(1) Look at the first submitted genome of SARS-CoV-2 (**GenBank**)

<https://www.ncbi.nlm.nih.gov/nuccore/MN908947>

How many CDS (right column: select 'Protein') ?

(2) Look at the reference genome of SARS-CoV-2 (**RefSeq**)

https://www.ncbi.nlm.nih.gov/nuccore/NC_045512

How many CDS (right column: select 'Protein') ?

(3) SARS-CoV-2 proteome in **UniProtKB**

<https://www.uniprot.org/uniprotkb/?query=proteome:UP000464024>

How many proteins (CDS) ?



(1) Look at the first submitted genome of SARS-CoV-2

<https://www.ncbi.nlm.nih.gov/nucore/MN908947>

Items: 10

[ORF10 protein \[Severe acute respiratory syndrome coronavirus 2\]](#)

1. 38 aa protein

Accession: QH42199.1 Gt: 1798172433

[Nucleotide](#) [PubMed](#) [Taxonomy](#)

[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

Identical sequence - different gene model

Note: ORF10 is 'not' in UniProtKB !

(2) Look at the reference genome of SARS-CoV-2 (RefSeq)

https://www.ncbi.nlm.nih.gov/nucore/NC_045512

Items: 12

[ORF7b \[Severe acute respiratory syndrome coronavirus 2\]](#)

1. 43 aa protein


Accession: YP_009725318.1 Gt: 1820616061

[BioProject](#) [Nucleotide](#) [PubMed](#) [Taxonomy](#)

[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

(3) SARS-CoV-2 proteome in UniProtKB

<https://www.uniprot.org/uniprot/?query=proteome:UP000464024>

 Reviewed (Swiss-Prot) (17)



A0A663DJA2 · ORF10_SARS2

Putative ORF10 protein · Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV-2) · Gene: ORF10 · 38 amino acids · Uncertain · Annotation score: 2/5

[Entry](#) [Feature viewer](#) [Publications](#) [External links](#) [History](#)

[BLAST](#) [Align](#) [Download](#) [Add](#) [Add a publication](#) [Entry feedback](#)

Functionⁱ

Caution



Could be the product of a pseudogene. Probably does not encode a functional protein. No subgenomic RNA was detected that could encode the protein (PubMed:33722935).

It has been shown to be non-essential in vivo and in vitro (PubMed:33301543).

There are no similar proteins in other betacoronavirus (PubMed:33301543). 3 Publications

Names & Taxonomyⁱ

Protein namesⁱ

Recommended name | Putative ORF10 protein 1 Publication

Gene namesⁱ

Name | ORF10 Imported

Organism namesⁱ

Organism | Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV-2) Imported



-> 5 additional protein sequences in UniProtKB
(proteome:UP000464024) NOT NC_045512

UniProtKB 5 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
<input type="checkbox"/> P0DTD2	ORF9B_SARS2	ORF9b protein[...]	9b	Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV-2)	97 AA
<input type="checkbox"/> P0DTD3	ORF9C_SARS2	Putative ORF9c protein[...]	9c	Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV-2)	73 AA
<input type="checkbox"/> P0DTF1	ORF3B_SARS2	Putative ORF3b protein[...]		Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV-2)	22 AA
<input type="checkbox"/> P0DTG1	ORF3C_SARS2	ORF3c protein[...]		Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV-2)	41 AA
<input type="checkbox"/> P0DTG0	ORF3D_SARS2	Putative ORF3d protein		Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV-2)	57 AA

Manually reviewed:
Direct contact with authors
+ publication (reviewed)

UniProtKB, protein sequence databases and sequence annotation

Protein sequence and annotation: overview

Nucleic acid sequence databases

INSDC, Ensembl, RefSeq

UniProtKB

UniProtKB/Swiss-Prot

UniProtKB/TrEMBL

Biochemical data (Rhea & ChEBI)

Gene Ontology

UniProt web sites and tools

NCBI Protein sequence databases

RefSeq



www.uniprot.org

~227 millions of proteins/records

derived from ~ 1,290,000 different species

8 million unique visitors/year

New release every 8 weeks



Release 2022_03 | Statistics



UniProt consortium : EMBL-EBI   

EBI : European Bioinformatics Institute (UK)

SIB : Swiss Institute of Bioinformatics (CH)

PIR : Protein Information Resource (USA)



Find your protein

UniProtKB Advanced | List Search

Examples: Insulin, APP, Human, P05067, organism_id:9606

UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. [Cite UniProt](#)

Accessing UniProt programmatically? Have a look at the [new API documentation](#).
If you still need it, the [legacy version of the website](#) is available until the 2022_04 release.

<h3>Proteins</h3> UniProt Knowledgebase Reviewed (Swiss-Prot) 568,002 Unreviewed (TrEMBL) 226,771,948	<h3>Species</h3> Proteomes Protein sets for species with sequenced genomes from across the tree of life	<h3>Protein Clusters</h3> UniRef Clusters of protein sequences at 100%, 90% & 50% identity	<h3>Sequence Archive</h3> UniParc Non-redundant archive of publicly available protein sequences seen across different databases
---	--	---	--

Supporting Data	Taxonomy	Keywords	Literature Citations
Human diseases	Cross-referenced databases	Subcellular locations	Automatic annotations: UniRule & ARBA

www.uniprot.org

UniProt databases

Proteins UniProt Knowledgebase

 Reviewed
(Swiss-Prot)
568,002

 Unreviewed
(TrEMBL)
226,771,948

Species Proteomes

Protein sets for species with sequenced
genomes from across the tree of life

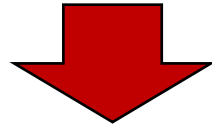
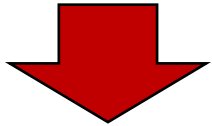
Protein Clusters UniRef

Clusters of protein sequences at 100%,
90% & 50% identity

Sequence Archive UniParc

Non-redundant archive of publicly available
protein sequences seen across different
databases

UniProt databases



Proteins UniProt Knowledgebase


Reviewed
(Swiss-Prot)
568,002


Unreviewed
(TrEMBL)
226,771,948

Species Proteomes

Protein sets for species with sequenced
genomes from across the tree of life

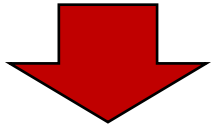
Protein Clusters UniRef

Clusters of protein sequences at 100%,
90% & 50% identity

Sequence Archive UniParc

Non-redundant archive of publicly available
protein sequences seen across different
databases

UniProt databases



Proteins UniProt Knowledgebase

 Reviewed
(Swiss-Prot)
568,002

 Unreviewed
(TrEMBL)
226,771,948

Species Proteomes

Protein sets for species with sequenced
genomes from across the tree of life

Protein Clusters UniRef

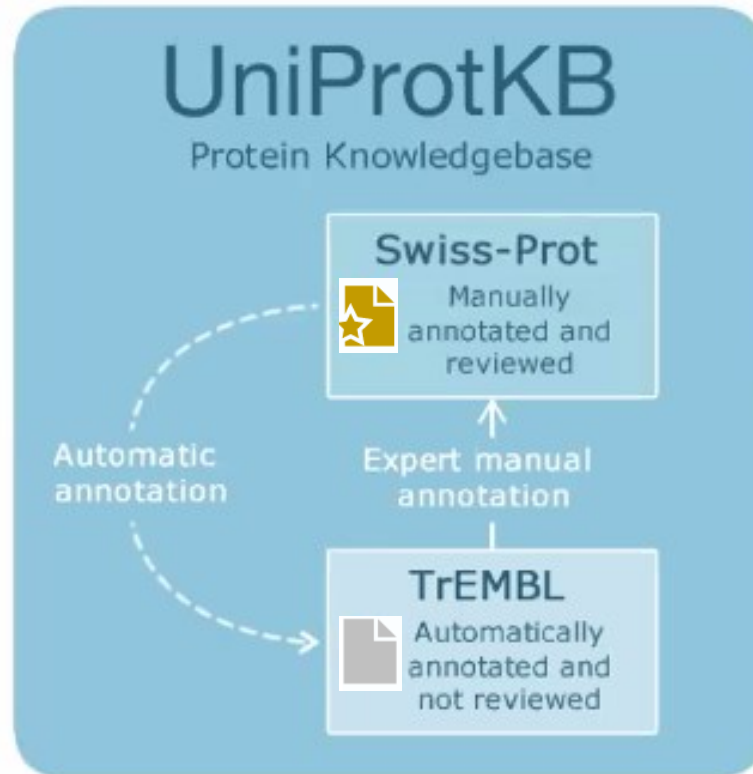
Clusters of protein sequences at 100%,
90% & 50% identity

Sequence Archive UniParc

Non-redundant archive of publicly available
protein sequences seen across different
databases

UniProtKB

Protein Knowledgebase



UniProtKB is composed of 2 sections



UniProtKB/Swiss-Prot

Reviewed - Expertly annotated

Records with information extracted from literature and curator-evaluated computational analysis.

One gene / one record



UniProtKB/TrEMBL

Unreviewed – Computationally analyzed

Records that await full manual annotation...

One protein / one record

some redundancy...

What is the % of expert curated protein entries in UniProtKB ?

Make a choice:

- 100
- 50
- 10
- 0.5



Vote

Results

Share

<https://strawpoll.com/polls/40Zmd7DaKga>

UniProtKB is composed of 2 sections



Major differences in the protein sequence and annotation accuracy !



Reviewed
(Swiss-Prot)
568,002

0.25 % of
UniProtKB protein
sequences



Unreviewed
(TrEMBL)
226,771,948

99.75 % of
UniProtKB protein
sequences

Does UniProtKB contain all protein sequences ?

UniProtKB excludes the following protein sequences:

1. Most non-germline immunoglobulins and T-cell receptors
2. Synthetic sequences
3. Most patent application sequences
4. Small fragments encoded from nucleotide sequence (<8 amino acids)
5. Pseudogenes
6. Sequences from redundant proteomes
7. Sequences from proteomes that NCBI genomes/RefSeq considers to be low quality assemblies, i.e. excluded proteomes
8. Fusion/truncated proteins
9. Not real proteins

https://www.uniprot.org/help/uniprotkb_coverage

UniProtKB entry content: overview

[Entry](#)

[Feature viewer](#)

[Publications](#)

[External links](#)

[History](#)

Protein sequence & biological knowledge

```
MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAENITTGCAEHC  
SLNENITVPDTKVNIFYAWKRMEVGGQAVEVWQGLALLSEAVLRGQALLVNSSQPWEPLQL  
HVDKAVSGLRSLTTLRRLGAQKEAISPDAASAAPLRTITADTFRKLFRVYSNFLRGKL  
KLYTGEACRTGDR
```

[Function](#)

[Names & Taxonomy](#)

[Subcellular Location](#)

[Disease & Variants](#)

[PTM/Processing](#)

[Expression](#)

[Interaction](#)

[Structure](#)

[Family & Domains](#)

[Sequence & Isoforms](#)

[Similar Proteins](#)

UniProtKB, protein sequence databases and sequence annotation

Protein sequence and annotation: overview

Nucleic acid sequence databases

INSDC, Ensembl, RefSeq

UniProtKB

UniProtKB/Swiss-Prot

UniProtKB/TrEMBL

Biochemical data (Rhea & ChEBI)

Gene Ontology

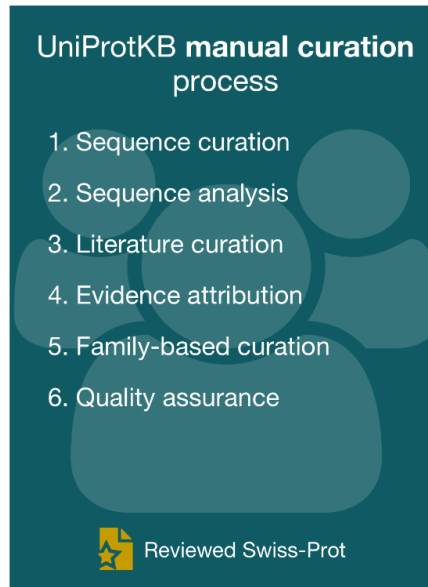
UniProt web sites and tools

NCBI Protein sequence databases

RefSeq

What is expert biocuration?

“Biocuration requires a combination of human intelligence, well-designed software tools, and advanced computational methods for literature identification and triage.”



“The SIB Swiss Institute of Bioinformatics’ resources: focus on curated databases”

<https://doi.org/10.1093/nar/gkv1310>

Video

‘Insert expert biocuration in UniProtKB/Swiss-Prot’

<https://www.youtube.com/watch?v=wvrCJtJnDbo>

Part 1 Introduction

00:30 What is UniProtKB?

01:22 What is UniProtKB/Swiss-Prot?

01:56 What type of information can be found in a Swiss-Prot record?

03:23 What is expert biocuration?

04:57 NUDT12 as an example (biological context)

Part 2 Demo

06:56 The curation editor

08:36 Import the records corresponding to the mammalian genes NUDT12

09:03 Discover the entry view (mouse NUDT12)

11:00 Protein sequence curation

12:04 Literature triage

14:00 Update of the UniProtKB/Swiss-Prot entry (mouse NUDT12)

22:42 Family-based curation propagation

25:50 Quality assurance



UniProtKB/Swiss-Prot

(1) Protein sequence

One entry – one gene – one or several protein sequences

UniProtKB/Swiss-Prot: sequence

- Validate a ‘consensus’ (**canonical**) sequence, which map to the genome sequence
- Validate and annotate the alternative **isoform** sequences, which map to the genome sequence
- At least 15% of UniProtKB/Swiss-Prot entries required manual curation effort to “correct” the sequences.

Typical problems

- unsolved conflicts (gene prediction)
- uncorrected initiation sites
- frameshifts
- other ‘problems’



- ✓ ~ 15 % of UniProtKB/Swiss-Prot entries required curation effort to “correct” the protein sequences.
- ✓ ~89 % of human UniProtKB/Swiss-Prot entries required curation effort to correct or confirm the protein sequences (CCDS and MANE select (mRNA)).

“Need for high quality sequences for learning from the language of protein (AI/ML)”

Deep Dive into Machine Learning Models for Protein Engineering

Yuting Xu,* Deeptak Verma, Robert P. Sheridan, Andy Liaw, Junshui Ma, Nicholas M. Marshall, John McIntosh, Edward C. Sherer, Vladimir Svetnik, and Jennifer M. Johnston*

Cite This: *J. Chem. Inf. Model.* 2020, 60, 10007

Read

Article Recomm

Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences

Alexander Rives^{a,b,1,2}, Joshua Meier^{a,1}, Tom Sercu^{a,1}, Siddharth Goyal^{a,1}, Zeming Lin^b, Jason Liu^a, Demi Guo^{c,3}, Myle Ott^a, C. Lawrence Zitnick^a, Jerry Ma^{a,d,e,3}, and Rob Fergus^b

^aFacebook AI Research, New York, NY 10003; ^bDepartment of Computer Science, Harvard University, Cambridge, MA 02138; ^cBooth School of Business, University of Chicago, Chicago, IL 60637; and ^dYale Law School, New Haven, CT 06511

scientific reports

Embeddings from deep learning transfer GO annotations beyond homology

Maria Littmann^{1,2,6}, Burkhard Rost^{1,3,4,5}, Michael Heinzinger^{1,2,6}, Christian Dallago^{1,2}, Tobias Olenyi¹ &

OPEN

Check for updates

Research

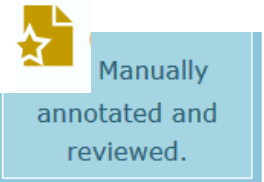
AlphaFold: a solution to a 50-year-old grand challenge in biology

November 30, 2020



<https://pubs.acs.org/doi/pdf/10.1021/acs.jcim.0c00073>


UniProtKB/Swiss-Prot: sequence



- Function
- Names & Taxonomy
- Subcellular Location
- Disease & Variants
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence & Isoforms
- Similar Proteins

Sequence & Isoformsⁱ

[BLAST 5 isoforms](#) [Align 5 isoforms](#)

Sequence statusⁱ 

Sequence processingⁱ | The displayed sequence is further processed into a mature form.

This entry describes 5 isoformsⁱ produced by **Alternative splicing & Alternative initiation**.


P04062-1



This isoform has been chosen as the **canonical** sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

Name Long

See also sequence in [UniParc](#) or sequence clusters in [UniRef](#)

Note Major isoform.  1 Publication

Tools  [Download](#)  [Add](#) [Highlight](#)  [Copy sequence](#)

Length 536

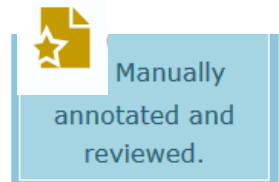
Last updated 2004-11-09 v3

Mass (Da) 59,716

Checksumⁱ FA1E15684344A0E6

```
MEFSSPSREE 10 CPKPLSRVSI 20 MAGSLTGLLL 30 LQAVSWASGA 40 RPCIPKSFY 50 SSVVCVCNAT 60 YCDSFDPPTF 70 PALGTFSRYE 80 STRSGRMEL 90
SMGPIQANHT 100 GTGLLLTLQP 110 EQKFQVKVGF 120 GGAMTDAAAL 130 NILALSPPAQ 140 NLLKSYFSE 150 EGIGYNIIRV 160 PMASCFDSIR 170 TYTYADTPDD 180
```

UniProtKB/Swiss-Prot: sequence



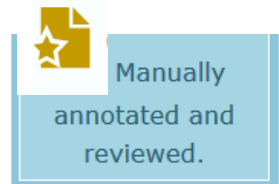
What is the canonical sequence?

Each UniProtKB/Swiss-Prot entry contains all curated protein products encoded by a given gene in a given species or strain. For each UniProtKB/Swiss-Prot entry, we choose a canonical (or representative) sequence for display that should conform to at least one of the following criteria:

1. It is functional;
2. It is widely expressed;
3. It is encoded by conserved exons found in orthologous sequences;
4. It is identical to consensus sequences chosen by other resources and genome curation efforts such as CCDS and MANE (see also [UniProt's human proteome](#));
5. In the absence of any information, we choose the longest sequence.

Sequences chosen according to these criteria generally allow the description of the majority of functionally important domains, motifs, sites, and post-translational modifications, naturally occurring variants with functional and clinical significance, and other sequence features.

UniProtKB/Swiss-Prot: sequence Source



- Function
- Names & Taxonomy
- Subcellular Location
- Disease & Variants
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence & Isoforms
- Similar Proteins

Sequence databases

CCDS	CCDS1102.1 [P04062-1] CCDS53373.1 [P04062-4] CCDS53374.1 [P04062-5]	RefSeq	NP_000148.2 [NM_000157.3] [P04062-1] NP_001005741.1 [NM_001005741.2] [P04062-1] NP_001005742.1 [NM_001005742.2] [P04062-1] NP_001165282.1 [NM_001171811.1] [P04062-4] NP_001165283.1 [NM_001171812.1] [P04062-5]
PIR	A94068 EUHUGC I52980 I52980 I67792 I67792		

SEQUENCE	PROTEIN	MOLECULE TYPE	STATUS
M16328 (EMBL) GenBank DDBJ	AAA35873.1 (EMBL) GenBank DDBJ	mRNA	
K02920 (EMBL) GenBank DDBJ	AAA35877.1 (EMBL) GenBank DDBJ	mRNA	
J03059 (EMBL) GenBank DDBJ	AAC63056.1 (EMBL) GenBank DDBJ	Genomic DNA	
D13286 (EMBL) GenBank DDBJ	BAA02545.1 (EMBL) GenBank DDBJ	mRNA	
D13287 (EMBL) GenBank DDBJ	BAA02546.1 (EMBL) GenBank DDBJ	mRNA	
AF023268 (EMBL) GenBank DDBJ	AAC51820.1 (EMBL) GenBank DDBJ	Genomic DNA	
AK291911 (EMBL) GenBank DDBJ	BAF84600.1 (EMBL) GenBank DDBJ	mRNA	
AK298900 (EMBL) GenBank DDBJ	BAH12898.1 (EMBL) GenBank DDBJ	mRNA	
AK300829 (EMBL) GenBank DDBJ	BAH13357.1 (EMBL) GenBank DDBJ	mRNA	
BC003356 (EMBL) GenBank DDBJ	AAH03356.1 (EMBL) GenBank DDBJ	mRNA	
M19285 (EMBL) GenBank DDBJ	AAA35880.1 (EMBL) GenBank DDBJ	mRNA	
M18916 (FMBI) GenBank DDBJ	AAA35878.1 (FMBI) GenBank DDBJ	Genomic DNA	Sequence problems.

Genome annotation databases

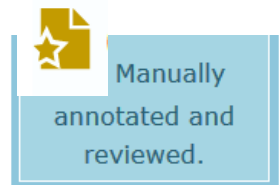
Ensembl	ENST00000327247.9 [ENSP00000314508.5] [ENSG00000177628.16] [P04062-1] ENST00000368373.8 [ENSP00000357357.3] [ENSG00000177628.16] [P04062-1] ENST00000427500.7 [ENSP00000402577.2] [ENSG00000177628.16] [P04062-5] ENST00000428024.3 [ENSP00000397986.2] [ENSG00000177628.16] [P04062-4]	GeneID	2629
		KEGG	hsa:2629
		MANE-Select	ENST00000368373.8 [ENSP00000357357.3] [NM_000157.4] [NP_000148.2]
		UCSC	uc001fjh.4 [human] [P04062-1]

Used to construct the UniProtKB canonical sequence and additional isoforms



P04062 · GLCM_HUMAN





UniProtKB/Swiss-Prot: sequence

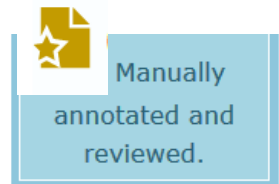
- No evidence nor quality statement for the protein sequence
 - No 'corresponding' nucleotide sequence

How do I get the nucleotide sequence that corresponds to the canonical UniProtKB sequence?

You cannot! Although more than 95% of the known protein sequences derive from DNA translation, there is no **single** nucleic acid reference sequence for a given UniProtKB/Swiss-Prot protein sequence.

https://www.uniprot.org/help/sequence_origin

UniProtKB/Swiss-Prot: sequence



- Function
- Names & Taxonomy
- Subcellular Location
- Disease & Variants
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence & Isoforms
- Similar Proteins

Automatically mapped to the UniProtKB record

Automatically mapped to the UniProtKB record

Sequence databases

CCDS	CCDS1102.1 ↗ [P04062-1]
	CCDS53373.1 ↗ [P04062-4]
	CCDS53374.1 ↗ [P04062-5]
PIR	A94068 ↗ EUHUGC
	I52980 ↗ I52980
	I67792 ↗ I67792

RefSeq	NP_000148.2 ↗ NM_000157.3 ↗ [P04062-1]
	NP_001005741.1 ↗ NM_001005741.2 ↗ [P04062-1]
	NP_001005742.1 ↗ NM_001005742.2 ↗ [P04062-1]
	NP_001165282.1 ↗ NM_001171811.1 ↗ [P04062-4]
	NP_001165283.1 ↗ NM_001171812.1 ↗ [P04062-5]

SEQUENCE	PROTEIN	MOLECULE TYPE	STATUS
M16328 (EMBL ↗ GenBank ↗ DDBJ ↗)	AAA35873.1 (EMBL ↗ GenBank ↗ DDBJ ↗)	mRNA	
K02920 (EMBL ↗ GenBank ↗ DDBJ ↗)	AAA35877.1 (EMBL ↗ GenBank ↗ DDBJ ↗)	mRNA	
J03059 (EMBL ↗ GenBank ↗ DDBJ ↗)	AAC63056.1 (EMBL ↗ GenBank ↗ DDBJ ↗)	Genomic DNA	
D13286 (EMBL ↗ GenBank ↗ DDBJ ↗)	BAA02545.1 (EMBL ↗ GenBank ↗ DDBJ ↗)	mRNA	
D13287 (EMBL ↗ GenBank ↗ DDBJ ↗)	BAA02546.1 (EMBL ↗ GenBank ↗ DDBJ ↗)	mRNA	
AF023268 (EMBL ↗ GenBank ↗ DDBJ ↗)	AAC51820.1 (EMBL ↗ GenBank ↗ DDBJ ↗)	Genomic DNA	
AK291911 (EMBL ↗ GenBank ↗ DDBJ ↗)	BAF84600.1 (EMBL ↗ GenBank ↗ DDBJ ↗)	mRNA	
AK298900 (EMBL ↗ GenBank ↗ DDBJ ↗)	BAH12898.1 (EMBL ↗ GenBank ↗ DDBJ ↗)	mRNA	
AK300829 (EMBL ↗ GenBank ↗ DDBJ ↗)	BAH13357.1 (EMBL ↗ GenBank ↗ DDBJ ↗)	mRNA	
BC003356 (EMBL ↗ GenBank ↗ DDBJ ↗)	AAH03356.1 (EMBL ↗ GenBank ↗ DDBJ ↗)	mRNA	
M19285 (EMBL ↗ GenBank ↗ DDBJ ↗)	AAA35880.1 (EMBL ↗ GenBank ↗ DDBJ ↗)	mRNA	
M18916 (FMBI ↗ GenBank ↗ DDBJ ↗)	AAA35878.1 (FMBI ↗ GenBank ↗ DDBJ ↗)	Genomic DNA	Sequence problems.

Used to construct the UniProtKB canonical sequence and additional isoforms

Genome annotation databases

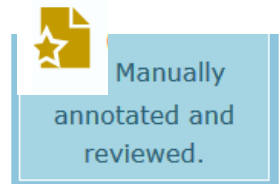
Ensembl	ENST00000327247.9 ↗ ENSP00000314508.5 ↗ ENSG00000177628.16 ↗ [P04062-1]
	ENST00000368373.8 ↗ ENSP00000357357.3 ↗ ENSG00000177628.16 ↗ [P04062-1]
	ENST00000427500.7 ↗ ENSP00000402577.2 ↗ ENSG00000177628.16 ↗ [P04062-5]
	ENST00000428024.3 ↗ ENSP00000397986.2 ↗ ENSG00000177628.16 ↗ [P04062-4]

GeneID	2629 ↗
KEGG	hsa:2629 ↗
MANE-Select	ENST00000368373.8 ↗ ENSP00000357357.3 ↗ NM_000157.4 ↗ NP_000148.2 ↗
UCSC	uc001fjh.4 ↗ human [P04062-1]

Automatically mapped to the UniProtKB record



UniProtKB/Swiss-Prot: sequence Source



- Function
- Names & Taxonomy
- Subcellular Location
- Disease & Variants
- PTM/Processing
- Expression
- Interaction
- Structure
- Family & Domains
- Sequence & Isoforms
- Similar Proteins

Sequence databases

CCDS	CCDS1102.1 [P04062-1] CCDS53373.1 [P04062-4] CCDS53374.1 [P04062-5]	RefSeq	NP_000148.2 [NM_000157.3] [P04062-1] NP_001005741.1 [NM_001005741.2] [P04062-1] NP_001005742.1 [NM_001005742.2] [P04062-1] NP_001165282.1 [NM_001171811.1] [P04062-4] NP_001165283.1 [NM_001171812.1] [P04062-5]
PIR	A94068 [EUHUGC] I52980 [I52980] I67792 [I67792]		

SEQUENCE	PROTEIN	MOLECULE TYPE	STATUS
M16328 [EMBL] [GenBank] [DDBJ]	AAA35873.1 [EMBL] [GenBank] [DDBJ]	mRNA	
K02920 [EMBL] [GenBank] [DDBJ]	AAA35877.1 [EMBL] [GenBank] [DDBJ]	mRNA	
J03059 [EMBL] [GenBank] [DDBJ]	AAC63056.1 [EMBL] [GenBank] [DDBJ]	Genomic DNA	
D13286 [EMBL] [GenBank] [DDBJ]	BAA02545.1 [EMBL] [GenBank] [DDBJ]	mRNA	
D13287 [EMBL] [GenBank] [DDBJ]	BAA02546.1 [EMBL] [GenBank] [DDBJ]	mRNA	
AF023268 [EMBL] [GenBank] [DDBJ]	AAC51820.1 [EMBL] [GenBank] [DDBJ]	Genomic DNA	
AK291911 [EMBL] [GenBank] [DDBJ]	BAF84600.1 [EMBL] [GenBank] [DDBJ]	mRNA	
AK298900 [EMBL] [GenBank] [DDBJ]	BAH12898.1 [EMBL] [GenBank] [DDBJ]	mRNA	
AK300829 [EMBL] [GenBank] [DDBJ]	BAH13357.1 [EMBL] [GenBank] [DDBJ]	mRNA	
BC003356 [EMBL] [GenBank] [DDBJ]	AAH03356.1 [EMBL] [GenBank] [DDBJ]	mRNA	
M19285 [EMBL] [GenBank] [DDBJ]	AAA35880.1 [EMBL] [GenBank] [DDBJ]	mRNA	
M18916 [FMBI] [GenBank] [DDBJ]	AAA35878.1 [FMBI] [GenBank] [DDBJ]	Genomic DNA	Sequence problems.

Genome annotation databases

Ensembl	ENST00000327247.9 [ENSP00000314508.5] [ENSG00000177628.16] [P04062-1] ENST00000368373.8 [ENSP00000357357.3] [ENSG00000177628.16] [P04062-1] ENST00000427500.7 [ENSP00000402577.2] [ENSG00000177628.16] [P04062-5] ENST00000428024.3 [ENSP00000397986.2] [ENSG00000177628.16] [P04062-4]	GeneID	2629
		KEGG	hsa:2629
		MANE-Select	ENST00000368373.8 [ENSP00000357357.3] [NM_000157.4] [NP_000148.2]
		UCSC	uc001jjh.4 [human] [P04062-1]

Used to construct the UniProtKB canonical sequence and additional isoforms

<https://www.ncbi.nlm.nih.gov/refseq/MANE/>
Matched Annotation from NCBI and EMBL-EBI (MANE) (human)



P04062 · GLCM_HUMAN



Look at this [UniProtKB](#) entry (P04150)

Question

About the protein sequences

- 'Header section'

What is the status of this entry: reviewed by a biocurator, or unreviewed?

- 'Sequence & isoforms' section

How many different protein sequences (isoforms) are available for this gene?

What is the length of the canonical protein sequence ?

How many 'computationally mapped' UniProtKB/TrEMBL entries ?

- 'Sequence & isoforms' section (sequence databases)

Look at the RefSeq cross references: How many RefSeq entries? How many entries are mapped to the UniProtKB entries ?

Query UniProtKB with the Gene name.

Look at the length of the different protein sequences in order to have a first idea of the differences that might exist...

Map your UniProtKB entries to RefSeq

- select all Acs, MapIds to Sequence databases (RefSeq)

P04150 · GCR_HUMAN



Reviewed, Swiss-Prot entry

Names & Taxonomyⁱ

Protein namesⁱ

Recommended name	Glucocorticoid receptor
Short names	GR
Alternative names	Nuclear receptor subfamily 3 group C member 1

Gene namesⁱ

Name	NR3C1
Synonyms	GRL



Organism namesⁱ

Organism	Homo sapiens (Human)
Taxonomic identifier ⁱ	9606 NCBI ↗

Taxonomic lineageⁱ [cellular organisms](#) > [Eukaryota \(eucaryotes\)](#) > [Opisthokonta](#) > [Metazoa \(metazoans\)](#) > [Eumetazoa](#) > [Bilateria](#) > [Deuterostomia](#) > [Chordata \(chordates\)](#) > [Craniata](#) > [Vertebrata \(vertebrates\)](#) > [Gnathostomata](#) > [vertebrates](#) > [Teleostomi](#) > [Euteleostomi \(bony vertebrates\)](#) > [Sarcopterygii](#) > [Dipnotetrapodomorpha](#) > [Tetrapoda \(tetrapods\)](#) > [Amniota \(amniotes\)](#) > [Mammalia \(mammals\)](#) > [Theria](#) > [Eutheria \(placental mammals\)](#) > [Boreoeutheria](#) > [Euarchontoglires](#) > [Primates](#) > [Haplorrhini](#) > [Simiiformes](#) > [Catarrhini](#) > [Hominoidea \(apes\)](#) > [Hominidae \(great apes\)](#) > [Homininae](#) > [Homo](#)

Accessions

Primary accession	P04150
Secondary accessions	A0ZXF9 B0LPG8 D3DQF4 F5ATB7 P04151 More accessions

Proteomeⁱ

Identifier	UP000005640
Component	Chromosome 5



*A **proteome** is the set of proteins thought to be expressed by an organism (completely sequenced genomes). See later

Sequence isoformsⁱ

BLAST 16 isoforms [View 16 isoforms](#)

This entry describes 16 isoformsⁱ produced by **Alternative splicing & Alternative initiation**.

P04150-1

This isoform has been chosen as the **canonical** sequence. All positional information in this entry refers to it. This is also the sequence that appears in the downloadable versions of the entry.

Name Alpha
Synonyms Alpha-A, GR-alphaA

Note Predominant physiological form. [1 Publication](#)

See also sequence in [UniParc](#) or sequence clusters in [UniRef](#)

Tools [Download](#) [Add](#) [Highlight](#) [Copy sequence](#)

Length 777

Mass (Da) 85,659

Last updated 1986-11-01 v1

Checksumⁱ C5C90C9A5DD16AAB

```
MDSKESLTPG REENPSSVLA QERGDVMDFY KTLRGGATVK VSASSPSLAV ASQSDSKQRR LLVDFPKGSV SNAQQPDL SK AVLSLMLGYM GETETKVMGN DLGFPQQGQI SLSSGETDLK LLEESIANLN RSTSVPENPK SSASTAVSAA PTEKEFPKTH
SDVSSEQQHL KGGQTGTNGGN VKLYTTDQST FDILQDLEFS SGPSPKETNE SPWRSDLIID ENCLLSPLAG EDDSFLLGN SNEDCKPLIL PDTKPKIKDN GDLVLSPSN VTLPPQVKTEK EDFIELCTPG VIKQEKLGTV YQASFPGAN IIGNKMSAIS
VHGVSTSGGQ MYHYDMNTAS LSQQDQKPI FNVIPPIPVG SENMNRCKGS GDDNLTSLGT LNFPGRVFS NGYSSPSMRP DVSSPPSSSS TATTGPPPKL CLVCSDEASG CHYGLVTCGS CKVFFKRAVE GQHNYLCAGR NDCIIDKIRR KNCPCARYRK
CLQAGMNL EA RKTKKIKGI QQATTGVSQE TSENPKNKI VPATLPQLTP TLVLSLEVIE PEVLYAGYDS SVPDSTWRIM TTLNMLGGRQ VIAAVKAKA IPGFRNLHLD DQMTLLQYSW MFLMAFALGW RSYRQSSANL LCFAPDLIIN EQRMTPCMY
DQCXHMILYS SELHRLQVSY EEYLCMKTL LLSVSPKDG L KSQLFDEIR MTYIKELGKA IVKREGNSSQ NMQRFYQLTK LLDSMHEVVE NLLNYCFQTF LDKTMSIEFP EMLAEIITMQ IPKYSNGNIK KLLFHQK
```

P04150-2

Name Beta
Synonyms Beta-A

Note High constitutive expression by neutrophils may provide a mechanism by which these cells escape glucocorticoid-induced cell death and up-regulation by pro-inflammatory cytokines such as IL8 further enhances their survival in the presence of glucocorticoids during inflammation. [1 Publication](#)

See also sequence in [UniParc](#) or sequence clusters in [UniRef](#)

Differences from canonical [728-777](#): 728-777: VVENLLNYCFQIFLDKIMSI EFP EMLAEIITNQIPKYSNGNIKLLFHQK

→ NVMWLKPESTSHLLI [1 Publication](#)

Computationally mapped potential isoform sequencesⁱ

There are 3 potential isoforms mapped to this entry

BLAST [Align](#) [Add](#) [View all](#)

Entry	Entry name	Gene name	Length
<input type="checkbox"/> D6RDA9	D6RDA9_HUMAN	NR3C1	144
<input type="checkbox"/> A0A494C0P1	A0A494C0P1_HUMAN	NR3C1	746
<input type="checkbox"/> Q3MSN4	Q3MSN4_HUMAN	NR3C1	145

Computationally mapped potential isoform sequencesⁱ

There are 3 potential isoforms mapped to this entry

BLAST Align Add View all

Entry	Entry name	Gene name	Length
<input type="checkbox"/> D6RDA9	D6RDA9_HUMAN	NR3C1	144
<input type="checkbox"/> A0A494C0P1	A0A494C0P1_HUMAN	NR3C1	746
<input type="checkbox"/> Q3MSN4	Q3MSN4_HUMAN	NR3C1	145

Query UniProt: Gene:NR3C1

UniProtKB 17 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
<input type="checkbox"/> E5KQF6	E5KQF6_HUMAN	Glucocorticoid receptor[...]	NR3C1, hCG_37601	Homo sapiens (Human)	778 AA
<input type="checkbox"/> A0A5B9RIM3	A0A5B9RIM3_HUMAN	Glucocorticoid receptor[...]	NR3C1, GCR, GR	Homo sapiens (Human)	777 AA
<input type="checkbox"/> B6ZGU6	B6ZGU6_HUMAN	Glucocorticoid receptor[...]	NR3C1	Homo sapiens (Human)	777 AA
<input type="checkbox"/> F1D8N4	F1D8N4_HUMAN	Glucocorticoid receptor[...]	NR3C1	Homo sapiens (Human)	777 AA
<input type="checkbox"/> P04150	GCR_HUMAN	Glucocorticoid receptor[...]	NR3C1, GRL	Homo sapiens (Human)	777 AA
<input type="checkbox"/> A0A494C0P1	A0A494C0P1_HUMAN	Glucocorticoid receptor[...]	NR3C1	Homo sapiens (Human)	746 AA
<input type="checkbox"/> H6V745	H6V745_HUMAN	Glucocorticoid receptor[...]	NR3C1	Homo sapiens (Human)	745 AA
<input type="checkbox"/> E5KQF5	E5KQF5_HUMAN	Glucocorticoid receptor[...]	NR3C1, hCG_37601	Homo sapiens (Human)	742 AA
<input type="checkbox"/> F5ATB8	F5ATB8_HUMAN	Glucocorticoid receptor[...]	NR3C1	Homo sapiens (Human)	662 AA
<input type="checkbox"/> A0A411AGV5	A0A411AGV5_HUMAN	Nuclear receptor subfamily 3 group C member 1 variant hGR-B(93)	NR3C1, GCR, GR	Homo sapiens (Human)	416 AA
<input type="checkbox"/> A0A411AGW6	A0A411AGW6_HUMAN	Nuclear receptor subfamily 3 group C member 1 variant hGR-B(54)	NR3C1, GCR, GR	Homo sapiens (Human)	402 AA
<input type="checkbox"/> A0A411AGV9	A0A411AGV9_HUMAN	Nuclear receptor subfamily 3 group C member 1 variant hGR-B(77)	NR3C1, GCR, GR	Homo sapiens (Human)	395 AA
<input type="checkbox"/> Q3MSN1	Q3MSN1_HUMAN	Nuclear receptor subfamily 3, group C, member 1 (Glucocorticoid receptor)	NR3C1	Homo sapiens (Human)	145 AA
<input type="checkbox"/> Q3MSN4	Q3MSN4_HUMAN	Glucocorticoid receptor[...]	NR3C1	Homo sapiens (Human)	145 AA
<input type="checkbox"/> D6RDA9	D6RDA9_HUMAN	Glucocorticoid receptor	NR3C1	Homo sapiens (Human)	144 AA
<input type="checkbox"/> H6V744	H6V744_HUMAN	Glucocorticoid receptor	NR3C1	Homo sapiens (Human)	118 AA
<input type="checkbox"/> L8E9L6	L8E9L6_HUMAN	Alternative protein NR3C1	NR3C1	Homo sapiens (Human)	55 AA

17 UniProtKB entries

4 UniProtKB mapped to 33 RefSeq

ID mapping 33 results found for UniProtKB_AC-ID → RefSeq_Protein

[Overview](#) [Input Parameters](#) [API Request](#)

[Download](#) View: Cards Table

13 IDs were not mapped:

[Show IDs](#)

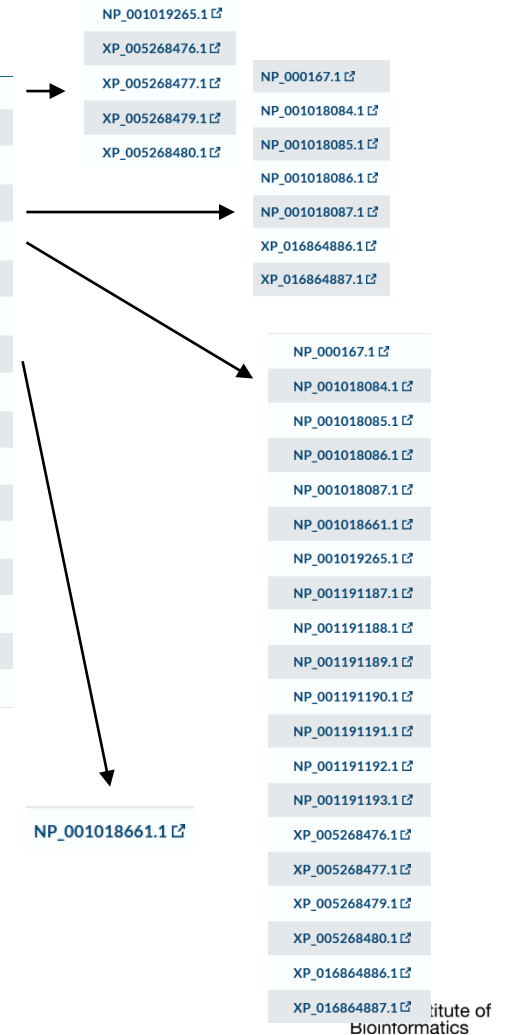
From	To
P04150	NP_000167.1
P04150	NP_001018084.1
P04150	NP_001018085.1
P04150	NP_001018086.1
P04150	NP_001018087.1
P04150	NP_001018661.1
P04150	NP_001019265.1
P04150	NP_001191187.1
P04150	NP_001191188.1
P04150	NP_001191189.1
P04150	NP_001191190.1
P04150	NP_001191191.1
P04150	NP_001191192.1

4 UniProtKB mapped to 33 RefSeq

Query UniProt: Gene:NR3C1

RefSeq

<input type="checkbox"/>	E5KQF6	E5KQF6_HUMAN	Glucocorticoid receptor[...]	NR3C1, hCG_37601	778 AA
<input type="checkbox"/>	A0A5B9RIM3	A0A5B9RIM3_HUMAN	Glucocorticoid receptor[...]	NR3C1, GCR, GR	777 AA
<input type="checkbox"/>	B6ZGU6	B6ZGU6_HUMAN	Glucocorticoid receptor[...]	NR3C1	777 AA
<input type="checkbox"/>	F1D8N4	F1D8N4_HUMAN	Glucocorticoid receptor[...]	NR3C1	777 AA
<input type="checkbox"/>	P04150	GCR_HUMAN	Glucocorticoid receptor[...]	NR3C1, GRL	777 AA
<input type="checkbox"/>	A0A494COP1	A0A494COP1_HUMAN	Glucocorticoid receptor[...]	NR3C1	746 AA
<input type="checkbox"/>	H6V745	H6V745_HUMAN	Glucocorticoid receptor[...]	NR3C1	745 AA
<input type="checkbox"/>	E5KQF5	E5KQF5_HUMAN	Glucocorticoid receptor[...]	NR3C1, hCG_37601	742 AA
<input type="checkbox"/>	F5ATB8	F5ATB8_HUMAN	Glucocorticoid receptor[...]	NR3C1	662 AA
<input type="checkbox"/>	A0A411AGV5	A0A411AGV5_HUMAN	Nuclear receptor subfamily 3 group C member 1 variant hGR-B(93)	NR3C1, GCR, GR	416 AA
<input type="checkbox"/>	A0A411AGW6	A0A411AGW6_HUMAN	Nuclear receptor subfamily 3 group C member 1 variant hGR-B(54)	NR3C1, GCR, GR	402 AA
<input type="checkbox"/>	A0A411AGV9	A0A411AGV9_HUMAN	Nuclear receptor subfamily 3 group C member 1 variant hGR-B(77)	NR3C1, GCR, GR	395 AA
<input type="checkbox"/>	Q3MSN1	Q3MSN1_HUMAN	Nuclear receptor subfamily 3, group C, member 1 (Glucocorticoid receptor)	NR3C1	145 AA
<input type="checkbox"/>	Q3MSN4	Q3MSN4_HUMAN	Glucocorticoid receptor[...]	NR3C1	145 AA
<input type="checkbox"/>	D6RDA9	D6RDA9_HUMAN	Glucocorticoid receptor	NR3C1	144 AA
<input type="checkbox"/>	H6V744	H6V744_HUMAN	Glucocorticoid receptor	NR3C1	118 AA
<input type="checkbox"/>	L8E9L6	L8E9L6_HUMAN	Alternative protein NR3C1	NR3C1	55 AA



+ 20 RefSeq entries...

- Some 'mapped' to the UniProtKB entry...
 - Several RefSeq entries for the same UniProtKB sequence
- Some predicted (XP_...)

```
RefSeq | NP_000167.1 ↗ NM_000176.2 ↗ [P04150-1]
        | NP_001018084.1 ↗ NM_001018074.1 ↗ [P04150-1]
        | NP_001018085.1 ↗ NM_001018075.1 ↗ [P04150-1]
        | NP_001018086.1 ↗ NM_001018076.1 ↗ [P04150-1]
        | NP_001018087.1 ↗ NM_001018077.1 ↗ [P04150-1]
        | NP_001018661.1 ↗ NM_001020825.1 ↗ [P04150-2]
        | NP_001019265.1 ↗ NM_001024094.1 ↗ [P04150-3]
        | NP_001191187.1 ↗ NM_001204258.1 ↗ [P04150-8]
        | NP_001191188.1 ↗ NM_001204259.1 ↗ [P04150-11]
        | NP_001191189.1 ↗ NM_001204260.1 ↗ [P04150-12]
        | NP_001191190.1 ↗ NM_001204261.1 ↗ [P04150-13]
        | NP_001191191.1 ↗ NM_001204262.1 ↗ [P04150-14]
        | NP_001191192.1 ↗ NM_001204263.1 ↗ [P04150-15]
        | NP_001191193.1 ↗ NM_001204264.1 ↗ [P04150-16]
        | XP_005268476.1 ↗ XM_005268419.3 ↗
        | XP_005268477.1 ↗ XM_005268420.4 ↗
        | XP_005268479.1 ↗ XM_005268422.3 ↗ [P04150-3]
        | XP_005268480.1 ↗ XM_005268423.3 ↗ [P04150-3]
        | XP_016864886.1 ↗ XM_017009397.1 ↗
        | XP_016864887.1 ↗ XM_017009398.1 ↗
        | Less RefSeq links
```

The definition of redundancy varies
among databases

The definition of 'redundancy' varies among databases

UniProtKB/Swiss-Prot

one record – one gene – one or several protein sequences

UniProtKB/TrEMBL

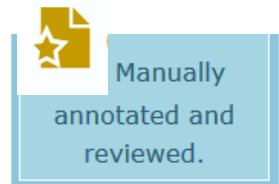
one record – one protein sequence

RefSeq

One record – one mRNA sequence – one protein sequence

Due to these different concepts,
it is not possible to draw conclusions
on the quality and completeness of a database
according to the number of entries.

UniProtKB/Swiss-Prot: sequence



One entry – one gene – one species

One or several protein sequences (isoforms) per entry

canonical & isoform

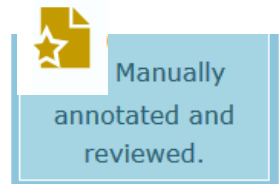
Example: GCR_HUMAN (Sequence 16 +)

<https://www.uniprot.org/uniprotkb/P04150/entry#sequences>



gene-centric / protein-centric

UniProtKB/Swiss-Prot: sequence

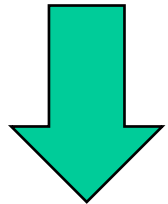


Release 2022_03 of 03-Aug-2022 of UniProtKB/Swiss-Prot contains 568002 sequence entries, curated from 287285 unique references and comprising 205171419 amino acids

521 sequences have been added since release 2022_02, the sequence data of 46 existing entries has been updated and the annotations of 377918 entries have been revised.

Number of fragments: 9289

Number of additional sequences produced by alternative splicing, initiation or promoter usage, or ribosomal frameshifting: 40825



<http://web.expasy.org/docs/relnotes/relstat.html>

Beware



The Swiss-Prot isoform sequences are not included in all datasets

Examples:

- Complete proteome* -> download Fasta (canonical & isoform)
- Blast@ NCBI (NCBI protein (nr))



BLAST Align Download Add to basket Columns

Entry	Entr
<input type="checkbox"/>	P31946 1433

Download selected (0)
 Download all (68511)
Format: FASTA (canonical & isoform)
Preview first 10 Go

Gene names
WHAB

*A **proteome** is the set of proteins thought to be expressed by an organism (completely sequenced genomes). See this afternoon



UniProtKB/Swiss-Prot

(2) Biological knowledge / annotation

Knowledge:

- comprehensive summary that provides a complete overview of the information available
- Free text or standardized vocabularies to facilitate consistent retrieval whenever possible



04

```
/translation="MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLQRYLL  
EAKEENITTCAEHCSLNENITVPDTRVNFYANKRMEVGGQAVEVWQGLALLSEAVL  
RGQALLVNSSQPWEPLQLHVDRVSGLRSLTTLRLALGAQKEAISPPDAASAAPLRTI  
TADTFRKLFRVYSNFLRGKLLKLYTGEACRTGDR'LLRALGAQKEAISPPDAAS/
```

Biocuration

From Wikipedia, the free encyclopedia

Biocuration is the field of [life sciences](#) research dedicated to translating and integrating biomedical knowledge from scientific articles to interoperable databases.^{[1][2]} The biocuration of biomedical knowledge is made possible by the cooperative work of biocurators, [software developers](#) and [bioinformaticians](#).^[1]

= adding biological information (annotation) mainly to a protein sequence.

- **expert biocurators** (reviewed) - source: publication & prediction and validation
- **automated** (unreviewed) - source: prediction

- **Free text**
- **Controlled vocabulary**, i.e. Keywords, in-house CV...
- **Ontology**, i.e. Gene Ontology, ChEBI, ...

UniProtKB: source of annotation evidence statements



Reviewed
(Swiss-Prot)
568,002

1 Publication

By Similarity

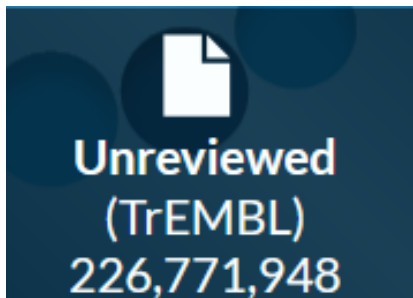
Combined Sources

Curated

1 Automatic Annotation

Imported

Could change in the future...



Unreviewed
(TrEMBL)
226,771,948

1 Automatic Annotation

Imported


InterPro Annotation


ECO Ontology:
standard ontology for evidence information


Source of manual annotation/Evidence statements


- Selected Publication (experimental)
- Another UniProtKB entry (orthologs):
- Curator-evaluated computational analysis
- An entry from another database:
- Combined sources

 1 Publication

 By Similarity

 1 Automatic Annotation


 Imported


 Combined Sources


Source of manual annotation/Evidence statements


- Selected Publication (experimental)
- Another UniProtKB entry (orthologs):
- Curator-evaluated computational analysis
- An entry from another database:
- Combined sources

 1 Publication

 By Similarity

 1 Automatic Annotation

 Imported

 Combined Sources

What is expert biocuration?

nature chemical biology

Structural and mechanistic insights into the function of mammalian Nudt12

Ewa Grudzien-Nogalska^{1,3}, Yixiang Liang Tong^{1,2*} and Megerditch Kiledjian^{1,*}

We recently demonstrated that Nudt12 is a deNADding enzyme that hydrolyzes NAD-capped RNAs. Here we reveal the molecular basis of the deNADding activity of mouse Nudt12 in complex with the endogenous NAD-capped RNA in cells, and its endogenous role in cellular energetics. Furthermore, exposure to NAD⁺ increases cellular NAD levels that are selectively responsive to Nudt12.

The redox cofactor nicotinamide adenine dinucleotide (NAD⁺) was recently reported to be covalently linked to the 5' end of mammalian mRNAs.

ARTICLES
<https://doi.org/10.1038/s41589-019-0719-0>

Cell Reports

Decapping Enzyme NUDT12 Partners with BLMH for Cytoplasmic Surveillance of NAD-Capped RNAs

Graphical Abstract

Review

Discovery of m⁷G-cap in eukaryotic mRNA

By Yasuhiro FURUICHI^{1,†}

(Communicated by Takao SEKIBA)

Abstract: Terminal structure analysis of mammalian genome RNA in the early 1970s at the National Institutes of Health revealed a methylated nucleotide in the 5' end of double-stranded RNA viruses that contain RNA polymerase II (RNAP II) promoters. This methylated nucleotide (m⁷G) generates a unique blocked acceptor site for RNAP II in the transcription start site (TSS) region. The m⁷GpppN cap is essential for the initiation of transcription. Here, we identify NUDT12 as a decapping enzyme for NAD-capped RNAs and demonstrate that NUDT12, in partnership with BLMH, promote RNA decay through the 5' to 3' direction. We reveal the existence of a NUDT12 complex that facilitates the recruitment of BLMH to distinct RNAs.

Authors
Hao Wu, Lingyun Li, Kuan-Ming Chen, ..., Fabienne Fleury-Olela, A. McCarthy, Ramesh S. Pillai

Correspondence
pillai@unige.ch

Summary
Eukaryotic mRNAs generally possess a 5'-end m⁷G cap that promotes their translation and stability. However, mammalian mRNAs can also carry a 5'-end nicotinamide adenine dinucleotide (NAD⁺) cap that, in contrast to the m⁷G cap, does not support translation but instead promotes

Literature triage

LitSuggest U.S. National Library of Medicine
National Center for Biotechnology Information TUTORIAL

ANONYMOUS [Logout](#)

PROJECTS

- Your First Project
- Ex: GWAS Catalog
- Ex: Protein Functions
- New Project 439
- Rhea no generic 1234
- New Project 1310
- [+ New Project](#)

LitSuggest is a web-based system for biomedical literature recommendation and curation.

Advanced machine learning and information retrieval techniques are utilized for finding and ranking publications pertinent to a topic of interest.

Curators evaluate around 50,000 to 70,000 papers per year during the course of their curation work for UniProtKB/Swiss-Prot.

On expert curation and scalability: UniProtKB/Swiss-Prot as a case study:

[Publication](#)

Going from unstructured (publication) to structured data.

“Expert curation is an essential part of the scientific process, one that adds significant value to research data and enables state of the art machine learning and artificial intelligence.”

Alan Bridge, director of Swiss-Prot

Most aspects of the complex process of biocuration cannot be replaced by machine learning !

Going from unstructured (publication) to structured data.

PubMed=16595657

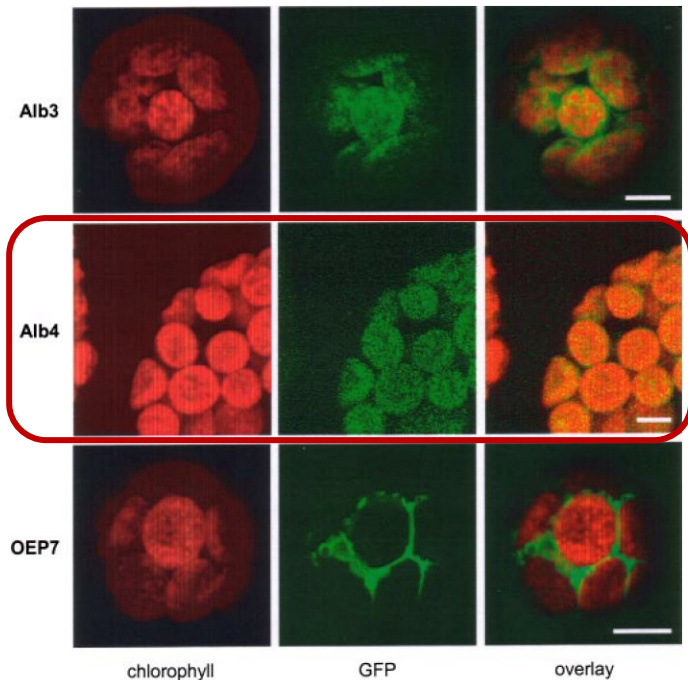


FIGURE 3. **Subcellular localization of Alb3-GFP and Alb4-GFP fusion proteins.** *Arabidopsis* mesophyll protoplasts were transiently transformed with constructs for Alb3- or Alb4-GFP. Maximum intensity signals from confocal images are shown for chlorophyll autofluorescence, GFP fluorescence, and an overlay of both. OEP7-GFP is included as a

corresponds to the predicted length of the Alb4 mRNA. Even after prolonged exposure of the blots treated with the Alb4 probe, no signal could be found at ~3.5 kb, the predicted size of the Artemis transcript.

Alb4 Is a Thylakoid Membrane Protein—Alb4 is predicted to have a chloroplast transit peptide with a processing site after amino acid resi-

due 45 based on the ChloroP prediction program (33). To test this prediction, *Arabidopsis* mesophyll protoplasts were transiently transformed with constructs for Alb3-GFP or Alb4-GFP. Merging of the GFP and autofluorescence images indicated a thylakoid localization of Alb4. The GFP distribution for Alb4 is similar to that of Alb3 and not to that of outer envelope protein AtOEP7 (Fig. 3). To test this assumption, *in vitro* translated radiolabeled Alb4 was imported into isolated pea

FROM:

PubMed=16595657

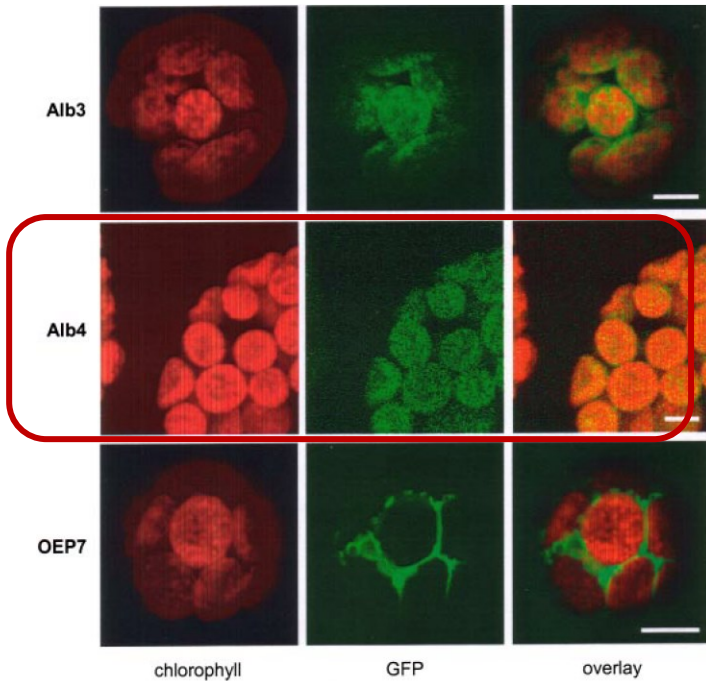


FIGURE 3. **Subcellular localization of Alb3-GFP and Alb4-GFP fusion proteins.** *Arabidopsis* mesophyll protoplasts were transiently transformed with constructs for Alb3- or Alb4-GFP. Maximum intensity signals from confocal images are shown for chlorophyll autofluorescence, GFP fluorescence, and an overlay of both. OEP7-GFP is included as a

corresponds to the predicted length of the Alb4 mRNA. Even after prolonged exposure of the blots treated with the Alb4 probe, no signal could be found at ~3.5 kb, the predicted size of the Artemis transcript.

Alb4 Is a Thylakoid Membrane Protein—Alb4 is predicted to have a chloroplast transit peptide with a processing site after amino acid resi-

due 45 based on the ChloroP prediction program (33). To test this prediction, *Arabidopsis* mesophyll protoplasts were transiently transformed with constructs for Alb3-GFP or Alb4-GFP. Merging of the GFP and autofluorescence images indicated a thylakoid localization of Alb4. The GFP distribution for Alb4 is similar to that of Alb3 and not to that of outer envelope protein AtOEP7 (Fig. 3). To test this assumption, *in vitro* translated radiolabeled Alb4 was imported into isolated pea

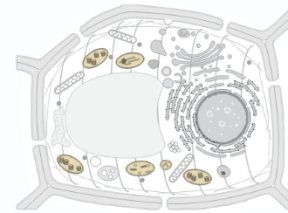
TO:

UniProtKB Q9FYL3

Subcellular Locationⁱ

UniProt Annotation

GO Annotation



Plastid, chloroplast thylakoid membrane

2 Publications; Multi-pass membrane protein

1 Automatic Annotation


Experimental data
Controlled vocabulary (CV)


Web: https://www.uniprot.org/uniprotkb/Q9FYL3/entry#subcellular_location


Source of manual annotation/Evidence statements

- Selected Publication (experimental)
- Another UniProtKB entry (orthologs):
- **Curator-evaluated computational analysis**
- An entry from another database:
- Combined sources

 1 Publication

 By Similarity

 1 Automatic Annotation

 Imported








 Combined Sources

Table 1. Sequence analysis tools used during the UniProtKB manual curation process

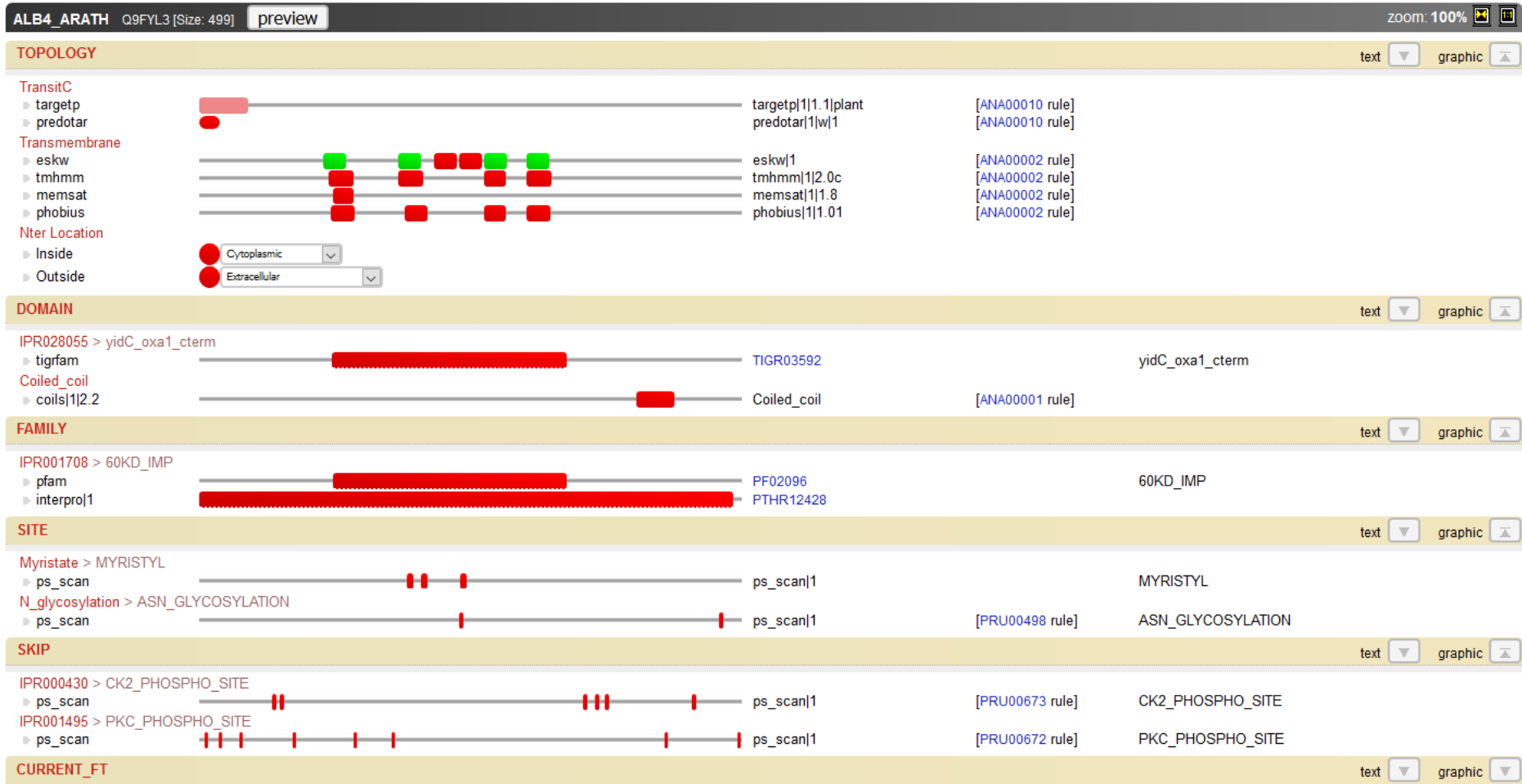
	Program	Version	Prediction
	Topology		
	Signal P (6)	3.0	Presence and location of signal peptides
	TargetP (6)	1.1	Presence and location of transit peptides
	Predotar (7)	1.03	Mitochondrial, plastid or ER targeting sequences
	ESKW* (8)	UniProt-modified version 1.0	Transmembrane domains
	MEMSAT (9)	UniProt-modified version 1.8a	Transmembrane domains
	TMHMM (10)	2.0	Transmembrane domains
	Phobius (11)	Unknown	Discriminates transmembrane and signal regions
	Post-translational modifications		
	GPI-predictor (12)	1.0	GPI lipid anchor sites
	NetNGlyc (13)	1.0	N-glycosylation sites
	NetOGlyc (14)	3.1	O-glycosylation sites
	NMT Predictor (15)	1.0	N-terminal myristoylation sites
	Sulfinator (16)	1.0	Tyrosine sulfation sites
	Domains		
	ps_scan	1.0	Internal PROSITE profile, pattern and rule scanning program
	InterPro (17)	Uses latest versions of InterPro and InterProScan	Retrieves non-PROSITE motif matches using InterPro database or InterProScan
	Coils (18)	2.2	Coiled-coil regions
	polyAA	1.0	Internal program which identifies homopolymeric stretches of amino acids
	REPEAT (19)	1.1	Identifies the following repeats: Ankyrin, Armadillo, HAT, HEAT, Kelch, Leucine-rich, PFTA, PFTB, RCC1, TPR, WD40

*ESKW = transmembrane prediction algorithm by Eisenberg, Schwarz, Komaromy and Wall

+ disordered regions: [MobiDB-lite](#)

Protein sequence analysis: in-house resource

Curator-evaluated computational analysis



FROM:

PubMed=16595657

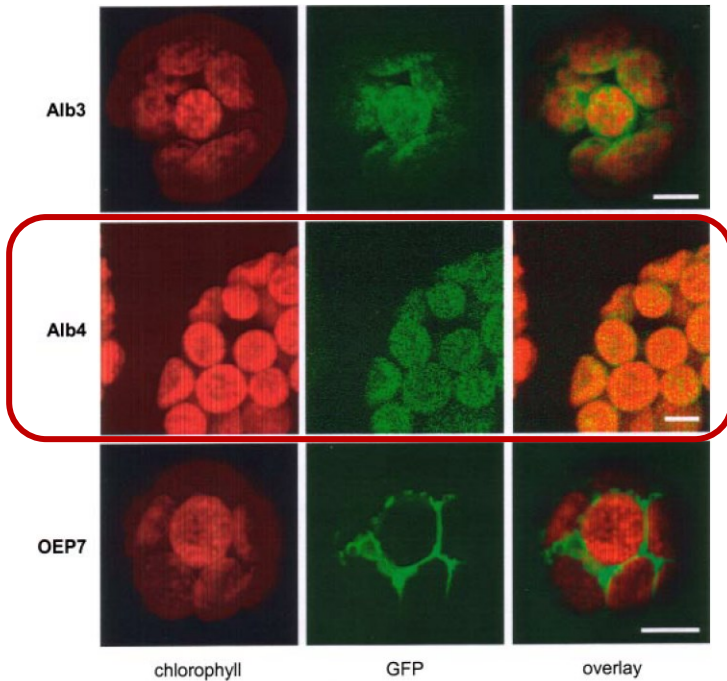


FIGURE 3. **Subcellular localization of Alb3-GFP and Alb4-GFP fusion proteins.** *Arabidopsis* mesophyll protoplasts were transiently transformed with constructs for Alb3- or Alb4-GFP. Maximum intensity signals from confocal images are shown for chlorophyll autofluorescence, GFP fluorescence, and an overlay of both. OEP7-GFP is included as a

corresponds to the predicted length of the Alb4 mRNA. Even after prolonged exposure of the blots treated with the Alb4 probe, no signal could be found at ~3.5 kb, the predicted size of the Artemis transcript

Alb4 Is a Thylakoid Membrane Protein—Alb4 is predicted to have a chloroplast transit peptide with a processing site after amino acid resi-

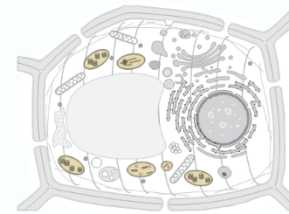
due 45 based on the ChloroP prediction program (33). To test this prediction, *Arabidopsis* mesophyll protoplasts were transiently transformed with constructs for Alb3-GFP or Alb4-GFP. Merging of the GFP and autofluorescence images indicated a thylakoid localization of Alb4. The GFP distribution for Alb4 is similar to that of Alb3 and not to that of outer envelope protein AtOEP7 (Fig. 3). To test this assumption, *in vitro* translated radiolabeled Alb4 was imported into isolated pea

TO:

UniProtKB Q9FYL3

Subcellular Locationⁱ

UniProt Annotation GO Annotation



Plastid, chloroplast thylakoid membrane
 2 Publications; Multi-pass membrane protein
 1 Automatic Annotation

Curator-evaluated prediction

Curator-evaluated prediction

▶ Transmembrane	115-135	Helical	1 Automatic Annotation
▶ Transmembrane	184-204	Helical	1 Automatic Annotation
▶ Transmembrane	263-283	Helical	1 Automatic Annotation
▶ Transmembrane	302-322	Helical	1 Automatic Annotation

UniProtKB/Swiss-Prot: PTM annotation

Glycosylation

Glycosylation is described using a strict controlled vocabulary.
Experimentally proven or curator-evaluated prediction

▶ Glycosylation	424	N-linked (GlcNAc...) (complex) asparagine	5 Publications	Combined Sources
▶ Glycosylation	506	N-linked (GlcNAc...) asparagine	3 Publications	Combined Sources

<https://www.uniprot.org/help/carbohydr>

Phosphorylation

Phosphorylation is described using a strict controlled vocabulary.
Only experimentally determined phosphorylation sites (MS).

▶ Modified residue	27	Phosphothreonine; by PLK1	2 Publications
▶ Modified residue	28	Phosphotyrosine; by SRC and HCK	1 Publication
▶ Modified residue	99	Phosphotyrosine; by ABL1	1 Publication

https://www.uniprot.org/help/post-translational_modification

UniProtKB/Swiss-Prot: 3D structure annotation

Manual annotation of entries with 3D-structures has high priority

- 3D-structures yield detailed information about the interactions of a protein with its ligands (substrates, ions, cofactors or regulatory molecules), and so help to identify active site residues.
- 3D-structures pinpoint the exact position of a residue that causes a genetic disease when it is mutated.

https://www.uniprot.org/help/3d-structure_annotation_in_swiss-prot

Look at this [UniProtKB entry \(P04150\)](#)

Question

About the biological knowledge

- 'Function' section

What is the function of the protein?

Where does the information come from?

- 'PTM/Processing' section

How many phosphorylated sites?

How many sites have been **experimentally proven** to be phosphorylated?

- 'Structure' section (3D structure databases):

Are there 3D structures available for this protein? Do they 'cover' the complete protein sequence?

- Look at the different tracks of the **Feature viewer** (including 'proteomics' and 'variants')

Look at the DNA binding domain in the 3D structure (*Hint: click on the domain 418-493*)

Which PDB entry allows to 'visualize' this domain in the 3D structure?

Source(s) of annotation

Functionⁱ

Receptor for glucocorticoids (GC) (PubMed:[27120390](#)).

Has a dual mode of action: as a transcription factor that binds to glucocorticoid response elements (GRE), both for nuclear and mitochondrial DNA, and as a modulator of other transcription factors. Affects inflammatory responses, cellular proliferation and differentiation in target tissues. Involved in chromatin remodeling (PubMed:[9590696](#)).

Plays a role in rapid mRNA degradation by binding to the 5' UTR of target mRNAs and interacting with PNRC2 in a ligand-dependent manner which recruits the RNA helicase UPF1 and the mRNA-decapping enzyme DCP1A, leading to RNA decay (PubMed:[25775514](#)).

Could act as a coactivator for STAT5-dependent transcription upon growth hormone (GH) stimulation and could reveal an essential role of hepatic GR in the control of body growth (By similarity). By Similarity **1**

Manual assertion inferred from sequence similarity (Inferred from sequence or structural similarity)ⁱ

P06537

 P06537 · GCR_MOUSE

3 Publications

2

Manual assertion based on experiment (Inferred from experiment)ⁱ

Glucocorticoid receptor interacts with PNRC2 in a ligand-dependent manner to recruit UPF1 for rapid mRNA degradation.

Cho H., Park O.H., Park J., Ryu I., Kim J., Ko J., Kim Y.K.

[View abstract](#)



[PubMed](#)

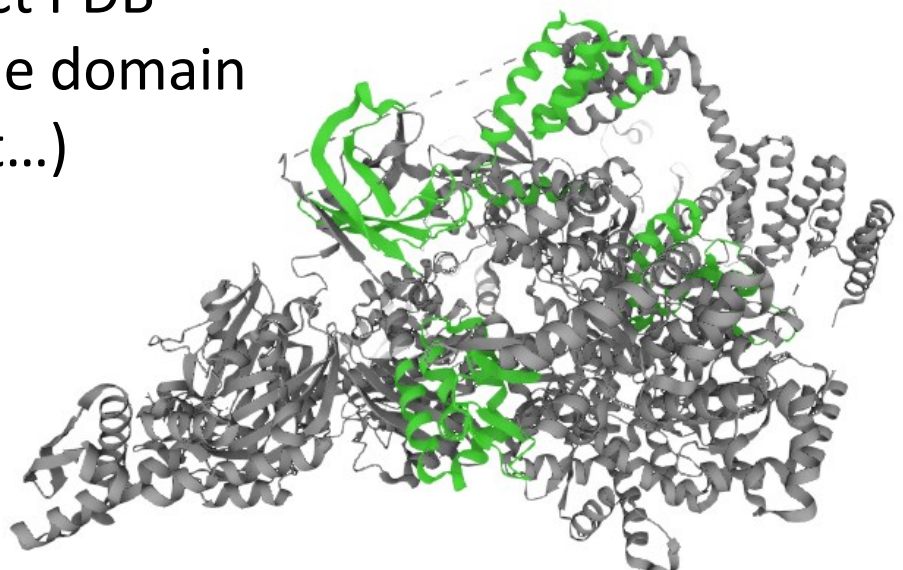
[Europe PMC](#)

Proc. Natl. Acad. Sci. U.S.A. 112:E1540-

E1549 (2015)



Choose the correct PDB entry (covering the domain you are looking at...)



PDB	7KW7	EM	3.57 Å	F	1-777	PDBe · RCSB-PDB · PDBj · PDBsum
AlphaFold	AF-P04150-F1	Predicted			1-777	AlphaFold

UniProtKB, protein sequence databases and sequence annotation

Protein sequence and annotation: overview

Nucleic acid sequence databases

INSDC, Ensembl, RefSeq

UniProtKB

UniProtKB/Swiss-Prot

UniProtKB/TrEMBL

Biochemical data (Rhea & ChEBI)

Gene Ontology

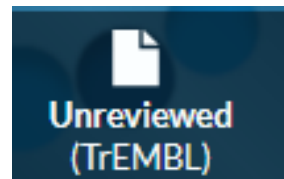
UniProt web sites and tools

NCBI Protein sequence databases

RefSeq

UniProtKB/TrEMBL

Protein sequence



Protein sequence


- The quality of the protein sequences is dependent on the information provided by the submitter of the original nucleotide entry (EMBL-ENA CDS) or of the gene prediction pipeline (i.e. Ensembl, RefSeq).
- **100% identical sequences (same length, same organism are merged automatically).**


gene-centric / **protein-centric**





















One protein sequence per entry

Some redundancy with UniProtKB/Swiss-Prot

(gene:GYPA) AND (taxonomy_id:9606)

 Reviewed (Swiss-Prot) (1)

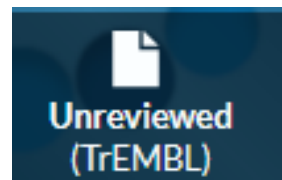
 Unreviewed (TrEMBL) (49)

<input type="checkbox"/>	A0A0C4DFT7	 A0A0C4DFT7_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	150 AA
<input type="checkbox"/>	A0A7G1PFV2	 A0A7G1PFV2_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	150 AA
<input type="checkbox"/>	B8Q183	 B8Q183_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	150 AA
<input type="checkbox"/>	P02724	 GLPA_HUMAN	Glycophorin-A[...]	GYPA, GPA	Homo sapiens (Human)	150 AA
<input type="checkbox"/>	Q58HE7	 Q58HE7_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	150 AA
<input type="checkbox"/>	X5M4Z9	 X5M4Z9_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	150 AA
<input type="checkbox"/>	A0A087WU29	 A0A087WU29_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	148 AA
<input type="checkbox"/>	A0A2R8Y7F9	 A0A2R8Y7F9_HUMAN	Glycophorin-A	GYPA	Homo sapiens (Human)	145 AA
<input type="checkbox"/>	E9PD10	 E9PD10_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	137 AA
<input type="checkbox"/>	Q8WWP1	 Q8WWP1_HUMAN	Glycophorin A	GYPA	Homo sapiens (Human)	132 AA
<input type="checkbox"/>	Q8WWP2	 Q8WWP2_HUMAN	Glycophorin A	GYPA	Homo sapiens (Human)	132 AA
<input type="checkbox"/>	Q8WWP3	 Q8WWP3_HUMAN	Glycophorin A	GYPA	Homo sapiens (Human)	132 AA
<input type="checkbox"/>	Q8WWP4	 Q8WWP4_HUMAN	Glycophorin A	GYPA	Homo sapiens (Human)	132 AA
<input type="checkbox"/>	E7EQF3	 E7EQF3_HUMAN	Glycophorin	GYPA	Homo sapiens (Human)	118 AA
<input type="checkbox"/>	E9PH25	 E9PH25_HUMAN	Glycophorin-A	GYPA	Homo sapiens (Human)	105 AA
<input type="checkbox"/>	K9JI14	 K9JI14_HUMAN	Glycophorin-A	GYPA	Homo sapiens (Human)	104 AA
<input type="checkbox"/>	Q13030	 Q13030_HUMAN	Glycophorin Erik I-IV[...]	GYPA, GPErik, hCG_2026259	Homo sapiens (Human)	85 AA
<input type="checkbox"/>	A0A8E8D7U9	 A0A8E8D7U9_HUMAN	GPA	GYPA	Homo sapiens (Human)	77 AA
<input type="checkbox"/>	A0A8E8D9B7	 A0A8E8D9B7_HUMAN	GPA	GYPA	Homo sapiens (Human)	77 AA
<input type="checkbox"/>	G8CW02	 G8CW02_HUMAN	Glycophorin A	GYPA	Homo sapiens (Human)	77 AA

(...)

UniProtKB/TrEMBL

Biological knowledge / annotation



ENA/GenBank/DDBJ and UniProtKB

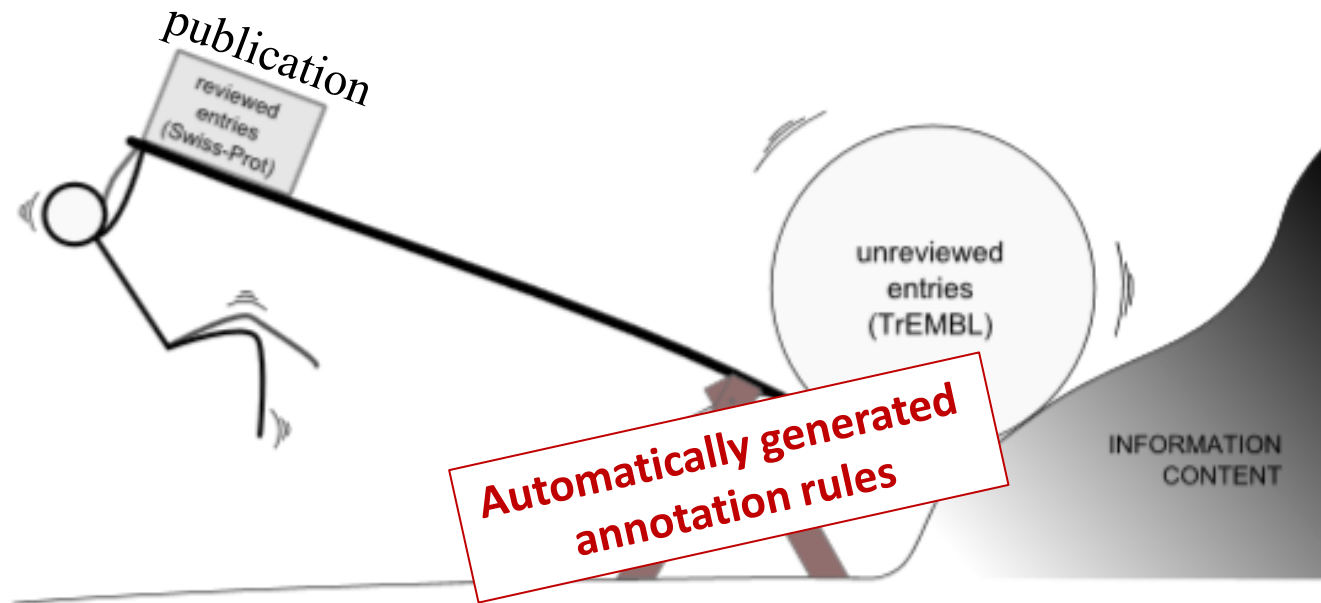
What is transferred to UniProtKB (and other protein sequence databases):

- The protein sequence (translated CDS)
- Publication provided by the submitting author
- Gene and protein names
- EC number (see later)
- Origin of the sequence (tissues)
- Taxonomy

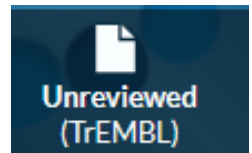


More this afternoon

Ivo Pedruzzi



Source of automated annotation



Automated generated rules (ARBA)

Multiclass learning system trained on expertly annotated entries in UniProtKB/Swiss-Prot



1 Automatic Annotation

Manually generated rules (UniRule)

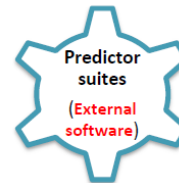
Maintains a set of manual annotation rules
UniRule = PIR + HAMAP + Rulebase



1 Automatic Annotation

Sequence analysis methods (SAM)

Signal, transmembrane, coils prediction, disordered region



1 Automatic Annotation

InterPro

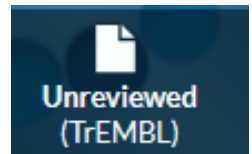
Domains & GO terms



InterPro Annotation



Source of automated annotation



Automated generated rules (ARBA)

Multiclass learning system trained on expertly annotated entries in UniProtKB/Swiss-Prot



1 Automatic Annotation

Manually generated rules (UniRule)

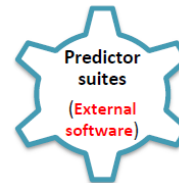
Maintains a set of manual annotation rules
UniRule = PIR + HAMAP + Rulebase



1 Automatic Annotation

Sequence analysis methods (SAM)

Signal, transmembrane, coils prediction, disordered region



1 Automatic Annotation

InterPro

Domains & GO terms



InterPro Annotation

F1MSM3 · F1MSM3_BOVIN

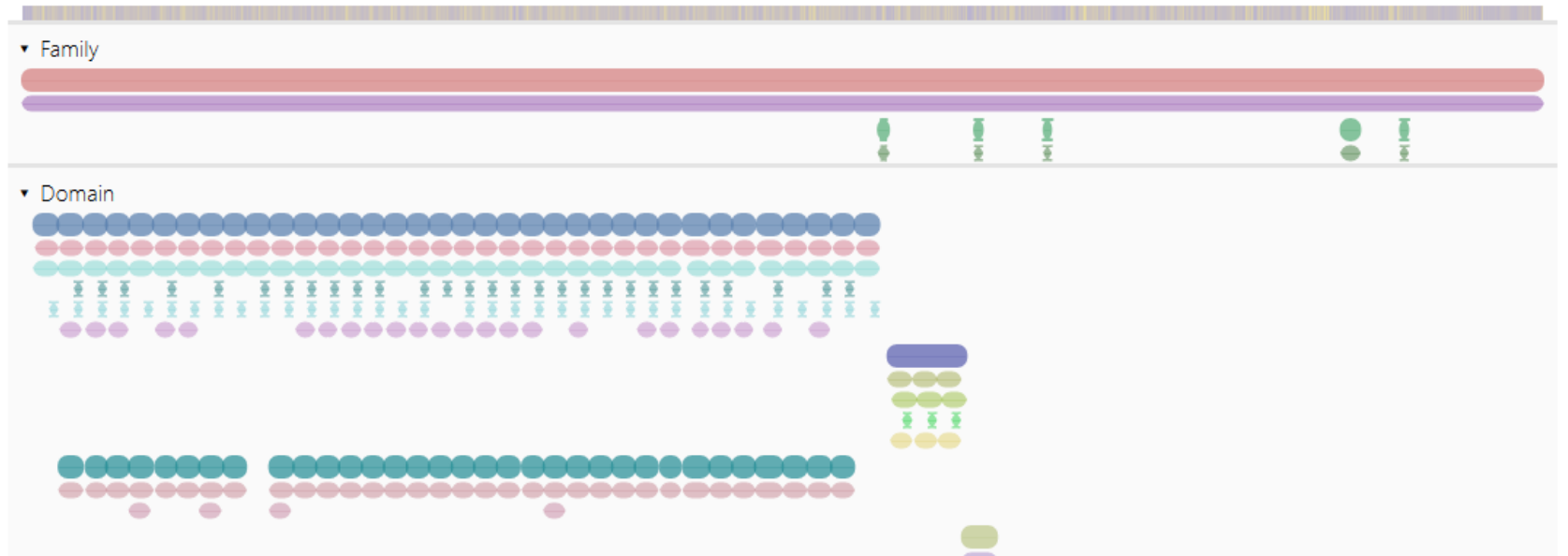
Neurogenic locus notch homolog protein 1 · **Bos taurus (Bovine)** · Gene: NOTCH1 · 2533 amino acids · Inferred from homology · Annotation score: 5/5

```
>tr|F1MSM3|F1MSM3_BOVIN Uncharacterized protein OS=Bos taurus GN=NOTCH1 PE=4 SV=2
MPPLAPLLCLALLPALAARGLRCSQPGETCLNGGKCEVFPNGTEACICGGAFAGQQCQA
PNPCLSAAPCKNGGTCHTTTEREGLVDYVCGCRLGFSGPLCLTPRDHACLASPLNGGTCDL
LTLTEYKCLCTPGWSSGKTCQQADPCASNPCANGGQCLPFEASYSICHPGPFHGPTCRQDV
NECSQSPGLCHHGTTCLNEVGSYRCVCRPTHGPHCELPHYVPCSPSPQNGGTCRPTGDT
THECACLPGFTGQNCENIIDDCPGNSCKNGGACVDGVNTYNCRCPEWTGQYCTEDVDEC
QLMPNACQNGGTCNHTGGYNCVCVNGWTGEDCSENIDDCASASCFQATGCHDRVASFYC
ECPHGRGTGLLCHLNDAICSNPCNEGSNCNTNPNVNGKAICTCPSGYTGPACSDVDECSLG
ANPCEHAGKCIINTLGSFECQCLQGYTGPRCEIDVNECVSNPCQNDATCLDQIGEFQCICM
PGYEGHLHCEVNTDECASSPCLQNGRCLDKINEFVCECPTGFTGHLGQYDVDECASTPCKN
GAKCLDGPNTYTCVCTEGYTGPHCEVDIDECDPDPCHYGSCKDGVATFTCLCQPGYTGHH
CESNINECHSQPCRHHGGTCQDRDNAYLCFCLKGTGPNCEINLDDCASNPCDSGTCLDKI
DGYECACEPGYTGSMCNINIDECADSPCHNGGTCEDGINGFTCRCPEGYHDPTCLSEVNE
CSSNPCIHGACRDSLNGYKCCDCDPSWGANCDVNNDECESNPCINGGTCCKDMTSGYVCAC
REGFSGPNCQTNINECASNPCLNQGTCIDDVAGYKCNCLLPTYGATCEVVLLAPCAPGPCR
NGGECRESEDIYESFCACPAWQGGTCEIDINECVKSPCRAGASCQNTNGSYRCHCQAGY
TGRNCETDIDDCRPNPCHNGGSCDTGINTAFCDCLPGFQGAFCEDINECASSPCRNGAN
CTDCVDSYTCPTGFSGIHCENNTPDCTESSCFNGGTCVDGINSFTCLCPPGFTGSYQ
HDVNECDSPRCLHGGTCHDSYGYTCTCPQGYTGLNCQTLVRWCDSPPCKNDGRQWNTNA
LYRCECHSGWTLGYCDVPSVSEVAARQQGVNVTHLCRNGGLCMNAGNTHRCHCQAGYTG
SYCEEQVDECSPPCQNGATCTDYPGGYSCECVAGYHGVCNSEEVNECLSQPCRNGGTCI
DLTNTYKCSCPRTQGVHCEINVDDCNPPIDPVSRRGPKCFNNGTCVDQVGGYSCSCPFGF
VGERCEGDVNECLSNPCDARGTQNCVQHVNAFHCECRAGHTGRRCESVINGCKDRPCKNG
GSCAVASNTARGFICKCPAGFEGATCENDARSCGSLRCLNGGTCIAGPRSPPTCLCLGPFPT
GPECQFPASSPCVGGNPNQGVCEPTAESPFYRCRCPAKFNGLLCHILDYSFGGGVGLD
IPPPQIEETCELPGREEAGNKVCSLQCNSHACGWDGGDCSLDFDDPWQNTQSLQCWKY
FSNGRCDSQCNAGCLFDGDFCQRAEGQCNPPLYDQYCKDHFDRGHCDQGCNSAECEWDGL
DCAEHVPERLAAGTLVLVLMPEQLRNRSLHFLRELSRLLHTNVVFKRDASGQQMIFPY
YGQEPHCRQGSAPRSVGVSTTHALLVLDKASPGHCAPPGLFSLIVYLEIDNRQCVQSSS
QCFQSATDVAAFLGALASLGLSNIYKIEAVQSETVEPPPPPLHFMVAVVAVVLLFFV
GCGVLLSRKRRRQHGQLWFPFGFKVSEASKKKRREPLGEDSVGLKPLKNSSDGALMDDNQ
NEWGDEGLEAKKFRFEEPVVLPDLDQTDHRQWTQQHLDAADLRVSAMAPFPQGEADAD
CMDVNVVRGPDGFTPLMIAASCSSGGLETGNSEEEEDAPAVISDFIYQGASLHNQTDRTGET
ALHLAARYSRSDAAKRLLEASADANIQDNMGRTPHAAVVSADAQGVFQILIRNRATDLDA
RMHDGTTPLILAAARLAVEGMLLEDLINSHADVNAVDDLKGSALHAAAAVNNVEAAVLLKN
GANKDMQNNKEETPLFLAAREGSYETAKVLLDHFANRDITDHMDRLPRDIAQERMHHDIIV
RLLEYSLVRSPLHAGATLGGTPTLSPPLCSPNGYLGNLKPPMQGKKARKPSTKGLACGG
KEPKDLKARRKKSQDGGKCLLDSSVMSPVDSLESHPGYLSDVASPPLLPSPFPQSPSVP
LNHLPGMPETHLGVSHLSVAAKPEMAVLSGGSRLEAFEAGPPRLSHLPVASTSTILGSGG
SGGSGAVNFTVGGAAAGLNGQCEWLSRLQNLVQYQYPLRGGVTPGTLSTQAAAGLQHGTV
GPLHAPALSQVMTYQALPSTRLASQPHLVQPQNLQMQPPSMPPQPNLQPHLGVSSAASG
HLGRSFLGGELSQAQDMQPLGPGNLAHAHTVLPQDQVLPSTSLPSTLAPPTMAPPMTAQFL
TPPSQHSYSSSPVDNTPSHQLQVPEHPFLTPSPESPDQWSSSSPHSNIIDWSEGISSPPT
SVPSQIAHVPEAFK
```


Protein family membership

- ▼ **F** Notch (IPR008297)
 - F** Neurogenic locus Notch 1 (IPR022362)

Entry matches to this protein ⁱ



- F** IPR008297
PIRSF002279
- F** IPR022362
PR01984
- D** IPR000742
SM00181
PS50026
PS01186
PS00022
PF00008
- D** IPR000800
SM00004
PS50258
PR01452
PF00066
- D** IPR001881
SM00179
PF07645
- D** IPR010660
SM01338
PF06816
- D** IPR011656
SM01339
PF07684
- D** IPR020683
PS50297
PF12796
- D** IPR024600
SM01334

GO terms

Biological Process

- Notch signaling pathway (GO:0007219) [ⓘ]
- cell differentiation (GO:0030154) [ⓘ]
- regulation of developmental process (GO:0050793) [ⓘ]
- multicellular organism development (GO:0007275) [ⓘ]
- regulation of transcription, DNA-templated (GO:0006355) [ⓘ]

Molecular Function

- protein binding (GO:0005515) [ⓘ]
- calcium ion binding (GO:0005509) [ⓘ]
- signaling receptor activity (GO:0038023) [ⓘ]

Cellular Component

- integral component of membrane (GO:0016021) [ⓘ]

F1MSM3 · F1MSM3_BOVIN

Neurogenic locus notch homolog protein 1 · *Bos taurus* (Bovine) · Gene: NOTCH1 · 2533 amino acids · Inferred from homology · Annotation score: 5/5

GO Annotationsⁱ

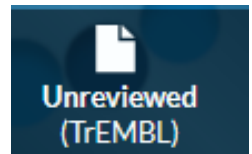
Molecular Function	calcium ion binding ↗	IEA:InterPro	
Molecular Function	chromatin DNA binding ↗	IEA:Ensembl	
Molecular Function	enzyme binding ↗	IEA:Ensembl	
Molecular Function	enzyme inhibitor activity ↗	IEA:Ensembl	
Molecular Function	identical protein binding ↗	IEA:Ensembl	
Molecular Function	Notch binding ↗	Manual Assertion Based On Experiment	IBA:GO_Central
Molecular Function	signaling receptor activity ↗	IEA:InterPro	
Molecular Function	transcription coactivator activity ↗	IEA:Ensembl	

Family & Domainsⁱ

▶ Domain	20-59	EGF-like	InterPro Annotation
▶ Domain	60-100	EGF-like	InterPro Annotation
▶ Domain	103-140	EGF-like	InterPro Annotation
▶ Domain	141-177	EGF-like	InterPro Annotation
▶ Domain	179-217	EGF-like	InterPro Annotation
▶ Domain	219-256	EGF-like	InterPro Annotation
▶ Domain	258-294	EGF-like	InterPro Annotation



Source of automated annotation



Automated generated rules (ARBA)

Multiclass learning system trained on expertly annotated entries in UniProtKB/Swiss-Prot



1 Automatic Annotation

Manually generated rules (UniRule)

Maintains a set of manual annotation rules

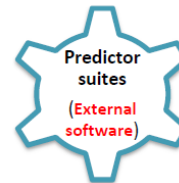
UniRule = PIR + HAMAP + Rulebase



1 Automatic Annotation

Sequence analysis methods (SAM)

Signal, transmembrane, coils prediction, disordered region



1 Automatic Annotation

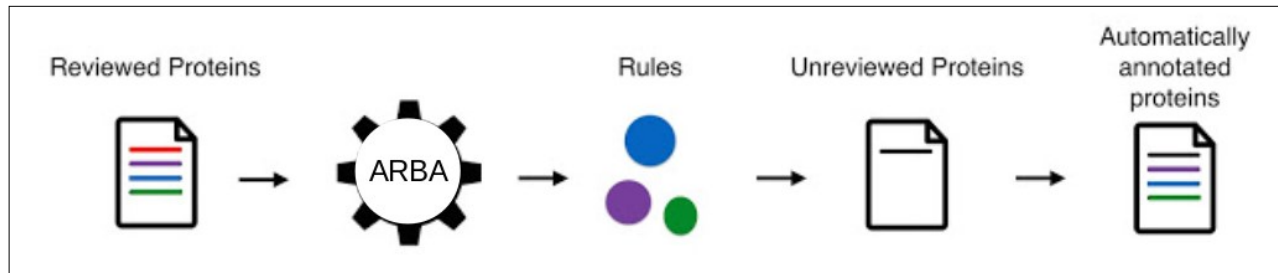
InterPro

Domains & GO terms



InterPro Annotation

Source of automated annotation/ARBA



ARBA currently generates around 27'000 models, resulting in annotation for more than 109 million protein sequences (Machine learning)

Annotation is not 'stable' !

Source of automated annotation/ARBA


ARBA 27,337 results (Number of rules)


[Download](#) View: Cards Table [Customize columns](#) [Share](#) ▼

ARBA ID	Statistics	Taxonomic scope
<input type="checkbox"/> ARBA00000001	2,481 UniProtKB entries 2,481 unreviewed UniProtKB entries	Metazoa
<input type="checkbox"/> ARBA00000002	1,669 UniProtKB entries 1,669 unreviewed UniProtKB entries	Togaviridae

Number of annotated entries

Total (2022_04)

 Reviewed (Swiss-Prot)
(568,002)

 Unreviewed (TrEMBL)
(226,771,948)



[BLAST](#) [Align](#) [Peptide search](#) [ID mapping](#) [SPARQL](#)

UniProtKB ▼

(source:arba)

Status

 Unreviewed (TrEMBL)
(108,867,323)

UniProtKB 108,867,323 results

Source of automated annotation/ARBA

ARBA - ARBA00000001

[Download](#) [View proteins](#)

IF

InterPro signature | [IPR040234](#)

taxon | Metazoa

IPR040234 Glutamyl-peptide cyclotransferase-like

THEN

catalytic activity

N-terminal L-glutamyl-[peptide] = N-terminal 5-oxo-L-prolyl-[peptide] + NH₄⁺

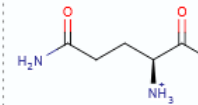
EC:2.3.2.5 (UniProtKB | ENZYME | Rhea)

Source: Rhea 23652

[^ Hide Rhea reaction](#)

N-terminal L-glutamyl-
[peptide]

RHEA-COMP:11846



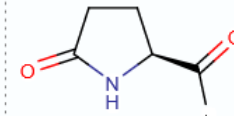
N-terminal
L-glutamine residue

CHEBI:64722

[zoom](#)

N-terminal 5-oxo-L-prolyl-
[peptide]

RHEA-COMP:11736



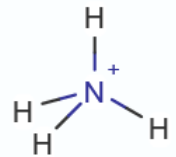
N-terminal 5-oxo-
L-proline residue

CHEBI:87215

[zoom](#)

NH₄⁺

CHEBI:28938

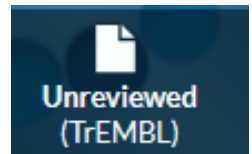


[zoom](#)

Annotated UniProtKB entries

[Browse all 2,481 entries](#)

Source of automated annotation



Automated generated rules (ARBA)

Multiclass learning system trained on expertly annotated entries in UniProtKB/Swiss-Prot



1 Automatic Annotation

Manually generated rules (UniRule)

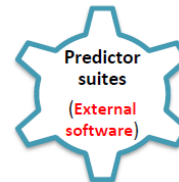
Maintains a set of manual annotation rules
UniRule = PIR + HAMAP + Rulebase



1 Automatic Annotation

Sequence analysis methods (SAM)

Signal, transmembrane, coils prediction, disordered region



1 Automatic Annotation

InterPro

Domains & GO terms



InterPro Annotation

Source of automated annotation/UniRule

UniProt BLAST Align Peptide search ID mapping SPARQL UniRule Advanced | List Search

Superkingdom

- Eukaryota (4,595)
- Bacteria (4,305)
- Archaea (1,528)
- Viruses (675)

UniRule 8,338 results (Number of rules)

Download View: Cards Table Customize columns Share

UniRule ID	Statistics	Taxonomic scope	Annotation covered	Predicted protein name
<input type="checkbox"/> UR000000032	7,756 UniProtKB entries 7,756 unreviewed UniProtKB entries	Fungi	protein name similarity keyword	Glucanase[...]
<input type="checkbox"/> UR000000033	2,997 UniProtKB entries 2,997 unreviewed UniProtKB entries	Bacteria	protein name similarity keyword	Glucanase[...]
<input type="checkbox"/> UR000000052	146,399 UniProtKB entries 146,399 unreviewed UniProtKB entries	Viridiplantae Rhodophyta Stramenopiles Haptophyceae Pyrenomonadales 2 more taxons	protein name catalytic activity cofactor subcellular location similarity 2 more annotations	Ribulose biphosphate carboxylase large chain[...]

Number of annotated entries

Total (2022_04)

Reviewed (Swiss-Prot)
(568,002)

Unreviewed (TrEMBL)
(226,771,948)

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB (source:PIR*) OR (source:HAMAP) OR (source:RuleBase)

Status

- Reviewed (Swiss-Prot) (315,287)
- Unreviewed (TrEMBL) (30,613,324)

Popular organisms

UniProtKB 53,801,987 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names
<input type="checkbox"/> O43716	GATC_HUMAN	Glutamyl-tRNA(Gln) amidotransferase	GATC, 15E1.2

Source of automated annotation/UniRule

UniRule - UR000000052

[Download](#) [View proteins](#)

Source ID: RU000302

IF	THEN																								
<table><tr><td>Pfam signature</td><td>PF00016</td></tr><tr><td>Pfam signature</td><td>PF02788</td></tr><tr><td>taxon</td><td>Viridiplantae, Rhodophyta, Stramenopiles, Haptophyceae, Pyrenomonadales, Euglenaceae, or Dinophyceae</td></tr><tr><td>gene location</td><td>Plastid</td></tr></table>	Pfam signature	PF00016	Pfam signature	PF02788	taxon	Viridiplantae, Rhodophyta, Stramenopiles, Haptophyceae, Pyrenomonadales, Euglenaceae, or Dinophyceae	gene location	Plastid	<table><tr><td>protein name</td><td>Ribulose biphosphate carboxylase large chain[...]</td></tr><tr><td>catalytic activity</td><td>2 (2R)-3-phosphoglycerate + 2 H⁺ = CO₂ + D-ribulose 1,5-bisphosphate + H₂O EC:4.1.1.39 (UniProtKB ENZYME Rhea) Source: Rhea 23124 View Rhea reaction</td></tr><tr><td></td><td>D-ribulose 1,5-bisphosphate + O₂ = (2R)-3-phosphoglycerate + 2-phosphoglycolate + 2 H⁺ Source: Rhea 36631 View Rhea reaction</td></tr><tr><td>cofactor</td><td>Mg(2+) (UniProtKB Rhea CHEBI:18420) Binds 1 Mg²⁺ ion per subunit.</td></tr><tr><td>subcellular location</td><td>Plastid, chloroplast</td></tr><tr><td>similarity</td><td>Belongs to the RuBisCO large chain family</td></tr><tr><td>subunit</td><td>Heterohexamer of 8 large chains and 8 small chains</td></tr><tr><td>keyword</td><td>Calvin cycle Carbon dioxide fixation Chloroplast Lyase Magnesium Metal-binding Monooxygenase Oxidoreductase Photorespiration Photosynthesis</td></tr></table>	protein name	Ribulose biphosphate carboxylase large chain[...]	catalytic activity	2 (2R)-3-phosphoglycerate + 2 H ⁺ = CO ₂ + D-ribulose 1,5-bisphosphate + H ₂ O EC:4.1.1.39 (UniProtKB ENZYME Rhea) Source: Rhea 23124 View Rhea reaction		D-ribulose 1,5-bisphosphate + O ₂ = (2R)-3-phosphoglycerate + 2-phosphoglycolate + 2 H ⁺ Source: Rhea 36631 View Rhea reaction	cofactor	Mg(2+) (UniProtKB Rhea CHEBI:18420) Binds 1 Mg ²⁺ ion per subunit.	subcellular location	Plastid, chloroplast	similarity	Belongs to the RuBisCO large chain family	subunit	Heterohexamer of 8 large chains and 8 small chains	keyword	Calvin cycle Carbon dioxide fixation Chloroplast Lyase Magnesium Metal-binding Monooxygenase Oxidoreductase Photorespiration Photosynthesis
Pfam signature	PF00016																								
Pfam signature	PF02788																								
taxon	Viridiplantae, Rhodophyta, Stramenopiles, Haptophyceae, Pyrenomonadales, Euglenaceae, or Dinophyceae																								
gene location	Plastid																								
protein name	Ribulose biphosphate carboxylase large chain[...]																								
catalytic activity	2 (2R)-3-phosphoglycerate + 2 H ⁺ = CO ₂ + D-ribulose 1,5-bisphosphate + H ₂ O EC:4.1.1.39 (UniProtKB ENZYME Rhea) Source: Rhea 23124 View Rhea reaction																								
	D-ribulose 1,5-bisphosphate + O ₂ = (2R)-3-phosphoglycerate + 2-phosphoglycolate + 2 H ⁺ Source: Rhea 36631 View Rhea reaction																								
cofactor	Mg(2+) (UniProtKB Rhea CHEBI:18420) Binds 1 Mg ²⁺ ion per subunit.																								
subcellular location	Plastid, chloroplast																								
similarity	Belongs to the RuBisCO large chain family																								
subunit	Heterohexamer of 8 large chains and 8 small chains																								
keyword	Calvin cycle Carbon dioxide fixation Chloroplast Lyase Magnesium Metal-binding Monooxygenase Oxidoreductase Photorespiration Photosynthesis																								

Annotated UniProtKB entries

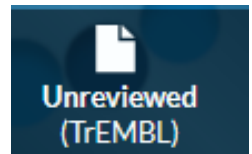
[Browse all 146,399 entries](#)



More information on
the automated gene / protein
annotation pipelines (& HAMAP) this
afternoon

Ivo Pedruzzi

Source of automated annotation



Automated generated rules (ARBA)

Multiclass learning system trained on expertly annotated entries in UniProtKB/Swiss-Prot



1 Automatic Annotation

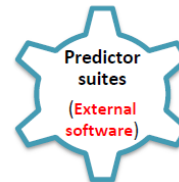
Manually generated rules (UniRule)

Maintains a set of manual annotation rules
UniRule = PIR + HAMAP + Rulebase



1 Automatic Annotation

→ **Sequence analysis methods (SAM)**
Signal, transmembrane, coils prediction, disordered region



1 Automatic Annotation

InterPro

Domains & GO terms



InterPro Annotation

Source of automated annotation/SAM

SAM - Sequence Analysis Methods for automatic annotation

UniProt's [Automatic Annotation pipeline](#) has been designed to enhance the **unreviewed records** (in UniProtKB/TrEMBL) by enriching them with automatic classification and annotation. In this context, we use a suite of Sequence Analysis Methods (SAM) to annotate extra sequence-specific information, some of which are also applied to reviewed records (in UniProtKB/Swiss-Prot).

Methods

Predictions of sequence features such as [Signal](#), [Transmembrane](#), [Coiled coil](#) and [intrinsically disordered](#) regions (the latter described in [Region](#) and [Compositional bias](#) annotations) are generated using the following software from external providers:

- [TMHMM](#)
- [SignalP](#)
- [Phobius](#)
- [Coils](#)
- [MobiDB-lite](#)

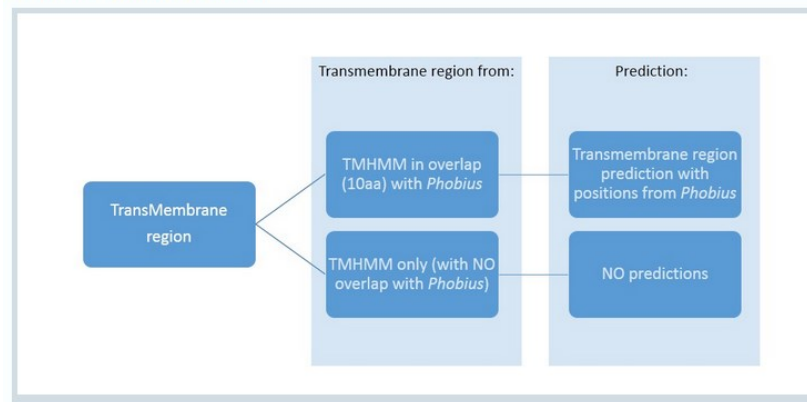
These methods are applied to UniProtKB sequences by [InterPro](#) to predict sequence features. More annotations (mainly [keywords](#)) are then added automatically to enrich the generated predictions. The new predictions are propagated to all the UniProtKB/TrEMBL records that do not already contain such feature predictions from the [UniRule](#) automatic annotation system.

Transmembrane region

TMHMM and Phobius predictors are used to infer transmembrane regions. If there is an overlap of at least 10 amino acids between TMHMM and Phobius results, the transmembrane region is annotated using the sequence ranges predicted by Phobius. Otherwise, if there is no such overlap, no predictions are generated.


See also


- [Transmembrane regions in reviewed entries](#)

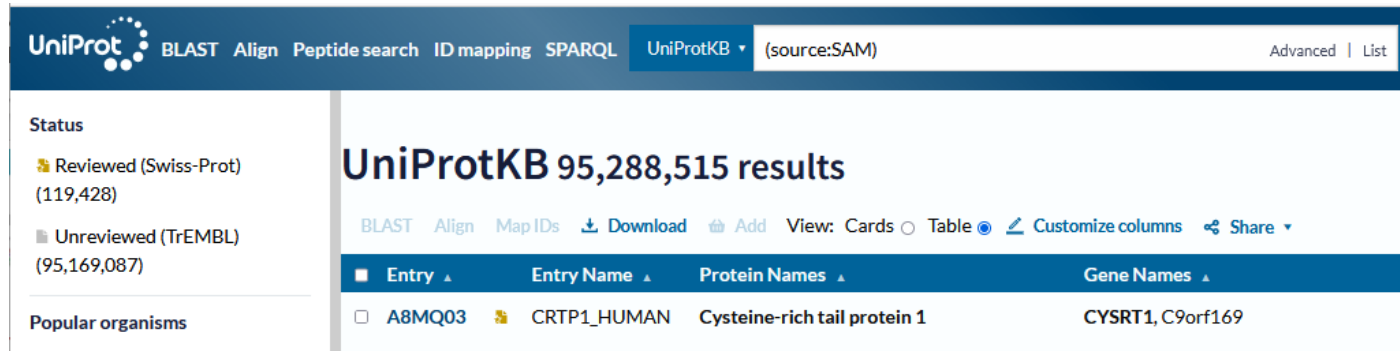


Source of automated annotation/SAM

Total (2022_04)

 Reviewed (Swiss-Prot)
(568,002)



 Unreviewed (TrEMBL)
(226,771,948)



The screenshot shows the UniProtKB search interface. The top navigation bar includes links for BLAST, Align, Peptide search, ID mapping, and SPARQL. The search bar contains 'UniProtKB' and '(source:SAM)'. The main content area displays 'UniProtKB 95,288,515 results'. Below this, there are options for BLAST, Align, Map IDs, Download, Add, View (Cards, Table), Customize columns, and Share. A table of results is shown with columns for Entry, Entry Name, Protein Names, and Gene Names. The first entry is A8MQ03, CRTP1_HUMAN, Cysteine-rich tail protein 1, and CYSRT1, C9orf169.

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB (source:SAM) Advanced | List


Status

-  Reviewed (Swiss-Prot) (119,428)
-  Unreviewed (TrEMBL) (95,169,087)

Popular organisms

UniProtKB 95,288,515 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names
<input type="checkbox"/> A8MQ03	 CRTP1_HUMAN	Cysteine-rich tail protein 1	CYSRT1, C9orf169

Source of automated annotation/SAM

Prediction of intrinsic disorder in proteins (IDP) : **ModiDB-lite**

▶ Region	1-23	Disordered	1 Automatic Annotation
▶ Compositional bias	130-149	Polar residues	1 Automatic Annotation
▶ Region	130-183	Disordered	1 Automatic Annotation
▶ Compositional bias	166-183	Polar residues	1 Automatic Annotation
		Disordered	1 Automatic Annotation
▶ Region	394-415	Automatic assertion according to rules (Automatically inferred from sequence model) ⁱ MobiDB-lite	

About 50 % of the UniProtKB proteins contain disordered regions !

These regions (IDP) are involved in:

- assembly of different subcellular structures,
- reaction crucible,
- sequestration,
- packaging for transport.

MobiDB-lite: fast and highly specific consensus prediction of intrinsic disorder in proteins FREE

Marco Necci, [Damiano Piovesan](#), [Zsuzsanna Dosztányi](#), [Silvio C.E Tosatto](#) ✉


Bioinformatics, Volume 33, Issue 9, 1 May 2017, Pages 1402–1404, <https://doi.org/10.1093/bioinformatics/btx015>


Source of automated annotation/UnProtein

NEW: release 2022_04: **Protein names**, machine learning and Google

UniProt BLAST [Align](#) [Peptide search](#) [ID mapping](#) [SPARQL](#) UniProtKB [source:Google](#)

Total (2022_04)

 Reviewed (Swiss-Prot)
(568,002)

 Unreviewed (TrEMBL)
(226,771,948)

UniProtKB 49,292,040 results

[BLAST](#) [Align](#) [Map IDs](#) [Download](#) [Add](#) View: [Cards](#) [Table](#) [Customize columns](#) [Share](#)

Entry	Entry Name	Protein Names	Gene Names
<input type="checkbox"/> A0A494C176	<input type="checkbox"/> A0A494C176_HUMAN	Secreted protein	
<input type="checkbox"/> E7ENX8	<input type="checkbox"/> E7ENX8_HUMAN	TAP2	
<input type="checkbox"/> H3BNH8			
<input type="checkbox"/> A0A087WZY1			
<input type="checkbox"/> A0A3B3ISS9			
<input type="checkbox"/> A0A0J9YYC4			
<input type="checkbox"/> O70125			
<input type="checkbox"/> Q9CUL0			
<input type="checkbox"/> Q3TRB0	<input type="checkbox"/> Q3TRB0_MOUSE	ALMS1 protein	Akap6
<input type="checkbox"/> Q9CVW5	<input type="checkbox"/> Q9CVW5_MOUSE	Collagen alpha-1(I) chain-like	Sptbn4, Spnb4
<input type="checkbox"/> Q3UHI8	<input type="checkbox"/> Q3UHI8_MOUSE	A-kinase anchor protein 6	Akap6
<input type="checkbox"/> Q8CF07	<input type="checkbox"/> Q8CF07_MOUSE	NK6 homeobox 2	Nkx6-2
<input type="checkbox"/> Q3UHR2	<input type="checkbox"/> Q3UHR2_MOUSE	A-kinase anchor protein 6	Akap6
<input type="checkbox"/> Q3V4A4	<input type="checkbox"/> Q3V4A4_MOUSE	Vesicle transport protein	Pygo2
<input type="checkbox"/> Q8CCV7	<input type="checkbox"/> Q8CCV7_MOUSE	Secreted protein	Meg3, Gtl2
<input type="checkbox"/> Q3TZQ0	<input type="checkbox"/> Q3TZQ0_MOUSE	Pericentrin-like	Cenpf
<input type="checkbox"/> Q3UZZ2	<input type="checkbox"/> Q3UZZ2_MOUSE	HES1	Hes1

Recommended name

TAP2

1 Automatic Annotation

Automatic assertion according to rules (Automatically inferred from sequence model)ⁱ

UnProtein

UniProtKB news

Google protein name predictions

UniProt has collaborated with the groups of Max Bileschi and Lucy Colwell at [Google Research](#) to predict names for UniProtKB/TrEMBL proteins. The UniProt 2021_02 release data were used to train a model called ProtNLM based on the [T5X framework](#). The model uses a shared vocabulary that encodes both protein sequences and their text descriptions (T5 methodology, [Raffel et al. 2020](#)). Free-text UniProt protein name(s) are produced as output. Expert biocurators manually evaluated a subset of model-predicted protein names chosen at random and informed model-building with stratified confidence scores. An automated verification tool also checked whether a predicted name occurs as a substring of the full UniProt entry for any protein belonging to the same UniRef50 2022_01 cluster.

Starting from release 2022_04, these name predictions will be used as the recommended name for all TrEMBL entries whose name would otherwise be "Uncharacterized protein" (49,292,040 entries in this release). The source of these protein names is indicated in their [evidence](#) tags and can be used to retrieve the corresponding entries with this [query](#).

<https://www.uniprot.org/release-notes/2022-10-12-release>

Look at this UniProtKB entry (F1D8N4)

Note: the protein sequence found in this entry is 100 % identical to the canonical protein sequence of the UniProtKB/Swiss-Prot entry P04150 (previous question)

- 'Header section'

What is the status of this entry: reviewed by a biocurator or unreviewed?
What is the annotation score?

- 'Names&Taxonomy' section

What are the name(s) of the gene and the name(s) of the protein?
What is the source of these name assignments?

- 'Subcellular location' and 'Family and domain' sections

- Look at the source of information

- Sequence section

How many protein sequence ?

- 'Cross references' section

Look (in GenBank) for the data available on the nucleic acid sequence.
Where does the protein sequence come from?
How many RefSeq entries have the same protein sequence?

Let's compare these Swiss-Prot and TrEMBL records
same gene, same protein sequence

P04150	GCR_HUMAN	1	MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVSASSPSLAVASQSDSKQRR	60
F1D8N4	F1D8N4_HUMAN	1	MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVSASSPSLAVASQSDSKQRR	60
P04150	GCR_HUMAN	61	LLVDFPKGSVSNAAQPDLSKAVSLSMGLYMGETETKVMGNDLGFPQQGQISLSSGETDLK	120
F1D8N4	F1D8N4_HUMAN	61	LLVDFPKGSVSNAAQPDLSKAVSLSMGLYMGETETKVMGNDLGFPQQGQISLSSGETDLK	120
P04150	GCR_HUMAN	121	LLEESIANLNRSTSVPENPKSSASTAVSAAPEKEFPKTHSDVSSEQQHLKGQTGTNGGN	180
F1D8N4	F1D8N4_HUMAN	121	LLEESIANLNRSTSVPENPKSSASTAVSAAPEKEFPKTHSDVSSEQQHLKGQTGTNGGN	180
P04150	GCR_HUMAN	181	VKLYTTDQSTFDILQDLEFSSGSPGKETNESPWRSDLLIDENCLLSPLAGEDDSFLLEGN	240
F1D8N4	F1D8N4_HUMAN	181	VKLYTTDQSTFDILQDLEFSSGSPGKETNESPWRSDLLIDENCLLSPLAGEDDSFLLEGN	240
P04150	GCR_HUMAN	241	SNEDCKPLILPPTTKPKIKDNGDLVLSPPSNVTLPQVKTEKEDFIELCTPGVIKQEKLGTV	300
F1D8N4	F1D8N4_HUMAN	241	SNEDCKPLILPPTTKPKIKDNGDLVLSPPSNVTLPQVKTEKEDFIELCTPGVIKQEKLGTV	300
P04150	GCR_HUMAN	301	YCQASFPG/TSGGQMYHYDMNTASLSQQQDQKPIFNVIPPPIPVG	360
F1D8N4	F1D8N4_HUMAN	301	YCQASFPG/TSGGQMYHYDMNTASLSQQQDQKPIFNVIPPPIPVG	360
P04150	GCR_HUMAN	361	SENWNRCCQGS/SDVSSPPSSSSTATIGPPPKL	420
F1D8N4	F1D8N4_HUMAN	361	SENWNRCCQGS/SDVSSPPSSSSTATIGPPPKL	420
P04150	GCR_HUMAN	421	CLVCSDEASGCHYGVLTCGSCKV/IRRNCPACRYRK	480
F1D8N4	F1D8N4_HUMAN	421	CLVCSDEASGCHYGVLTCGSCKV/IRRNCPACRYRK	480
P04150	GCR_HUMAN	481	CLQAGMNLARKTKKKIKIGIQQATTGVSQETSENPGN/ALPQLTPTLVSLLEVIE	540
F1D8N4	F1D8N4_HUMAN	481	CLQAGMNLARKTKKKIKIGIQQATTGVSQETSENPGN/ALPQLTPTLVSLLEVIE	540
P04150	GCR_HUMAN	541	PEVLYAGYDSSVPDSTWRIMTTLNMLGGRQVIAAVKWAKAIPGFRNLHLDQMTLLQYSW	600
F1D8N4	F1D8N4_HUMAN	541	PEVLYAGYDSSVPDSTWRIMTTLNMLGGRQVIAAVKWAKAIPGFRNLHLDQMTLLQYSW	600
P04150	GCR_HUMAN	601	MFLMAFALGWRSYRQSSANLLCFAPDLIINEQRMTLPCMYDQCKHMLYVSSSELHRLQVSY	660
F1D8N4	F1D8N4_HUMAN	601	MFLMAFALGWRSYRQSSANLLCFAPDLIINEQRMTLPCMYDQCKHMLYVSSSELHRLQVSY	660
P04150	GCR_HUMAN	661	EEYLCMKTLLLLSSVPKDGKLSQELFDEIRMTYIKELGKAIVKREGNSSQNWRQFYQLTK	720
F1D8N4	F1D8N4_HUMAN	661	EEYLCMKTLLLLSSVPKDGKLSQELFDEIRMTYIKELGKAIVKREGNSSQNWRQFYQLTK	720
P04150	GCR_HUMAN	721	LLDSMHEVVENLLNYCFQTFLDKTMSEIFPEMLAEIITNQIPKYSNGNIKKLLFHQK	777
F1D8N4	F1D8N4_HUMAN	721	LLDSMHEVVENLLNYCFQTFLDKTMSEIFPEMLAEIITNQIPKYSNGNIKKLLFHQK	777

100 % identical
Some redundancy...



P04150
Swiss-Prot

Names & Taxonomyⁱ

Protein namesⁱ

Recommended name	Glucocorticoid receptor
Short names	GR
Alternative names	Nuclear receptor subfamily 3 group C member 1

Gene namesⁱ

Name	NR3C1
Synonyms	GRL

Organism namesⁱ

Organism	Homo sapiens (Human)
Taxonomic identifier ⁱ	9606 NCBI ↗



F1D8N4
TrEMBL

Names & Taxonomyⁱ

Protein namesⁱ

Recommended name	Glucocorticoid receptor	1 Automatic Annotation
Alternative names	Nuclear receptor subfamily 3 group C member 1	1 Automatic Annotation

Gene namesⁱ

Name	NR3C1	Imported
------	-------	----------

Organism namesⁱ






















Organism	Homo sapiens (Human)	Imported
Taxonomic identifier ⁱ	9606 NCBI ↗	

Function¹

Receptor for glucocorticoids (GC) (PubMed:[27120390](#)).

Has a dual mode of action: as a transcription factor that binds to glucocorticoid response elements (GRE), both for nuclear and mitochondrial DNA, and as a modulator of other transcription factors. Affects inflammatory responses, cellular proliferation and differentiation in target tissues. Involved in chromatin remodeling (PubMed:[9590696](#)).

Plays a role in rapid mRNA degradation by binding to the 5' UTR of target mRNAs and interacting with PNRC2 in a ligand-dependent manner which recruits the RNA helicase UPF1 and the mRNA-decapping enzyme DCP1A, leading to RNA decay (PubMed:[25775514](#)).

Could act as a coactivator for STAT5-dependent transcription upon growth hormone (GH) stimulation and could reveal an essential role of hepatic GR in the control of body growth (By similarity).                     

Function:



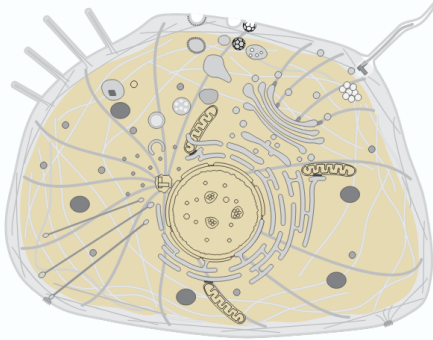
F1D8N4
TrEMBL

No CC Function

Molecular Function	DNA-binding transcription factor activity	IEA:InterPro
Molecular Function	identical protein binding	IEA:Ensembl
Molecular Function	nuclear glucocorticoid receptor activity	IEA:Ensembl
Molecular Function	protein kinase binding	IEA:Ensembl
Molecular Function	sequence-specific DNA binding	IEA:Ensembl
Molecular Function	steroid binding	IEA:UniProtKB-KW
Molecular Function	zinc ion binding	IEA:InterPro

Subcellular Locationⁱ

UniProt Annotation GO Annotation



Isoform Alpha

- Cytoplasm 5 Publications
- Nucleus 5 Publications
- Mitochondrion 1 Publication
- Cytoplasm, cytoskeleton, spindle 1 Publication
- Cytoplasm, cytoskeleton, microtubule organizing center, centrosome 1 Publication

After ligand activation, translocates from the cytoplasm to the nucleus. In the presence of NR1D1 shows a time-dependent subcellular localization, localizing to the cytoplasm at ZT8 and to the nucleus at ZT20 (By similarity).

Lacks this diurnal pattern of localization in the absence of NR1D1, localizing to both nucleus and the cytoplasm at ZT8 and ZT20 (By similarity). [By Similarity](#) [3 Publications](#)

Isoform Beta

- Nucleus 3 Publications
- Cytoplasm 2 Publications

Expressed predominantly in the nucleus with some expression also detected in the cytoplasm.

[2 Publications](#)



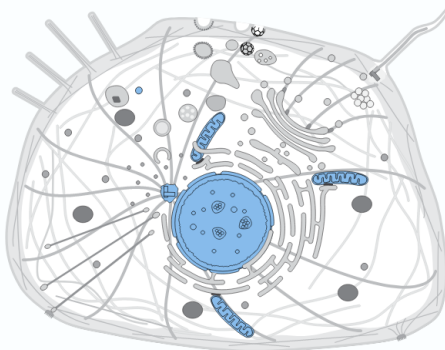
Isoform Alpha-B

- Nucleus 1 Publication
- Cytoplasm 1 Publication

After ligand activation, translocates from the cytoplasm to the nucleus. [1 Publication](#)

Subcellular Locationⁱ

UniProt Annotation GO Annotation



- Cytoplasm, cytoskeleton, microtubule organizing center, centrosome 1 Automatic Annotation
- Cytoplasm, cytoskeleton, spindle 1 Automatic Annotation
- Mitochondrion 1 Automatic Annotation
- Nucleus 2 Automatic Annotations



P04150
Swiss-Prot



F1D8N4
TrEMBL

UniProtKB/TrEMBL & Automated annotation



Important remarks

Differences between TrEMBL and Swiss-Prot



	TrEMBL	Swiss-Prot
annotation	automatic	manual
Annotation = complete ?	Partial annotation (~70 % of the entries)	As complete and systematic as possible
Set of sequences = complete ?	As complete as possible; does not contain Swiss-Prot sequences !	Complete set of sequences only for a few organisms
Number of entries	227'000'000	568'000
Number of species	1'200'000	12'000

When you compare biological information of given datasets of proteins beware the redundancy and the ratio of TrEMBL vs Swiss-Prot entries in your dataset: the results might not be only 'biological'!

Set of mouse proteins with N-glycosylation – UniProt 2022_03

N-glycosylation is not annotated (predicted) in UniProtKB/TrEMBL

	Swiss-Prot	TrEMBL	total
mouse	17,132	70,891	86,436
mouse proteome	17,120	38,195	55,315
mouse proteome + N-glycosylation	3,735	34	3,769
%	21.8 %	0.08 %	6,8 %

Query 1: (organism_id:10090)

Query 2: (proteome:UP000000589)

Query 3: (proteome:UP000000589) AND (ft_carbohyd:"N-linked (GlcNAc...)")

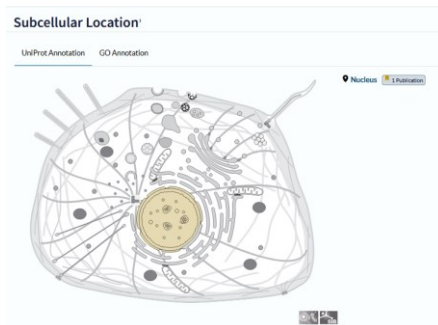
Mouse proteome: what is the % of nuclear proteins in UniProtKB/Swiss-Prot, in UniProtKB/TrEMBL, ?

1. Which % of the mouse proteome entries are in UniProtKB/Swiss-Prot & UniProtKB/TrEMBL ?

Query: (taxonomy_id:10090) AND (proteome:up000000589)

2. What is the % of nuclear proteins in UniProtKB/Swiss-Prot, in UniProtKB/TrEMBL

Query: (taxonomy_id:10090) AND (proteome:up000000589) AND (cc_scl_term:SL-0191)



Query: CC Subcellular location Nucleus (SL-0191)

1. Which % of the mouse proteome entries are in UniProtKB/Swiss-Prot & UniProtKB/TrEMBL ?

	Swiss-Prot	TrEMBL	total
mouse proteome	17,120	38.195	55,315
%	31 %	69 %	100 %

UniProt 2022_03

2. What is the % of mouse nuclear proteins in UniProtKB/Swiss-Prot, in UniProtKB/TrEMBL and in UniProtKB?

	Swiss-Prot	TrEMBL	total
mouse proteome	17,120	38.195	55,315
mouse proteome + nucleus	4,733	2,199	6,932
%	27,6 %	0.6 %	12.5 %

UniProtKB, protein sequence databases and sequence annotation

Protein sequence and annotation: overview

Nucleic acid sequence databases

INSDC, Ensembl, RefSeq

UniProtKB

UniProtKB/Swiss-Prot

UniProtKB/TrEMBL

Biochemical data (Rhea & ChEBI)

Gene Ontology

UniProt web sites and tools

NCBI Protein sequence databases

RefSeq

How often do you use Rhea?

by a guest · just now

Make a choice:

- I have never heard of Rhea
- weekly
- less
- never

Vote

 Results

 Share

<https://strawpoll.com/polls/GPgV3aNYBZa>

Rhea is a knowledgebase of biochemical transformations and transport reactions, based on the chemical ontology ChEBI.

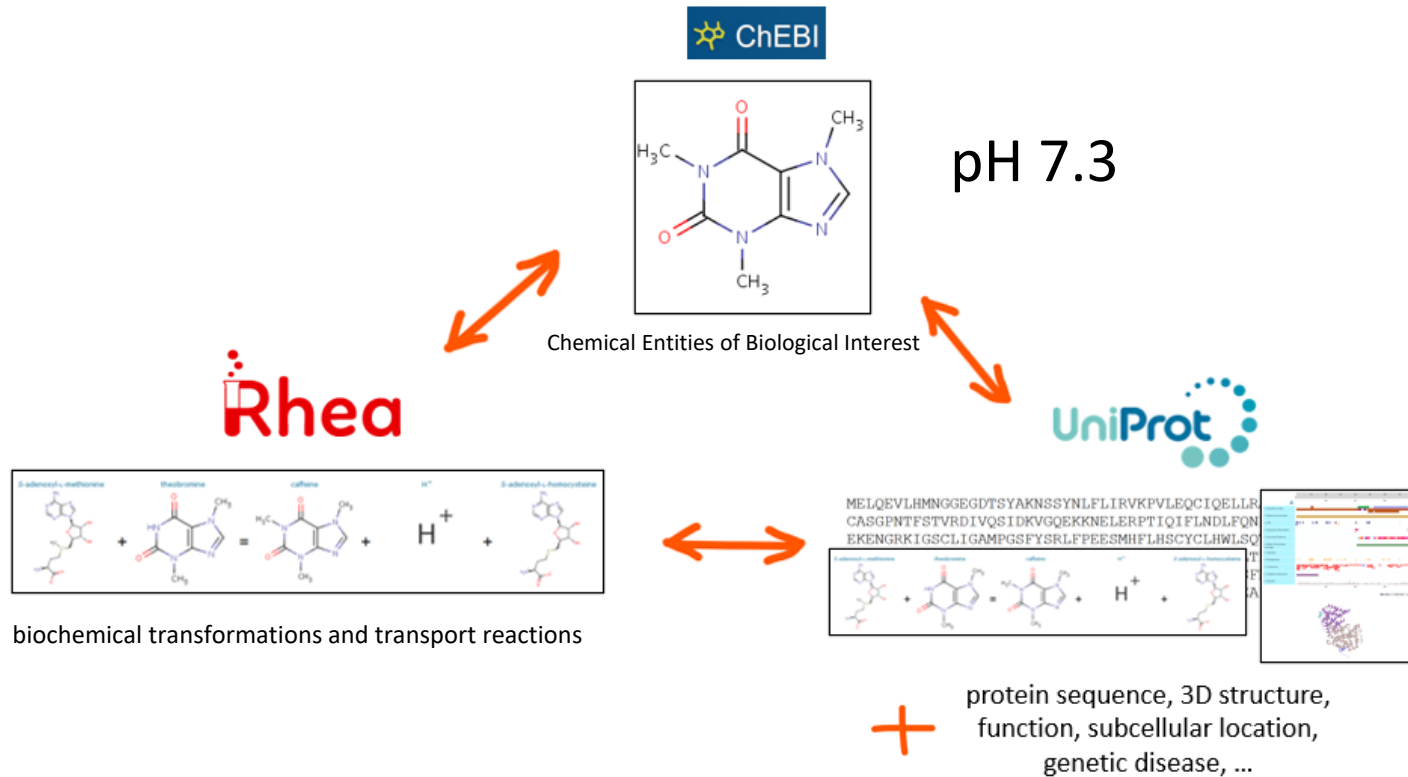
Chemical Entities of Biological Interest (**ChEBI**) is a freely available ontology of molecular entities focused on 'small' chemical compounds.

UniProtKB uses Rhea to link enzymes and transporters to explicit representations of the chemical structures of their substrates and products (metabolites).



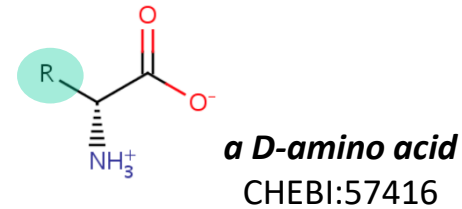
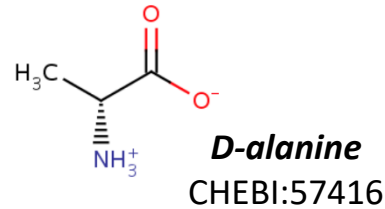
<https://www.rhea-db.org/>

The concepts summed up in one drawing

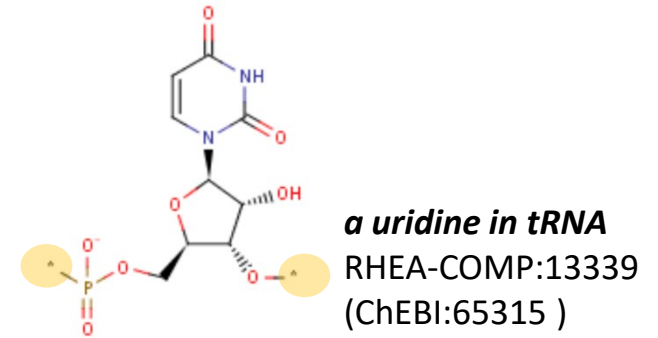
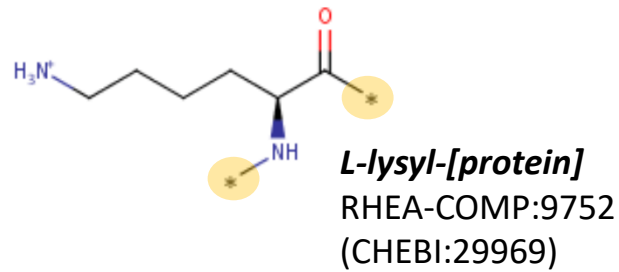


Rhea/ChEBI: reaction participants

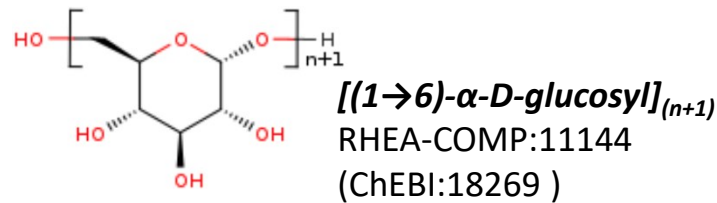
Small molecules



Macromolecules



Polymers

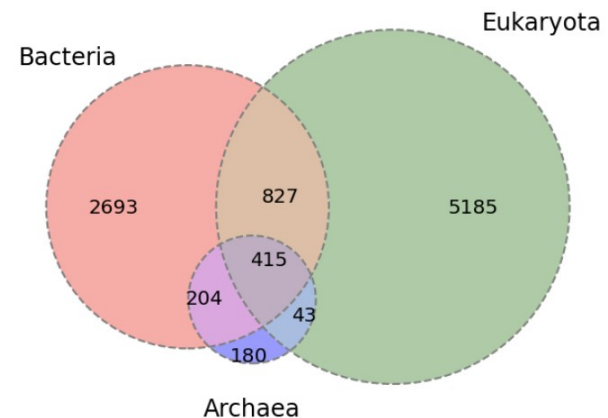


UniProtKB & Enzyme reactions (Rhea)

(biochemical transformations and transport reactions)

UniProtKB 2022_03	Swiss-Prot 568,002	%	TrEMBL 226,771,948	%
Proteins in UniProtKB linked to Rhea	226,101	~40	24,416,545	~11
Unique Rhea reactions in UniProtKB	10,540			
Unique ChEBI in UniProtKB	9,263			

Taxonomy distribution of reactions associated to sequences in UniProtKB/Swiss-Prot (2021_04 snapshot)

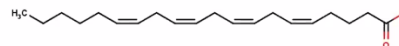


Exploiting small molecule data in UniProtKB

You can query UniProtKB with:

Names

Arachidonate



Identifiers

ChEBI:32395 (we index the ontology)

Structure descriptors

InChIKey:YZXBAPSDXZZRGB-DOFZRALJSA-M

The **International Chemical Identifier** (InChI [/ˈɪntʃiː/ IN-chee](#) or [/ˈɪŋkiː/ ING-kee](#)) is a textual **identifier** for **chemical substances**, designed to provide a standard way to encode molecular information and to facilitate the search for such information in databases and on the web.

https://en.wikipedia.org/wiki/International_Chemical_Identifier

If you want to look for **‘similar molecules’**: use Rhea

Exploiting small molecule data in UniProtKB

UniProtKB ▾ ChEBI:32395 Advanced | List Search

ChEBI identifier search - allows user to leverage the ontology to explore larger classes – arachidonate, polyunsaturated fatty acids, fatty acids, lipids and more.

UniProtKB ▾ inchikey:YZXBAPSDXZZRGB-DOFZRALJSA-M Advanced | List Search

InChIKey search – complete or partial.

Search is performed independently of the charge (microspecies at pH 7.3)

by default: hierarchical search

UniProtKB ▾ ChEBI:32395 Advanced | List Search

ChEBI identifier search – 12,611

[https://www.uniprot.org/uniprotkb?query=\(chebi:"CHEBI:32395"\)](https://www.uniprot.org/uniprotkb?query=(chebi:)

UniProtKB ▾ inchikey:YZXBAPSDXZZRGB-DOFZRALJSA-M Advanced | List Search

InChIKey search – complete – 12,611

<https://www.uniprot.org/uniprotkb?query=inchikey:YZXBAPSDXZZRGB-DOFZRALJSA-M>

UniProtKB ▾ inchikey:YZXBAPSDXZZRGB-DOFZRALJSA Advanced | List Search

InChIKey search – partial – 12,611

<https://www.uniprot.org/uniprotkb?query=inchikey:YZXBAPSDXZZRGB-DOFZRALJSA>

Charge: does not taken into account (microspecies at pH 7.3)

UniProt 2022_03

Exploiting small molecule data in UniProtKB

Proteins linked to arachidonate (using ChEBI ID):

[https://www.uniprot.org/uniprotkb?query=\(chebi:"CHEBI:32395"\)](https://www.uniprot.org/uniprotkb?query=(chebi:)

Human proteins linked to arachidonate (using NCBI TaxID):

[https://www.uniprot.org/uniprotkb?query=\(chebi:"CHEBI:32395"\) AND \(taxonomy_id:9606\)](https://www.uniprot.org/uniprotkb?query=(chebi:)

Human proteins linked to arachidonate (InChIKey variant):

[https://www.uniprot.org/uniprotkb?query=inchikey:YZXBAPSDXZZRGB AND \(taxonomy_id:9606\)](https://www.uniprot.org/uniprotkb?query=inchikey:YZXBAPSDXZZRGB AND (taxonomy_id:9606))

Human proteins linked to any lipid (using the ChEBI ontology):

[https://www.uniprot.org/uniprotkb?query=\(chebi:"CHEBI:18059"\) AND \(taxonomy_id:9606\)](https://www.uniprot.org/uniprotkb?query=(chebi:)

Exploiting ChEBI ontology in UniProtKB

Human proteins interacting with...

lipid (ChEBI:18059)



fatty acid anion (ChEBI:28868)



polyunsaturated fatty acid anion (ChEBI:76567)



arachidonate (CHEBI:32395)

- Reviewed (Swiss-Prot) (917)
- Unreviewed (TrEMBL) (1,310)
- Reviewed (Swiss-Prot) (254)
- Unreviewed (TrEMBL) (375)
- Reviewed (Swiss-Prot) (112)
- Unreviewed (TrEMBL) (100)
- Reviewed (Swiss-Prot) (75)
- Unreviewed (TrEMBL) (49)

(chebi:"CHEBI:XXXXX") AND (taxonomy_id:9606)

Exploiting small molecule data in UniProtKB

Human proteins linked to any lipid and with a 3D structure in PDB:

[https://www.uniprot.org/uniprotkb?query=\(chebi:"CHEBI:18059"\) AND \(taxonomy_id:9606\) AND \(database:PDB\)](https://www.uniprot.org/uniprotkb?query=(chebi:)

Human proteins linked to any lipid and found in the Golgi (experimental evidence):

[https://www.uniprot.org/uniprotkb?query=\(chebi:"CHEBI:18059"\) AND \(taxonomy_id:9606\) AND \(\(cc_scl_term:SL-0132\) AND \(ccev_scl_term:experimental\)\)](https://www.uniprot.org/uniprotkb?query=(chebi:)

<https://www.uniprot.org/help/query-fields>

UniProtKB and Rhea

How to mine enzyme data in UniProtKB ?

- (1) Look for caffeine in ChEBI (<https://www.ebi.ac.uk/chebi/>)
 - Look for its parent term (ChEBI ontology)
- (2) Look for enzymes involved in caffeine synthesis in UniProtKB
Query UniProtKB with 'Small molecule' using name, ChEBI ID or InChIKey and restrict your query to taxonomy Viridiplantae [33090]

Advanced Search

Searching in
UniProtKB

CHEBI ID
caffein Remove

Add Field

Type * in the search box to search for all values for the selected field.

- (3) Customize columns: add Rhea ID and Structure/3D

UniProtKB 11 results

BLAST Align Map IDs Download Add View: Cards Table **Customize columns** Share

- (4) Share your URL

UniProtKB 11 results

BLAST [Align](#) [Map IDs](#) [Download](#) [Add](#) View: [Cards](#) [Table](#) [Customize columns](#) [Share](#)

Entry	Entry Name	Protein Names	Gene Names	Organism	Rhea ID	3D structures
<input type="checkbox"/> Q9FZN8	TCS1_CAMSI	3,7-dimethylxanthine N-methyltransferase TCS1[...]	TCS1, TCS1A	Camellia sinensis (Tea)	24604 (UniProtKB Q Rhea E) 20944 (UniProtKB Q Rhea E) 10280 (UniProtKB Q Rhea E)	
<input type="checkbox"/> A4GE70	DXMT1_COFCA	3,7-dimethylxanthine N-methyltransferase[...]	DXMT, GSCOC_T00011062001	Coffea canephora (Robusta coffee)	20944 (UniProtKB Q Rhea E) 10280 (UniProtKB Q Rhea E) 24604 (UniProtKB Q Rhea E)	X-ray: 1
<input type="checkbox"/> Q8H0D2	DXMT1_COFAR	3,7-dimethylxanthine N-methyltransferase 1[...]	DXMT1, CS7	Coffea arabica (Arabian coffee)	20944 (UniProtKB Q Rhea E) 10280 (UniProtKB Q Rhea E) 24604 (UniProtKB Q Rhea E)	
<input type="checkbox"/> Q2HXI6	PCS1_CAMPL	7-methylxanthine methyltransferase PCS1[...]	PCS1, TCS1C	Camellia ptilophylla (Cocoa tea)	10280 (UniProtKB Q Rhea E) 24604 (UniProtKB Q Rhea E)	
<input type="checkbox"/> Q8H0D3	DXMT2_COFAR	3,7-dimethylxanthine N-methyltransferase 2[...]	DXMT2, CCS1, CS6	Coffea arabica (Arabian coffee)	24604 (UniProtKB Q Rhea E) 20944 (UniProtKB Q Rhea E) 10280 (UniProtKB Q Rhea E)	
<input type="checkbox"/> Q2HXL9	ICS1_CAMIR	7-methylxanthine methyltransferase ICS1[...]	ICS1, TCS1B	Camellia irrawadiensis (Burmese tea)	10280 (UniProtKB Q Rhea E) 20944 (UniProtKB Q Rhea E) 24604 (UniProtKB Q Rhea E)	
<input type="checkbox"/> A0A052PM92	TCS1F_CAMCR	Caffeine synthase 1[...]	TCS1F	Camellia crassicolumna (Evergreen tea)	20944 (UniProtKB Q Rhea E) 24604 (UniProtKB Q Rhea E)	
<input type="checkbox"/> A0A052PMA8	TCS1D_CAMTA	Caffeine synthase 1[...]	TCS1D	Camellia taliensis (Wild tea)	20944 (UniProtKB Q Rhea E) 24604 (UniProtKB Q Rhea E)	
<input type="checkbox"/> A0A6C0WW38	CKCS_CAMSB	3,7-dimethylxanthine N-methyltransferase TCS1[...]	CS	Camellia sinensis var. assamica (Assam tea) (Thea assamica)	10280 (UniProtKB Q Rhea E) 20944 (UniProtKB Q Rhea E) 24604 (UniProtKB Q Rhea E)	
<input type="checkbox"/> Q68CM3	TCS2_CAMSI	Probable caffeine synthase 2[...]	TCS2	Camellia sinensis (Tea)	24604 (UniProtKB Q Rhea E) 20944 (UniProtKB Q Rhea E) 10280 (UniProtKB Q Rhea E)	
<input type="checkbox"/> A0A052PM82	TCS3_CAMSI	Caffeine synthase 3[...]	TCS1E	Camellia sinensis (Tea)	20944 (UniProtKB Q Rhea E) 24604 (UniProtKB Q Rhea E)	

[https://www.uniprot.org/uniprotkb?query=\(chebi:"CHEBI:27732"\) AND \(taxonomy_id:33090\)&fields=accession,reviewed,id,protein_name,gene_names,organism_name,rhea,structure_3d&view=table](https://www.uniprot.org/uniprotkb?query=(chebi:)

UniProtKB, protein sequence databases and sequence annotation

Protein sequence and annotation: overview

Nucleic acid sequence databases

INSDC, Ensembl, RefSeq

UniProtKB

UniProtKB/Swiss-Prot

UniProtKB/TrEMBL

Biochemical data (Rhea & ChEBI)

Gene Ontology

UniProt web sites and tools

NCBI Protein sequence databases

RefSeq

How often do you use Gene Ontology ?

by a guest · just now

Make a choice:

- daily
- once a month
- less
- never

Vote

 Results

 Share

<https://strawpoll.com/polls/3RnYp7ENAYe>

Gene Ontology

- The Gene Ontology is a controlled vocabulary, a set of standard terms—words and phrases—used for indexing and retrieving information. *Same vocabulary for all species*
- GO defines the relationships between the terms (**hierarchy**), making it a structured vocabulary.
- Created by the GO consortium
- Contains ~43'000 terms.
- **Standardization of biological data/information**

GO:0009535   JSON

chloroplast thylakoid membrane


Cellular Component

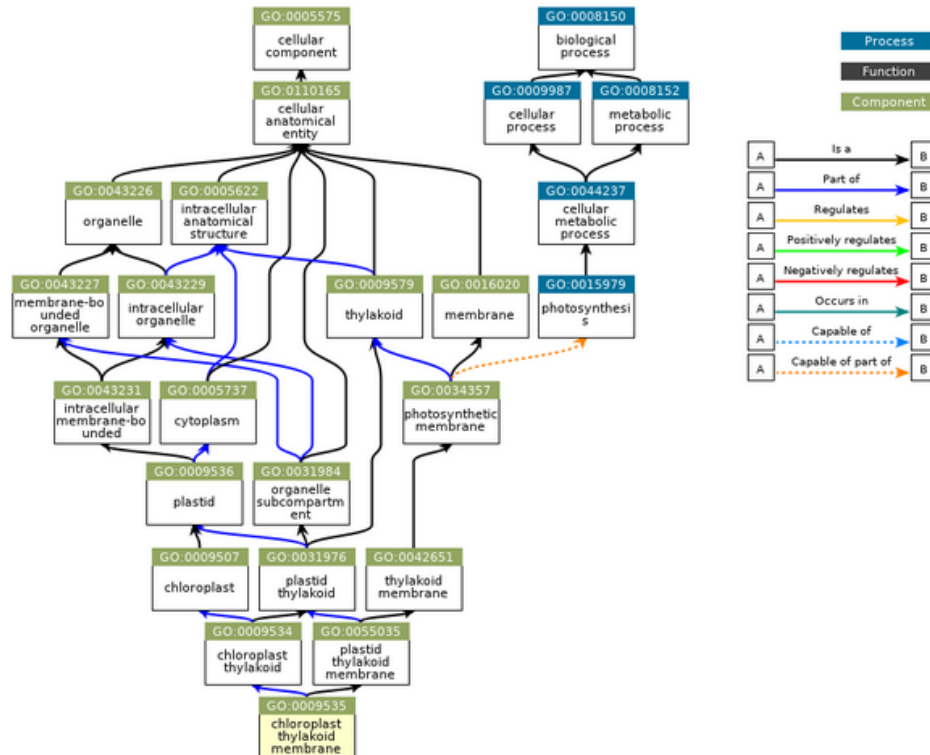
Definition ([GO:0009535 GONUTS page](#))

The pigmented membrane of a chloroplast thylakoid. An example of this component is found in *Arabidopsis thaliana*.

655,738 annotations

Ancestor Chart

Ancestor chart for GO:0009535 



Category

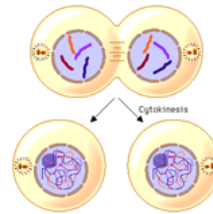
Definition

Hierarchy

3 categories of GO terms

1. Biological Process

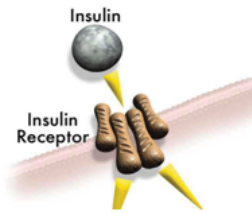
A commonly recognized series of events



- Cell division
- Mitosis
- Organelle fission

2. Molecular Function

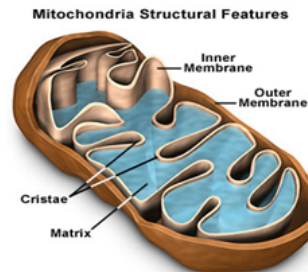
An elemental activity or task or job



- Protein kinase activity
- Insulin binding
- Insulin receptor activity

3. Cellular Component

Where a gene product is located



- Mitochondrion
- Mitochondrial matrix
- Mitochondrial membrane



Current release 2022-07-01: 43,558 GO terms | 7,483,496 annotations
1,480,259 gene products | 5,213 species (see statistics)

THE GENE ONTOLOGY RESOURCE

The mission of the GO Consortium is to develop a comprehensive, **computational model of biological systems**, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life.

The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.

- Any
- Ontology
- Gene Product

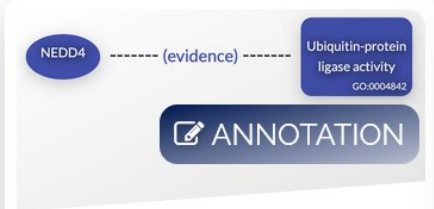
GO Enrichment Analysis

Powered by PANTHER

Hint: can use UniProt ID/AC, Gene Name, Gene Symbols, MOD IDs



The network of biological classes describing



Statements, based on specific, traceable



GO Causal Activity Model (GO-CAM) provides



Tools to curate, browse, search, visualize and

<http://www.geneontology.org>



Gene Ontology (GO) annotation

Gene products and species

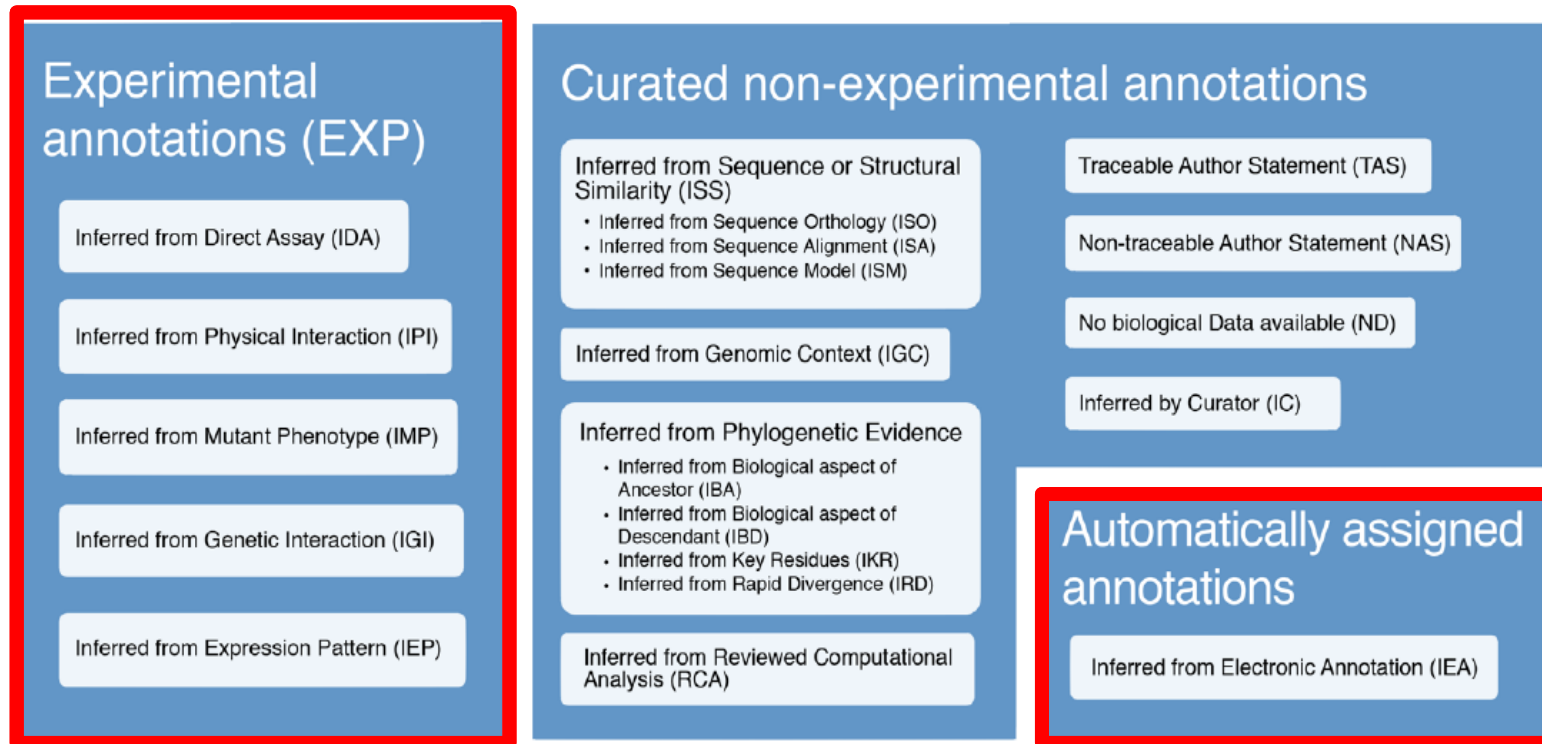
Property	Value
Annotated gene products	1,480,259
Annotated species	5,213
Annotated species with over 1,000 annotations	185

<http://geneontology.org/stats.html>

Gene Ontology (GO) annotation

- Annotation is the process of assigning/mapping GO terms to gene products...
- Automatically assigned vs Experimental (manual) annotation... (IEA vs EXP)

Annotations are supported by Evidence codes



EXP
~0.1 %

QuickGO (oct 2022)
Total : 1,023,603,628 annotations
IEA : 1,017,112,631 (> 99 %)
Non-IEA : 6,000,000 million (< 1%)

- IBA : 4,000,000 (0.4 %)
- EXP : 1,174,524 (0.1 %)

IEA
>99 %

Annotations are supported by Evidence codes

Experimental annotations (EXP)

Inferred from Direct Assay (IDA)

Inferred from Physical Interaction (IPI)

Inferred from Mutant Phenotype (IMP)

Inferred from Genetic Interaction (IGI)

Inferred from Expression Pattern (IEP)

Curated non-experimental annotations

Inferred from Sequence or Structural Similarity (ISS)

- Inferred from Sequence Orthology (ISO)
- Inferred from Sequence Alignment (ISA)
- Inferred from Sequence Model (ISM)

Inferred from Genomic Context (IGC)

Inferred from Phylogenetic Evidence

- Inferred from Biological aspect of Ancestor (IBA)
- Inferred from Biological aspect of Descendant (IBD)
- Inferred from Key Residues (IKR)
- Inferred from Rapid Divergence (IRD)

Inferred from Reviewed Computational Analysis (RCA)

Traceable Author Statement (TAS)

Non-traceable Author Statement (NAS)

No biological Data available (ND)

Inferred by Curator (IC)

Automatically assigned annotations

Inferred from Electronic Annotation (IEA)

EXP

~0.1 %

GO annotation: manual (experimental)

PUBMED 1: Protein kinase R (PKR) recognizes double stranded RNA in the cytoplasm

PUBMED 2: PKR acts as a kinase in the nucleus

GO annotation:

Molecular Function: -double stranded RNA binding (PUBMED 1)
-kinase activity (PUBMED 2)

Cellular Component: -nucleus (PUBMED 2)
-cytoplasm (PUBMED 1)

Experimental GO annotation is provided by different databases, including UniProtKB

Number of annotated scientific publications	162,935
---	---------

UniProt provides 60 % of experimental GO annotation

Annotations are supported by Evidence codes

Experimental annotations (EXP)

Inferred from Direct Assay (IDA)

Inferred from Physical Interaction (IPI)

Inferred from Mutant Phenotype (IMP)

Inferred from Genetic Interaction (IGI)

Inferred from Expression Pattern (IEP)

Curated non-experimental annotations

Inferred from Sequence or Structural Similarity (ISS)

- Inferred from Sequence Orthology (ISO)
- Inferred from Sequence Alignment (ISA)
- Inferred from Sequence Model (ISM)

Inferred from Genomic Context (IGC)

Inferred from Phylogenetic Evidence

- Inferred from Biological aspect of Ancestor (IBA)
- Inferred from Biological aspect of Descendant (IBD)
- Inferred from Key Residues (IKR)
- Inferred from Rapid Divergence (IRD)

Inferred from Reviewed Computational Analysis (RCA)

Traceable Author Statement (TAS)

Non-traceable Author Statement (NAS)

No biological Data available (ND)

Inferred by Curator (IC)

Automatically assigned annotations

Inferred from Electronic Annotation (IEA)

IEA
>99 %

Automatically assigned GO annotation

Inferred from Electronic Annotation (IEA) : > 98 %

External Mappings

- InterPro2GO
- EC2GO
- SwissProt Keywords
- UniProtKB Subcellular Localization

Automated Annotation by Orthology

- Ensembl Compara

F Zinc finger protein NOA36^{*} IPR010531

InterPro entry

Overview

Proteins	2k
Domain Architectures	22
Taxonomy	5k
Proteomes	777
RoseTTAFold	1
AlphaFold	2k
Genome3D	10

Short name: NOA36

Description

This family consists of several NOA36 proteins (also known as zinc finger protein 330) which contain 29 highly conserved cysteine residues. In mitosis it associates with centromeres and concentrates at the midbody in cytokinesis [1].

GO terms

Biological Process

None

Molecular Function

- zinc ion binding (GO:0008270)

Cellular Component



- nucleus (GO:0005634)

References

1. Molecular cloning of a zinc finger autoantigen transiently associated with nucleolus and mitotic centromeres and midbodies. These proteins with nine CXXC motifs are highly conserved from nematodes to humans. Bolivar J, Diaz I, Iglesias C, et al. *J. Biol. Chem.* 274, 36456-64, (1999). [View article](#)

Add your annotation

Contributing Member Database Entries

 Pfam: PF06524	 PANTHER: PTHR13214
--	---

Human Proteome & GO

In 2022:

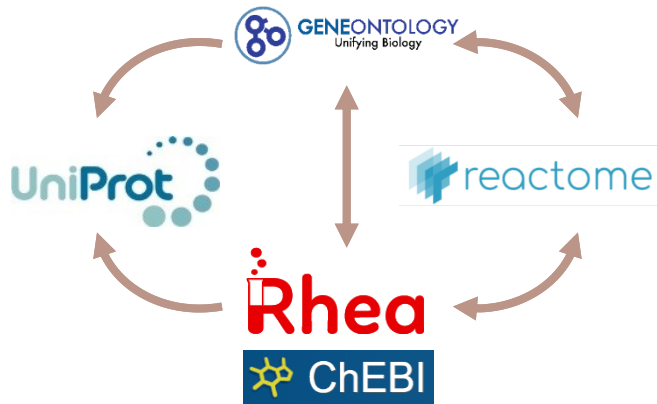
Complete GO annotation of human proteins containing a domain (PANTHER classification system).



The mission of the PANTHER knowledgebase is to support biomedical and other research by providing **comprehensive information about the evolution of protein-coding gene families**, particularly protein phylogeny, function and genetic variation impacting that function. [Learn more](#)

Cover 80 % of human genes.

Rhea2GO



```
! Generated on 2022-07-01T08:03Z from the ontology 'go' with data version: 'releases/2022-07-01'
!
RHEA:10000 > GO:pentanamidase activity ; GO:0050168
RHEA:10004 > GO:thiocyanate isomerase activity ; GO:0050335
RHEA:10008 > GO:peroxiredoxin activity ; GO:0051920
RHEA:10012 > GO:(R)-6-hydroxynicotine oxidase activity ; GO:0018530
RHEA:10016 > GO:sinapine esterase activity ; GO:0050285
RHEA:10020 > GO:saccharopine dehydrogenase (NADP+, L-glutamate-forming) activity ; GO:0004755
RHEA:10024 > GO:histone-lysine N-methyltransferase activity ; GO:0018024
RHEA:10028 > GO:D-glutamate oxidase activity ; GO:0047821
RHEA:10032 > GO:peptidyl-glutaminase activity ; GO:0050170
RHEA:10036 > GO:vanillyl-alcohol oxidase activity ; GO:0018465
RHEA:10040 > GO:3'(2'),5'-bisphosphate nucleotidase activity ; GO:0008441
RHEA:10044 > GO:cyclohexanol dehydrogenase activity ; GO:0018460
RHEA:10048 > GO:O-acetylhomoserine aminocarboxypropyltransferase activity ; GO:0003961
RHEA:10056 > GO:3-aminobutyryl-CoA ammonia-lyase activity ; GO:0047459
RHEA:10060 > GO:D-tryptophan N-acetyltransferase activity ; GO:0047835
RHEA:10064 > GO:D-alanine 2-hydroxymethyltransferase activity ; GO:0050413
RHEA:10068 > GO:poly(ribitol-phosphate) beta-glucosyltransferase activity ; GO:0047266
RHEA:10072 > GO:dimethylglycine N-methyltransferase activity ; GO:0052729
RHEA:10076 > GO:lactase activity ; GO:0000016
RHEA:10080 > GO:phloroglucinol reductase activity ; GO:0018510
RHEA:10084 > GO:protocatechuate 3,4-dioxygenase activity ; GO:0018578
RHEA:10088 > GO:sucrose 6F-alpha-galactotransferase activity ; GO:0047235
RHEA:10092 > GO:deoxynucleotide 3'-phosphatase activity ; GO:0047846
RHEA:10096 > GO:N-acyl homoserine lactone synthase activity ; GO:0061579
RHEA:10100 > GO:sterol esterase activity ; GO:0004771
RHEA:10104 > GO:gibberellin 3-beta-dioxygenase activity ; GO:0016707
RHEA:10108 > GO:salutaridine reductase (NADPH) activity ; GO:0047037
RHEA:10112 > GO:cystathionine beta-synthase activity ; GO:0004122
RHEA:10116 > GO:phosphogluconate dehydrogenase (decarboxylating) activity ; GO:0004616
RHEA:10124 > GO:phenylacetone monooxygenase activity ; GO:0033776
RHEA:10128 > GO:neoxanthin synthase activity ; GO:0034020
RHEA:10132 > GO:benzoate-CoA ligase activity ; GO:0018858
RHEA:10136 > GO:tartronate-semialdehyde synthase activity ; GO:0009028
RHEA:10140 > GO:prostaglandin-F synthase activity ; GO:0047017
RHEA:10144 > GO:3'-nucleotidase activity ; GO:0008254
RHEA:10148 > GO:blasticidin-S deaminase activity ; GO:0047711
RHEA:10152 > GO:carbamate kinase activity ; GO:0008804
```

Rhea2GO mapping: <http://current.geneontology.org/ontology/external2go/rhea2go>

with Harold Drabkin, Pascale Gaudet, Chris Mungall, Peter D'Eustachio, Paul Thomas, David Hill, Alex Ignatchenko

Gene Ontology (GO) some caveats...

1

Lack of consistency between species

Entry	Entry Name	Protein Names	Gene Names	Organism
P01588	EPO_HUMAN	Erythropoietin[...]	EPO	Homo sapiens (Human)
P07321	EPO_MOUSE	Erythropoietin	Epo	Mus musculus (Mouse)
P29676	EPO_RAT	Erythropoietin	Epo	Rattus norvegicus (Rat)
P48617	EPO_BOVIN	Erythropoietin	EPO	Bos taurus (Bovine)
Q2XNF5	EPO_DANRE	Erythropoietin[...]	epo	Danio rerio (Zebrafish) (Brachydanio rerio)

Gene Ontology - Biological Process

- acute-phase response
- aging
- blood circulation
- cellular hyperosmotic response
- embryo implantation
- More terms

- acute-phase response
- aging
- angiogenesis
- apoptotic process
- cardiac muscle tissue morphogenesis
- More terms

- acute-phase response
- aging
- apoptotic process
- cellular hyperosmotic response
- embryo implantation
- More terms

- cellular hyperosmotic response
- embryo implantation
- erythrocyte differentiation
- erythrocyte maturation
- erythropoietin-mediated signaling pathway
- More terms

- erythrocyte maturation
- hemopoiesis
- nucleate erythrocyte development
- pronephros development
- response to activity

Gene Ontology - Cellular Component

- cell body
- cell surface
- extracellular region
- extracellular space

- cell body
- cell surface
- extracellular region
- extracellular space

- cell body
- cell surface
- extracellular space

- cell surface
- extracellular space

- extracellular space
- perinuclear region of cytoplasm

Gene Ontology - Molecular Function

- cytokine activity
- erythropoietin receptor binding
- hormone activity
- protein kinase activator activity

- cytokine activity
- erythropoietin receptor binding
- hormone activity
- protein kinase activator activity

- cytokine activity
- erythropoietin receptor binding
- hormone activity
- protein kinase activator activity

- cytokine activity
- erythropoietin receptor binding
- hormone activity
- protein kinase activator activity

- cytokine activity
- erythropoietin receptor binding
- hormone activity

PAN GO project

The **Phylogenetic Annotation Project** performs annotation inferences across evolutionary related proteins based on known function of proteins within PANTHER phylogenetic family trees.

https://wiki.geneontology.org/Phylogenetic_Annotation_Project

2

Each database has its own 'rules' to import GO annotation from the GO databases

-> different set of GO terms depending of the database

UniProtKB 2 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism
<input type="checkbox"/> P01308	INS_HUMAN	Insulin	INS	Homo sa
<input type="checkbox"/> A6XGL2	A6XGL2_HUMAN	Insulin	INS	Homo sa

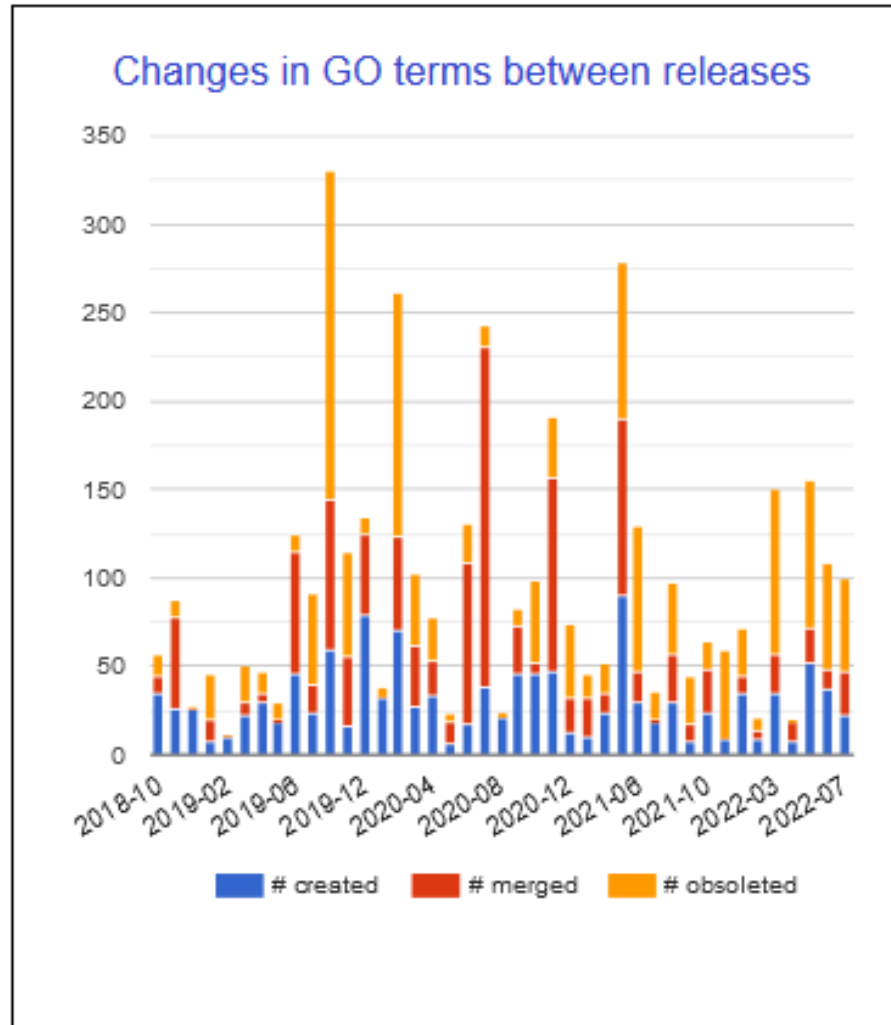
Gene Ontology - Biological Process

- [activation of protein kinase B activity](#)
- [acute-phase response](#)
- [alpha-beta T cell activation](#)
- [cell-cell signaling](#)
- [cognition](#)
- [More terms](#)

glucose metabolic process

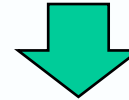
Cellular Component	Gene Ontology - Molecular Function
culum lumen	hormone activity
culum-Golgi intermediate compartment membrane	identical protein binding
ion	insulin receptor binding
ion	insulin-like growth factor receptor binding
ion	protease binding
ion	hormone activity

3



@UniProtKB: look for the entries corresponding to gene **ygbT**

UniProtKB 259 results







BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Add column GO molecular function, biological process and cellular component

Compare the GO annotation for the Swiss-Prot and the corresponding TrEMBL entries

Help links for Customize columns (old UniProtKB design videos):

- <https://www.uniprot.org/help/customize>
- <http://insideuniprot.blogspot.com/2015/03>
- <https://www.youtube.com/watch?v=ado1r8IDm3U> (at about 1'10)

Entry	Entry Name	Protein Names	Gene Names	Organism	Length	Gene Ontology - Biological Process	Gene Ontology - Cellular Component	Gene Ontology - Molecular Function
<input type="checkbox"/> Q46896	 CAS1_ECOLI	CRISPR-associated endonuclease Cas1[...]	ygbT, cas1, b2755, JW2725	Escherichia coli (strain K12)	305 AA	cellular response to DNA damage stimulus ↗ CRISPR-cas system ↗ defense response to virus ↗ DNA repair ↗ maintenance of CRISPR repeat elements ↗	cytoplasm ↗	5'-flap endonuclease activity ↗ crossover junction endodeoxyribonuclease activity ↗ DNA binding ↗ identical protein binding ↗ metal ion binding ↗
<input type="checkbox"/> A0A1E3GUF3	 A0A1E3GUF3_9HYPH	CRISPR-associated endonuclease Cas1[...]	ygbT, cas1, A6302_04375	Methylobrevis pamukkalensis	326 AA	defense response to virus ↗ maintenance of CRISPR repeat elements ↗		DNA binding ↗ endodeoxyribonuclease activity ↗ metal ion binding ↗
<input type="checkbox"/> A0A380MR50	 A0A380MR50_9GAMM	CRISPR-associated endonuclease Cas1[...]	ygbT, cas1, NCTC13337_01079	Suttonella ornithocola	325 AA	defense response to virus ↗ maintenance of CRISPR repeat elements ↗	integral component of membrane ↗	DNA binding ↗ endodeoxyribonuclease activity ↗ metal ion binding ↗
<input type="checkbox"/> A0A486VSH6	 A0A486VSH6_KLEPN	CRISPR-associated endonuclease Cas1[...]	ygbT, cas1, SAMEA4873563_03653	Klebsiella pneumoniae	306 AA	defense response to virus ↗ maintenance of CRISPR repeat elements ↗		DNA binding ↗ endodeoxyribonuclease activity ↗ metal ion binding ↗

Feedback 

UniProtKB, protein sequence databases and sequence annotation

Protein sequence and annotation: overview

Nucleic acid sequence databases

INSDC, Ensembl, RefSeq

UniProtKB

UniProtKB/Swiss-Prot

UniProtKB/TrEMBL

Biochemical data (Rhea & ChEBI)

Gene Ontology

UniProt web sites and tools

NCBI Protein sequence databases

RefSeq

UniProt website and tools

The UniProt web site – www.uniprot.org



- Powerful search engine, **google-like** and **easy-to-use**, but also supports very directed field searches
- Entry views, search result views and downloads are **customizable**
- The URL of a result page reflects the query; all pages and queries are **bookmarkable**, supporting programmatic access
- Tools: Blast, Align, Retrieve/ID mapping, Peptide search

Find your protein

UniProtKB ▾ [Advanced](#) | [List](#) [Search](#)

Examples: [Insulin](#), [APP](#), [Human](#), [P05067](#), [organism_id:9606](#)

UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. [Cite UniProt](#)™

Advanced Search ✕

Searching in
UniProtKB

Gene Name [GN] Remove

AND Taxonomy [OC] Q Remove

AND Keyword [KW] Q Remove

AND All Remove

[Add Field](#) Cancel Search

i Type * in the search box to search for all values for the selected field.

Release 2022_03 | Statistics 3 [Help](#)

ein

Advanced | List Search

[Feedback](#)

[Help](#)

protein sequence and functional information. [Cite UniProt](#)

AND / OR / NOT



Result pages: highly customizable (also available for Blast)

UniProtKB (taxonomy_id:9606) Advanced | List Search Help

Status

- Reviewed (Swiss-Prot) (20,399)
- Unreviewed (TrEMBL) (184,651)

Popular organisms

- Human (204,961)

Taxonomy

- 9606 ×

Filter by taxonomy

Proteins with

- 3D structure (7,818)
- Active site (4,043)
- Activity regulation (1,569)
- Allergen (6)
- Alternative products (isoforms) (10,635)

More items

Protein existence

UniProtKB 205,050 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
<input type="checkbox"/> A0A0C5B5G6	MOTSC_HUMAN	Mitochondrial-derived peptide MOTS-c[...]	MT-RNR1	Homo sapiens (Human)	16 AA
<input type="checkbox"/> A0A1B0GTW7	CIROP_HUMAN	Ciliated left-right organizer metallopeptidase[...]	CIROP, LMLN2	Homo sapiens (Human)	788 AA
<input type="checkbox"/> A0JNW5	UH1BL_HUMAN	UHRF1-binding protein 1-like[...]	UHRF1BP1L, KIAA0701, SHIP164	Homo sapiens (Human)	1,464 AA
<input type="checkbox"/> A0JP26	POTB3_HUMAN	POTE ankyrin domain family member B3	POTEB3	Homo sapiens (Human)	581 AA
<input type="checkbox"/> A0PK11	CLRN2_HUMAN	Clarin-2	CLRN2	Homo sapiens (Human)	232 AA
<input type="checkbox"/> A1A4S6	RHG10_HUMAN	Rho GTPase-activating protein 10[...]	ARHGAP10, GRAF2	Homo sapiens (Human)	786 AA
<input type="checkbox"/> A1A519	F170A_HUMAN	Protein FAM170A[...]	FAM170A, ZNFD	Homo sapiens (Human)	330 AA
<input type="checkbox"/> A1L190	SYCE3_HUMAN	Synaptonemal complex central element protein 3 [...]	SYCE3, C22orf41, THEG2	Homo sapiens (Human)	88 AA

Customize columns

Data **8** External links

DNA binding pH dependence

Miscellaneous 1

Interaction

Expression

Gene Ontology (GO)

Pathology & Biotech

Subcellular location 1

Intramembrane Topological domain

Subcellular location [CC] Transmembrane

PTM / Processing

Chain Initiator methionine Post-translational modification

Cross-link Lipidation Propeptide

Disulfide bond Modified residue Signal peptide

Glycosylation Peptide Transit peptide

Structure

3D Helix

[Reset to default](#) [Close](#)

Advanced | List Search Help

Customize columns Share

Accession	Organism	Length	Rhea ID	Subcellular Location
LN2	Homo sapiens (Human)	16 AA		Secreted Mitochondrion Nucleus Translocates to nucleus in a manner. to metabolize an AMPK-c
LN2	Homo sapiens (Human)	788 AA		Membrane pass type I membrane prote
LN2, SHIP164	Homo sapiens (Human)	1,464 AA		Cytoplasm, cyto Early endosome Primarily cytos Recruited to ear endosomes follo STX6 overexpre: Overexpression

[Feedback](#) [Help](#)

Creating URL for the UniProt web site REST API

UniProtKB (taxonomy_id:9606) Advanced | List Search

UniProtKB 205,050 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Length	Rhea ID	Subcellular Location
<input type="checkbox"/> A0A0C5B5G6	MOTSC_HUMAN	Mitochondrial-derived peptide MOTS-c[...]	MT-RNR1	Homo sapiens (Human)	16 AA		Secreted Mitochondrion Nucleus Translocates to nucleus in response to metabolic stress in an AMPK-dependent manner.
<input type="checkbox"/> A0A1B0GTW7	CIROP_HUMAN	Ciliated left-right organizer metalloproteinase[...]	CIROP, LMLN2	Homo sapiens (Human)	788 AA		Membrane protein
<input type="checkbox"/> A0JNW5	UH1BL_HUMAN	UHRF1-binding protein 1-like[...]	UHRF1BP1L, KIAA0701, SHIP164	Homo sapiens	1,464 AA		Cytoplasm, cytosol, Early endosome

`https://www.uniprot.org/uniprotkb?query=(taxonomy_id:9606)&fields=accession,reviewed,id,protein_name,gene_names,organism_name,length,rhea,cc_subcellular_location&view=table`

What is the query corresponding at this URL ?

[https://www.uniprot.org/uniprotkb?query=%28taxonomy id%3A9606%29+AND+%28proteome%3AUP000005640%29&fields=accession%2Creviewed%2Cid%2Cprotein name%2Cgene names%2Corganism name%2Ccc function%2Ccc disease&view=table](https://www.uniprot.org/uniprotkb?query=%28taxonomy%20id%3A9606%29+AND+%28proteome%3AUP000005640%29&fields=accession%2Creviewed%2Cid%2Cprotein%2Cgene%2Cnames%2Corganism%2Cname%2Ccc%2Cfunction%2Ccc%2Cdisease&view=table)

Status

- Reviewed (Swiss-Prot) (20,383)
- Unreviewed (TrEMBL) (59,357)

Popular organisms

- Human (79,740)

Taxonomy

9606 ✕

[Filter by taxonomy](#)

Proteins with

- 3D structure (7,613)
- Active site (2,893)
- Activity regulation (1,526)
- Allergen (6)
- Alternative products (isoforms) (10,635)
- More items

Protein existence

- Protein level (66,272)
- Predicted (8,160)
- Transcript level (2,571)
- Homology (2,128)
- Uncertain (609)

Annotation score

- 5 (14,334)
- 4 (2,232)
- 3 (3,886)
- 2 (13,294)
- 1 (45,994)

UniProtKB 79,740 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Function	Disease Involvement
<input type="checkbox"/> A0A0C5B5G6	MOTSC_HUMAN	Mitochondrial-derived peptide MOTS-c[-]	MT-RNR1	Homo sapiens (Human)	<p>Regulates insulin sensitivity and metabolic homeostasis (PubMed:25738459, PubMed:33468709).</p> <p>Inhibits the folate cycle, thereby reducing de novo purine biosynthesis which leads to the accumulation of the de novo purine synthesis intermediate 5-aminoimidazole-4-carboxamide (AICAR) and the activation of the metabolic regulator 5'-AMP-activated protein kinase (AMPK) (PubMed:25738459).</p> <p>Protects against age-dependent and diet-induced insulin resistance as well as diet-induced obesity (PubMed:25738459).</p> <p>In response to metabolic stress, translocates to the nucleus where it binds to antioxidant response elements (ARE) present in the promoter regions of a number of genes and plays a role in regulating nuclear gene expression in an NFE2L2-dependent manner and increasing cellular resistance to metabolic stress (PubMed:29983246).</p> <p>Increases mitochondrial respiration and levels of CPT1A and cytokines IL1B, IL6, IL8, IL10 and TNF in senescent cells (PubMed:29886458).</p> <p>Increases activity of the serine/threonine protein kinase complex mTORC2 and reduces activity of the PTEN phosphatase, thus promoting phosphorylation of AKT (PubMed:33554779).</p> <p>This promotes AKT-mediated phosphorylation of transcription factor FOXO1 which reduces FOXO1 activity, leading to reduced levels of MSTN and promotion of skeletal muscle growth (PubMed:33554779).</p> <p>Promotes osteogenic differentiation of bone marrow mesenchymal stem cells via the TGFβ/SMAD pathway (PubMed:30468456).</p> <p>Promotes osteoblast proliferation and osteoblast synthesis of type I collagens COL1A1 and COL1A2 via the TGFβ/SMAD pathway (PubMed:31081069).</p>	
<input type="checkbox"/> A0A1B0GTW7	CIROP_HUMAN	Ciliated left-right organizer metalloproteinase[-]	CIROP, LMLN2	Homo sapiens (Human)	Putative metalloproteinase that plays a role in left-right patterning process.	<p>Heterotaxy, visceral, 12, autosomal (HTX12)</p> <p>Note The disease is caused by variants affecting the gene represented in this entry</p> <p>Description A form of visceral heterotaxy, a complex disorder due to disruption of the normal left-right asymmetry of the thoracoabdominal organs. Visceral heterotaxy or situs ambiguus results in randomization of the placement of visceral organs, including the heart, lungs, liver, spleen, and stomach. The organs are oriented randomly with</p>

Feedback

Help

Status

- Reviewed (Swiss-Prot) (20,383)
- Unreviewed (TrEMBL) (59,357)

Popular organisms

- Human (79,740)

Taxonomy

9606

Filter by taxonomy

Proteins with

- 3D structure (7,613)
- Active site (2,893)
- Activity regulation (1,526)
- Allergen (6)
- Alternative products (isoforms) (10,635)
- More items

Protein existence

- Protein level (66,272)
- Predicted (8,160)
- Transcript level (2,571)
- Homology (2,128)
- Uncertain (609)

Annotation score

- 5 (14,334)
- 4 (2,232)
- 3 (3,886)
- 2 (13,294)
- 1 (45,994)

UniProtKB 79,740 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Function	Disease Involvement
<input type="checkbox"/> A0A0C5B5G6	MOTSC_HUMAN	Mitochondrial-derived peptide MOTS-c[-]	MT-RNR1	Homo sapiens (Human)	<p>Regulates insulin sensitivity and metabolic homeostasis (PubMed:25738459, PubMed:33468709).</p> <p>Inhibits the folate cycle, thereby reducing de novo purine biosynthesis which leads to the accumulation of the de novo purine synthesis intermediate 5-aminoimidazole-4-carboxamide (AICAR) and the activation of the metabolic regulator 5'-AMP-activated protein kinase (AMPK) (PubMed:25738459).</p> <p>Protects against age-dependent and diet-induced insulin resistance as well as diet-induced obesity (PubMed:25738459).</p> <p>In response to metabolic stress, translocates to the nucleus where it binds to antioxidant response elements (ARE) present in the promoter regions of a number of genes and plays a role in regulating nuclear gene expression in an NFE2L2-dependent manner and increasing cellular resistance to metabolic stress (PubMed:29983246).</p> <p>Increases mitochondrial respiration and levels of CPT1A and cytokines IL1B, IL6, IL8, IL10 and TNF in senescent cells (PubMed:29886458).</p> <p>Increases activity of the serine/threonine protein kinase complex mTORC2 and reduces activity of the PTEN phosphatase, thus promoting phosphorylation of AKT (PubMed:33554779).</p> <p>This promotes AKT-mediated phosphorylation of transcription factor FOXO1 which reduces FOXO1 activity, leading to reduced levels of MSTN and promotion of skeletal muscle growth (PubMed:33554779).</p> <p>Promotes osteogenic differentiation of bone marrow mesenchymal stem cells via the TGFβ/SMAD pathway (PubMed:30468456).</p> <p>Promotes osteoblast proliferation and osteoblast synthesis of type I collagens COL1A1 and COL1A2 via the TGFβ/SMAD pathway (PubMed:31081069).</p>	
<input type="checkbox"/> A0A1B0GTW7	CIROP_HUMAN	Ciliated left-right organizer metallopeptidase[-]	CIROP, LMLN2	Homo sapiens (Human)	Putative metalloproteinase that plays a role in left-right patterning process.	<p>Heterotaxy, visceral, 12, autosomal (HTX12)</p> <p>Note The disease is caused by variants affecting the gene represented in this entry</p> <p>Description A form of visceral heterotaxy, a complex disorder due to disruption of the normal left-right asymmetry of the thoracoabdominal organs. Visceral heterotaxy or situs ambiguus results in randomization of the placement of visceral organs, including the heart, lungs, liver, spleen, and stomach. The organs are oriented randomly with</p>

Download (different formats)

Download

Download selected (0)
Download all (79,740)

Format: Excel

Compressed: Yes

Customize columns: Reviewed, Entry Name, Protein names, Gene Names, Organism, Function [CC], Involvement in disease

Data 7 External links

Search for available columns

- Names & Taxonomy
- Sequences
- Function
- Miscellaneous
- Interaction
- Expression
- Gene Ontology (GO)
- Pathology & Biotech
- Subcellular location
- PTM / Processing
- Structure
- Publications
- Date of
- Family & Domains

Generate URL for API Preview 10 Cancel Download

5640

Advanced | List Search

Share

Organism	Function	Disease Involvement
Homo sapiens (Human)	Regulates insulin sensitivity and metabolic homeostasis (PubMed:25738459, PubMed:33468709). Inhibits the folate cycle, thereby reducing de novo purine biosynthesis which leads to the accumulation of the de novo purine synthesis intermediate 5-aminimidazole-4-carboxamide (AICAR) and the activation of the metabolic regulator 5'-AMP-activated protein kinase (AMPK) (PubMed:25738459). Protects against age-dependent and diet-induced insulin resistance as well as diet-induced obesity (PubMed:25738459). In response to metabolic stress, translocates to the nucleus where it binds to antioxidant response elements (ARE) present in the promoter regions of a number of genes and plays a role in regulating nuclear gene expression in an NFE2L2-dependent manner and increasing cellular resistance to metabolic stress (PubMed:29983246). Increases mitochondrial respiration and levels of CPT1A and cytokines IL1B, IL6, IL8, IL10 and TNF in senescent cells (PubMed:29886458). Increases activity of the serine/threonine protein kinase complex mTORC2 and reduces activity of the PTEN phosphatase, thus promoting phosphorylation of AKT (PubMed:33554779). This promotes AKT-mediated phosphorylation of transcription factor FOXO1 which reduces FOXO1 activity, leading to reduced levels of MSTN and promotion of skeletal muscle growth (PubMed:33554779). Promotes osteogenic differentiation of bone marrow mesenchymal stem cells via the TGF β /SMAD pathway (PubMed:30468456). Promotes osteoblast proliferation and osteoblast synthesis of type I collagens COL1A1 and COL1A2 via the TGF β /SMAD pathway (PubMed:31081069).	
Homo sapiens (Human)	Putative metalloproteinase that plays a role in left-right patterning process.	Heterotaxy, visceral, 12, autosomal (HTX12) Note The disease is caused by variants affecting the gene represented in this entry Description A form of visceral heterotaxy, a complex disorder due to disruption of the normal left-right asymmetry of the thoracoabdominal organs. Visceral heterotaxy or situs ambiguus results in randomization of the placement of visceral organs, including the heart, lungs, liver, spleen, and stomach. The organs are oriented randomly with

Highlight sequence annotation in alignment (multiple alignment)

Align results

Overview Trees Percent Identity Matrix Text Output Input Parameters API Request

BLAST Align Map IDs Download Add Resubmit

Highlight properties Select annotation View: Overview Wrapped

<input type="checkbox"/> sp P67970 INS_CHICK	MALWIRSLPLLALLVFSGPGTSYA	AANQHLCGSHLVEALYLVCGERGFFYS	PKARRDVEEQPLVSSP	66
<input type="checkbox"/> sp P01317 INS_BOVIN	MALWTRLRPLLALLALWPPPPARA	FVNQHLCGSHLVEALYLVCGERGFFYT	PKARREVEEGPQVGAL	66
<input type="checkbox"/> sp P01315 INS_PIG	MALWTRLRPLLALLALWAPAPAQA	FVNQHLCGSHLVEALYLVCGERGFFYT	PKARREAENPQAGAV	66
<input type="checkbox"/> sp P01321 INS_CANLF	MALWMRLLPLLALLALWAPAPTRA	FVNQHLCGSHLVEALYLVCGERGFFYT	PKARREVEDLQVRDV	66
<input type="checkbox"/> sp P01308 INS_HUMAN	MALWMRLLPLLALLALWGPDPAAA	FVNQHLCGSHLVEALYLVCGERGFFYT	PKTRREAEDLQVGQV	66
<input type="checkbox"/> sp Q91X13 INS_ICTTR	MALWTRLRPLLALLALLGPDPAQA	FVNQHLCGSHLVEALYLVCGERGFFYT	PKSRREVEEQQGGQV	66

P67970:Signal

<input type="checkbox"/> sp P67970 INS_CHICK	L - RGEAG - - VLPFQQE EYEKVKRGI	VEQCCHNTCSLYQLENYCN	107
<input type="checkbox"/> sp P01317 INS_BOVIN	ELAGGP GAGG - - - - L EGP P Q K R G I	VEQCCASVCSLYQLENYCN	105
<input type="checkbox"/> sp P01315 INS_PIG	ELGGG - - LGG L Q A L A L E G P P Q K R G I	VEQCCTSICSLYQLENYCN	108
<input type="checkbox"/> sp P01321 INS_CANLF	ELAGAPGEGGLQPLALEGALQKRG I	VEQCCTSICSLYQLENYCN	110
<input type="checkbox"/> sp P01308 INS_HUMAN	ELGGGPGAGSLQPLALEGSLQKRG I	VEQCCTSICSLYQLENYCN	110
<input type="checkbox"/> sp Q91X13 INS_ICTTR	ELGGGPGAGLPQPLALEMALQKRG I	VEQCCTSICSLYQLENYCN	110

P67970:Signal

Align results

Overview Trees Percent Identity Matrix Text Output Input Parameters API Request

BLAST Align Map IDs Download Add Resubmit

Highlight properties Showing "Natural variant" in "sp|P01308|INS_HUMAN" View: Overview Wrapped

```

 sp|P67970|INS_CHICK
 sp|P01317|INS_BOVIN
 sp|P01315|INS_PIG
 sp|P01321|INS_CANLF
 sp|P01308|INS_HUMAN
 sp|Q91X13|INS_ICTTR
MALWIRSLPLLLALLLVFSGPGTSYA AANQHLCGSHLVEALYLVCGERGFFYS PKARRDVEQPLVSSP 66
MALWTRRLRPLLLALLLALWPPPPARA AAVNQHLGSHLVEALYLVCGERGFFYTPKARREVEGPGV GAL 66
MALWTRLLPLLLALLLALWAPAPAQA AAVNQHLGSHLVEALYLVCGERGFFYTPKARREAE NPQAGAV 66
MALWMRLLPLLLALLLALWAPAPTRA AAVNQHLGSHLVEALYLVCGERGFFYTPKARREVE DLQVRDV 66
MALWMRLLPLLLALLLALWGPDPAA AAVNQHLGSHLVEALYLVCGERGFFYTPKTR REAE DLQVGQV 66
MALWTRLLPLLLALLLALWLPDPAQA AAVNQHLGSHLVEALYLVCGERGFFYTPKSR REVE EQQGGQV 66
  
```

P01308:Natural variant

```

 sp|P67970|INS_CHICK
 sp|P01317|INS_BOVIN
 sp|P01315|INS_PIG
 sp|P01321|INS_CANLF
 sp|P01308|INS_HUMAN
 sp|Q91X13|INS_ICTTR
L - RGEAG - - VLPFQQEEYEKVKRGIV EQCCHNTCSLYQLENYCN 107
ELAGGPGAGG - - - - LEGPPQKRGIVECCASVCSLYQLENYCN 105
ELGGG - - LGG LQALALEGPPQKRGIVECCCTSI CSLYQLENYCN 108
ELAGAPGEGGLQPLALEGALQKRGIVECCCTSI CSLYQLENYCN 110
ELGGGPGAGSLQPLALEGSLQKRGIVECCCTSI CSLYQLENYCN 110
ELGGGPGAGLPQPLALEMALQKRGIVECCCTSI CSLYQLENYCN 110
  
```

P01308:Natural variant

Natural variants in PNDM4

VARIANT ID	POSITION(S)	CHANGE	DESCRIPTION
VAR_063723	24	A>D	in PNDM4; dbSNP:rs80356663
VAR_063724	29	H>D	in PNDM4; dbSNP:rs121908272
VAR_063725	32	G>R	in PNDM4; dbSNP:rs80356664
VAR_063726	32	G>S	in PNDM4; dbSNP:rs80356664
VAR_063727	35	L>P	in PNDM4; dbSNP:rs121908273
VAR_063728	43	C>G	in PNDM4; dbSNP:rs80356666
VAR_063730	47	G>V	in PNDM4; dbSNP:rs80356667
VAR_063731	48	F>C	in PNDM4; dbSNP:rs80356668
VAR_063734	84	G>R	in PNDM4; uncertain pathological significance; dbSNP:rs121908274
VAR_063735	89	R>C	in PNDM4; dbSNP:rs80356669
VAR_063736	90	G>C	in PNDM4; dbSNP:rs80356670

Look for the spike protein from SARS-CoV-2, SARS-CoV and MERS1 (Middle East respiratory syndrome-related coronavirus)

Align these 3 protein sequences P59594, K9N5Q8, P0DTC2

- Look at the glycosylation sites (select annotation present in the P0DTC2 entry)
- Look at the natural variants

Spike proteins from different viruses

Natural variant annotation

sp|K9N5Q8|SPIKE_MERS1 INKCSRFLSDDRTEVPLVNAVQYSPCVSIVPST-VWEDGDYYRKQLSPLEGGGWLVASGSTVAMT 564
 sp|P59594|SPIKE_SARS RYLRRHGKLRPFERDISNVFSPDGKPCPT-PAQNCYW-----PLNDYGFYTTTIGIGYQPY 494
 sp|PODTC2|SPIKE_SARS2 R LFRKSNLKPFFERDISTEIQAGSTPCNGV EGFNCYF-----PLQSYGFQPTNGVGYQPY 508

PODTC2:Natural variant

sp|K9N5Q8|SPIKE_MERS1 EQLQMGFGITVQYGTDTNSVCPKLEFANDTKIASQLGNCVEYSLYGVSGRGVVFQNCCTAVGVRQQR 630
 sp|P59594|SPIKE_SARS R VVVLSPELL---NAPATVCGP-----KLSTDLIKNQCVNFNFGLTGTGVLTPSSSKRFQPFQQF 551
 sp|PODTC2|SPIKE_SARS2 R VVVLSPELL---HAPATVCGP-----KKSTNLVKNKCVNFNFGLTGTGVLTPSSSKRFQPFQQF 565

PODTC2:Natural variant

sp|K9N5Q8|SPIKE_MERS1 VYDAYQNLVGYYSDDGNYCLRACVSVVSVIYD--KETKTHATLFGSVACEHI SSTMSQYSRS 692
 sp|P59594|SPIKE_SARS GRDVSDFDTSVRDPKTSEILDISPCFSGGVSVITPGTNASSEVAVLYQDVNCTDVSSTIAIHAQQLT 616
 sp|PODTC2|SPIKE_SARS2 GRDIA DTT-DAVRDPQTLEILDITPCSFSGGVSVITPGTNTSNQVAVLYQDVNCTEVPVAIHADQLT 630

PODTC2:Natural variant

sp|K9N5Q8|SPIKE_MERS1 TRSMLKRRDSTYGPLQTPVGCVLGLVNSSLFVEDCKLPLGQSLCALPDTPTSTLTPRSVRSVPGEMR 758
 sp|P59594|SPIKE_SARS - - PAWR IYSTGNNVFQTAGCLIGAEHVD-TSYECDIPIGAGICASYHTVS- - - LRSTSQKSI 674
 sp|PODTC2|SPIKE_SARS2 - - PTWRVYSTGSNVFQTRAGCLIGAEHVN-NSYECDIPIGAGICASYQTQT-NSTRRARSVASQS I 692

PODTC2:Natural variant

sp|K9N5Q8|SPIKE_MERS1 LASIAFNHPIQV-DQLNSSYFKLSIPTNF SFGVTQEYIQTTIQKVTVDCKQYVCNGFQKCEQLLRE 823
 sp|P59594|SPIKE_SARS - - - VAYTMSLGA DSSIAVSNNTIAIPTNF S I S I T T E V M P V S M A K T S V D C N M Y I C G D S T E C A N L L L Q 737
 sp|PODTC2|SPIKE_SARS2 - - - IAYTMSLGA EN S I A V S N N S I A I P T N F T I S V T T E I L P V S M T K T S V D C T M Y I C G D S T E C S N L L L Q 755

PODTC2:Natural variant

sp|K9N5Q8|SPIKE_MERS1 YGQFC SKINQALHGANLRQDDSVRNLFASVKSSQSSPIIPGFGGDFNLTLLEPVSISTGSRARSARSA 889
 sp|P59594|SPIKE_SARS YGSFC TQLNRRALSGIAAEQDRNTRVFAQVQKMYKTPTLKYFGGF-NFSQILPDP--LK T KR SF 799
 sp|PODTC2|SPIKE_SARS2 YGSFCTQLNRALTGIAVEQDKNTQEVFAQVKQIYKTPPIKDFGGF-NFSQILPDP--SKPSKR SF 817

PODTC2:Natural variant

sp|K9N5Q8|SPIKE_MERS1 IEDLLFDKVTIADPGYMQGYDDCMQGGPASARDLICAQYVAGYKVL PPLMDVNMEAAAYTSSL LGS I 955
 sp|P59594|SPIKE_SARS IEDLLFNKVTLADAGFMKQYGECL--GDINARDLICAQKFENGLTVLPPLLTDDMIAAYTAALVSGT 863
 sp|PODTC2|SPIKE_SARS2 IEDLLFNKVTLADAGFIKQYGDCL--GDIAARDLICAQKFENGLTVLPPLLTDEMIAQYTSALLAGT 881

PODTC2:Natural variant

sp|K9N5Q8|SPIKE_MERS1 AGVGWTAGLSSFAAIPFAQSIFYRLNGVITQQVLSNQKLIANKFNQALGAMQTGFTTTNEAFHK 1021
 sp|P59594|SPIKE_SARS ATAGWTFGAGAAALQIPFAMQMAYRFNGIGVTQNVLYENQKQIANQFNKAISQIQESLTTTSTALGK 929
 sp|PODTC2|SPIKE_SARS2 ITSGWTFGAGAAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANKFNQALGAMQTGFTTTNEAFHK 947

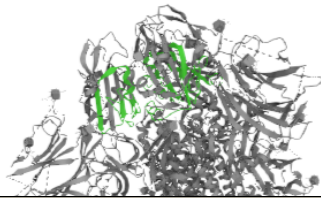
PODTC2:Natural variant

UniProtKB annotation & Feature viewer

P0DTC2 · SPIKE_SARS2

Spike glycoprotein · Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) (SARS-CoV-2) · Gene: S · 1273 amino acids · Evidence at protein level · Annotation score: (5/5)

Entry Feature viewer Publications External links History



SOURCE	IDENTIFIER	METHOD	RESOLUTION	CHAIN	POSITIONS	LINKS
-- Select --		-- Select --				
PDB	6LVN	X-ray	2.47 Å	A/B/C/D	1168-1203	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6LXT	X-ray	2.90 Å			PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6LZG	X-ray	2.50 Å	B	319-527	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6M0J	X-ray	2.45 Å	E	319-541	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6M17	EM	2.90 Å	E/F	319-541	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6M1V	X-ray	1.50 Å	A	917-966	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6VSB	EM	3.46 Å	A/B/C	1-1208	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6VW1	X-ray	2.68 Å	E/F	455-518	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6VXX	EM	2.80 Å	A/B/C	14-1211	PDBe · RCSB-PDB · PDBj · PDBsum
PDB	6VYB	EM	3.20 Å	A/B/C	14-1211	PDBe · RCSB-PDB · PDBj · PDBsum

ID/AC mapping

Retrieve/ID mapping

Enter your IDs or [load from a text file](#). Separate IDs by whitespace (space, tab, newline) or commas.

P31946 P62258 ALBU_HUMAN EFTU_ECOLI

From database

UniProtKB AC/ID ▾

Name your ID Mapping job

"my job title"

To database

UniProtKB ▾

Search database name 🔍

UniProt ▶

Sequence databases ▶

3D structure databases ▶

Protein-protein interaction databases ▶

Chemistry ▶

Protein family/group databases ▶

PTM databases ▶

Genetic variation databases ▶

2D gel databases ▶

Proteomic databases ▶

Which database do these identifiers correspond to?

NP_001018084 NP_001018085 NP_001018086
NP_001191191 NP_001191192 NP_001191193
XP_005268476 XP_005268477 XP_016864886
XP_016864887

Find the corresponding UniProtKB entries, using UniProt's ID mapping tool.

Do a multiple alignment of the UniProtKB entries.
How many different protein sequences ? Why these differences?



Protein

Protein ▾

NP_001018084

Advanced

GenPept ▾

glucocorticoid receptor isoform alpha [Homo sapiens]

NCBI Reference Sequence: NP_001018084.1

[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to: ☺

LOCUS NP_001018084 777 aa linear PRI 02-OCT-2022

RefSeq entries

UniProt BLAST Align Peptide search ID mapping SPARQL Tool results ▾ Advanced | List Search Help

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Popular organisms
Human (3)

Taxonomy
Filter by taxonomy

Proteins with
3D structure (1)
Alternative products (isoforms) (1)
Alternative splicing (1)
Beta strand (1)
Binary interaction (1)
More items

Protein existence
Protein level (1)
Transcript level (1)
Homology (1)

Annotation score
5 (1)
4 (1)
3 (1)

Sequence length
601 - 800 (3)

From	Entry	Entry Name	Protein Names	Gene Names	Organism	Length
<input type="checkbox"/> NP_001018084	P04150	GCR_HUMAN	Glucocorticoid receptor[...]	NR3C1, GRL	Homo sapiens (Human)	777 AA
<input type="checkbox"/> NP_001018084	F1D8N4	F1D8N4_HUMAN	Glucocorticoid receptor[...]	NR3C1	Homo sapiens (Human)	777 AA
<input type="checkbox"/> NP_001018085	P04150	GCR_HUMAN	Glucocorticoid receptor[...]	NR3C1, GRL	Homo sapiens (Human)	777 AA
<input type="checkbox"/> NP_001018085	F1D8N4	F1D8N4_HUMAN	Glucocorticoid receptor[...]	NR3C1	Homo sapiens (Human)	777 AA
<input type="checkbox"/> NP_001018086	P04150	GCR_HUMAN	Glucocorticoid receptor[...]	NR3C1, GRL	Homo sapiens (Human)	777 AA
<input type="checkbox"/> NP_001018086	F1D8N4	F1D8N4_HUMAN	Glucocorticoid receptor[...]	NR3C1	Homo sapiens (Human)	777 AA
<input type="checkbox"/> NP_001191191	P04150	GCR_HUMAN	Glucocorticoid receptor[...]	NR3C1, GRL	Homo sapiens (Human)	777 AA
<input type="checkbox"/> NP_001191192	P04150	GCR_HUMAN	Glucocorticoid receptor[...]	NR3C1, GRL	Homo sapiens (Human)	777 AA
<input type="checkbox"/> NP_001191193	P04150	GCR_HUMAN	Glucocorticoid receptor[...]	NR3C1, GRL	Homo sapiens (Human)	777 AA
<input type="checkbox"/> XP_005268476	P04150	GCR_HUMAN	Glucocorticoid receptor[...]	NR3C1, GRL	Homo sapiens (Human)	777 AA
<input type="checkbox"/> XP_005268476	E5KQF6	E5KQF6_HUMAN	Glucocorticoid receptor[...]	NR3C1, hCG_37601	Homo sapiens (Human)	778 AA
<input type="checkbox"/> XP_005268477	P04150	GCR_HUMAN	Glucocorticoid receptor[...]	NR3C1, GRL	Homo sapiens (Human)	777 AA
<input type="checkbox"/> XP_005268477	E5KQF6	E5KQF6_HUMAN	Glucocorticoid receptor[...]	NR3C1, hCG_37601	Homo sapiens (Human)	778 AA
<input type="checkbox"/> XP_016864886	P04150	GCR_HUMAN	Glucocorticoid receptor[...]	NR3C1, GRL	Homo sapiens (Human)	777 AA
<input type="checkbox"/> XP_016864886	F1D8N4	F1D8N4_HUMAN	Glucocorticoid receptor[...]	NR3C1	Homo sapiens (Human)	777 AA
<input type="checkbox"/> XP_016864887	P04150	GCR_HUMAN	Glucocorticoid receptor[...]	NR3C1, GRL	Homo sapiens (Human)	777 AA
<input type="checkbox"/> XP_016864887	F1D8N4	F1D8N4_HUMAN	Glucocorticoid receptor[...]	NR3C1	Homo sapiens (Human)	777 AA

Feedback Help

<input type="checkbox"/> sp (P04150)GCR_HUMAN	MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVVSASSPSLAVASQSDSKQRRLLVDFP	66
<input type="checkbox"/> tr (F1D8N4)F1D8N4_HUMAN	MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVVSASSPSLAVASQSDSKQRRLLVDFP	66
<input type="checkbox"/> tr (ESKQF6)ESKQF6_HUMAN	MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVVSASSPSLAVASQSDSKQRRLLVDFP	66
P04150:Chain		
<input type="checkbox"/> sp (P04150)GCR_HUMAN	KGSVSNAAQPPDLSKAVSLSMGLYMGETETKVMGNDLGFPPQQGQISLSSGETDLKLLLEESIANLNRS	132
<input type="checkbox"/> tr (F1D8N4)F1D8N4_HUMAN	KGSVSNAAQPPDLSKAVSLSMGLYMGETETKVMGNDLGFPPQQGQISLSSGETDLKLLLEESIANLNRS	132
<input type="checkbox"/> tr (ESKQF6)ESKQF6_HUMAN	KGSVSNAAQPPDLSKAVSLSMGLYMGETETKVMGNDLGFPPQQGQISLSSGETDLKLLLEESIANLNRS	132
P04150:Chain		
<input type="checkbox"/> sp (P04150)GCR_HUMAN	TSPVPENPKSSASTAVSAAAPTEKEFPKTHSDVSEQQHLKGQTGTNGGNVKLYTTDQSTFDILQDLE	198
<input type="checkbox"/> tr (F1D8N4)F1D8N4_HUMAN	TSPVPENPKSSASTAVSAAAPTEKEFPKTHSDVSEQQHLKGQTGTNGGNVKLYTTDQSTFDILQDLE	198
<input type="checkbox"/> tr (ESKQF6)ESKQF6_HUMAN	TSPVPENPKSSASTAVSAAAPTEKEFPKTHSDVSEQQHLKGQTGTNGGNVKLYTTDQSTFDILQDLE	198
P04150:Chain		
<input type="checkbox"/> sp (P04150)GCR_HUMAN	FSSGSPGKETNESPWRSDLLIDENCLLSPLAGEDDSFLLLEGNSNEDCKPLILPDTKPKIKDNGDLV	264
<input type="checkbox"/> tr (F1D8N4)F1D8N4_HUMAN	FSSGSPGKETNESPWRSDLLIDENCLLSPLAGEDDSFLLLEGNSNEDCKPLILPDTKPKIKDNGDLV	264
<input type="checkbox"/> tr (ESKQF6)ESKQF6_HUMAN	FSSGSPGKETNESPWRSDLLIDENCLLSPLAGEDDSFLLLEGNSNEDCKPLILPDTKPKIKDNGDLV	264
P04150:Chain		
<input type="checkbox"/> sp (P04150)GCR_HUMAN	LSSPSNVTLTPQVKTEKEDFIELCTPGVVKQEKLGTVYCCASFPGANIIGNKMSAISVHGVSTSGGQ	330
<input type="checkbox"/> tr (F1D8N4)F1D8N4_HUMAN	LSSPSNVTLTPQVKTEKEDFIELCTPGVVKQEKLGTVYCCASFPGANIIGNKMSAISVHGVSTSGGQ	330
<input type="checkbox"/> tr (ESKQF6)ESKQF6_HUMAN	LSSPSNVTLTPQVKTEKEDFIELCTPGVVKQEKLGTVYCCASFPGANIIGNKMSAISVHGVSTSGGQ	330
P04150:Chain		
<input type="checkbox"/> sp (P04150)GCR_HUMAN	MYHYDMNTASLSQQQDQKPIFNVIPIPVGSESNWNRCCQSGDDNLTSLGTLNFPGRVVFVSNGYSSP	396
<input type="checkbox"/> tr (F1D8N4)F1D8N4_HUMAN	MYHYDMNTASLSQQQDQKPIFNVIPIPVGSESNWNRCCQSGDDNLTSLGTLNFPGRVVFVSNGYSSP	396
<input type="checkbox"/> tr (ESKQF6)ESKQF6_HUMAN	MYHYDMNTASLSQQQDQKPIFNVIPIPVGSESNWNRCCQSGDDNLTSLGTLNFPGRVVFVSNGYSSP	396
P04150:Chain		
<input type="checkbox"/> sp (P04150)GCR_HUMAN	SMRPDVSSPPSSSSTATTGPPPCLCLVCSDEASGCHYGVLTCGSCKVFFKRAVEG-QHNYLCAGR	461
<input type="checkbox"/> tr (F1D8N4)F1D8N4_HUMAN	SMRPDVSSPPSSSSTATTGPPPCLCLVCSDEASGCHYGVLTCGSCKVFFKRAVEG-QHNYLCAGR	461
<input type="checkbox"/> tr (ESKQF6)ESKQF6_HUMAN	SMRPDVSSPPSSSSTATTGPPPCLCLVCSDEASGCHYGVLTCGSCKVFFKRAVEGRQHNYLCAGR	462
P04150:Chain		
<input type="checkbox"/> sp (P04150)GCR_HUMAN	DCIIDKIRRKNCACRYRKCLQAGMNLARKTKKKIKGIQQATTGVSQETSENPGNKTIVPATLPQ	527
<input type="checkbox"/> tr (F1D8N4)F1D8N4_HUMAN	DCIIDKIRRKNCACRYRKCLQAGMNLARKTKKKIKGIQQATTGVSQETSENPGNKTIVPATLPQ	527
<input type="checkbox"/> tr (ESKQF6)ESKQF6_HUMAN	DCIIDKIRRKNCACRYRKCLQAGMNLARKTKKKIKGIQQATTGVSQETSENPGNKTIVPATLPQ	528
P04150:Chain		
<input type="checkbox"/> sp (P04150)GCR_HUMAN	LTPTLVSLLEVEIEPEVLYAGYDSSVPDSTWRIMTTLNMLGGRQVIAAVKWAKAIPGFRNLHLDQ	593
<input type="checkbox"/> tr (F1D8N4)F1D8N4_HUMAN	LTPTLVSLLEVEIEPEVLYAGYDSSVPDSTWRIMTTLNMLGGRQVIAAVKWAKAIPGFRNLHLDQ	593
<input type="checkbox"/> tr (ESKQF6)ESKQF6_HUMAN	LTPTLVSLLEVEIEPEVLYAGYDSSVPDSTWRIMTTLNMLGGRQVIAAVKWAKAIPGFRNLHLDQ	594
P04150:Chain		
<input type="checkbox"/> sp (P04150)GCR_HUMAN	TLLQYSWMFLMAFALGWRYSYRQSSANLLCFAPDLIINEQRMTLPCMYDQCKHMLYVSSSELHRLQVS	659
<input type="checkbox"/> tr (F1D8N4)F1D8N4_HUMAN	TLLQYSWMFLMAFALGWRYSYRQSSANLLCFAPDLIINEQRMTLPCMYDQCKHMLYVSSSELHRLQVS	659
<input type="checkbox"/> tr (ESKQF6)ESKQF6_HUMAN	TLLQYSWMFLMAFALGWRYSYRQSSANLLCFAPDLIINEQRMTLPCMYDQCKHMLYVSSSELHRLQVS	660
P04150:Chain		
<input type="checkbox"/> sp (P04150)GCR_HUMAN	YEEYLCMKTLTLLSSVPKDGKLSQELFDEIRMTYIKELGKAIVKREGNSSQNWRFFYQLTKLLDSM	725
<input type="checkbox"/> tr (F1D8N4)F1D8N4_HUMAN	YEEYLCMKTLTLLSSVPKDGKLSQELFDEIRMTYIKELGKAIVKREGNSSQNWRFFYQLTKLLDSM	725
<input type="checkbox"/> tr (ESKQF6)ESKQF6_HUMAN	YEEYLCMKTLTLLSSVPKDGKLSQELFDEIRMTYIKELGKAIVKREGNSSQNWRFFYQLTKLLDSM	726
P04150:Chain		
<input type="checkbox"/> sp (P04150)GCR_HUMAN	HEVVENLLNYCFQTFDKTMSIEFPEMLAEIITNQIPKYSNGNIKKLLFHQK	777
<input type="checkbox"/> tr (F1D8N4)F1D8N4_HUMAN	HEVVENLLNYCFQTFDKTMSIEFPEMLAEIITNQIPKYSNGNIKKLLFHQK	777
<input type="checkbox"/> tr (ESKQF6)ESKQF6_HUMAN	HEVVENLLNYCFQTFDKTMSIEFPEMLAEIITNQIPKYSNGNIKKLLFHQK	778
P04150:Chain		



sp(P04150)GCR_HUMAN MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVSAASSPSLAVASQSDSKQRLLLVDFP 66
b1F1D0N4F1D0N4_HUMAN MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVSAASSPSLAVASQSDSKQRLLLVDFP 66
b1ESKQF0ESKQF0_HUMAN MDSKESLTPGREENPSSVLAQERGDVMDFYKTLRGGATVKVSAASSPSLAVASQSDSKQRLLLVDFP 66

P04150:Chain

sp(P04150)GCR_HUMAN KGSVSNAAQPPDLSKAVSLSMGLYMGETETKVMGNDLGFPPQQGQISLSSGETDLKLLLEESIANLNRS 132
b1F1D0N4F1D0N4_HUMAN KGSVSNAAQPPDLSKAVSLSMGLYMGETETKVMGNDLGFPPQQGQISLSSGETDLKLLLEESIANLNRS 132
b1ESKQF0ESKQF0_HUMAN KGSVSNAAQPPDLSKAVSLSMGLYMGETETKVMGNDLGFPPQQGQISLSSGETDLKLLLEESIANLNRS 132

P04150:Chain

sp(P04150)GCR_HUMAN TSVPENPKSSASTAVSAAAPTEKEFPKTHSDVSEQQHLKGQTGTNGGNVKLYTTDQSTFDILQDLE 198
b1F1D0N4F1D0N4_HUMAN TSVPENPKSSASTAVSAAAPTEKEFPKTHSDVSEQQHLKGQTGTNGGNVKLYTTDQSTFDILQDLE 198
b1ESKQF0ESKQF0_HUMAN TSVPENPKSSASTAVSAAAPTEKEFPKTHSDVSEQQHLKGQTGTNGGNVKLYTTDQSTFDILQDLE 198

P04150:Chain

sp(P04150)GCR_HUMAN FSSGSPGKETNESPWRSDDLIDENCLLSPLAGEDDSFLLLEGNSNEDCKPLILPDTKPKIKDNGDLV 264
b1F1D0N4F1D0N4_HUMAN FSSGSPGKETNESPWRSDDLIDENCLLSPLAGEDDSFLLLEGNSNEDCKPLILPDTKPKIKDNGDLV 264
b1ESKQF0ESKQF0_HUMAN FSSGSPGKETNESPWRSDDLIDENCLLSPLAGEDDSFLLLEGNSNEDCKPLILPDTKPKIKDNGDLV 264

10 RefSeq entries, 3 UniProtKB entries
2 different protein sequences (due to alternative splicing)

b1F1D0N4F1D0N4_HUMAN MYHYDMNTASLSQQQDQKPIFNVIIPPVPVGSNWNRCCGSGDDNLTSLGTLNFPGRTVFVSNGYSSP 396
b1ESKQF0ESKQF0_HUMAN MYHYDMNTASLSQQQDQKPIFNVIIPPVPVGSNWNRCCGSGDDNLTSLGTLNFPGRTVFVSNGYSSP 396

P04150:Chain

sp(P04150)GCR_HUMAN SMRPDVSSPPSSSSTATTGPPPKLCLVCSDEASGCHYGVLTCGSCKVFFKRAVEG-QHNYLCAGR 461
b1F1D0N4F1D0N4_HUMAN SMRPDVSSPPSSSSTATTGPPPKLCLVCSDEASGCHYGVLTCGSCKVFFKRAVEG-QHNYLCAGR 461
b1ESKQF0ESKQF0_HUMAN SMRPDVSSPPSSSSTATTGPPPKLCLVCSDEASGCHYGVLTCGSCKVFFKRAVEGRQHNYLCAGR 462

P04150:Chain

sp(P04150)GCR_HUMAN DCIIDKIRRKNCPCRYRKCLQAGMNLARKTKKKIKGIQQATTGVSQETSENPNGNKTIVPATLPQ 527
b1F1D0N4F1D0N4_HUMAN DCIIDKIRRKNCPCRYRKCLQAGMNLARKTKKKIKGIQQATTGVSQETSENPNGNKTIVPATLPQ 527
b1ESKQF0ESKQF0_HUMAN DCIIDKIRRKNCPCRYRKCLQAGMNLARKTKKKIKGIQQATTGVSQETSENPNGNKTIVPATLPQ 528

P04150:Chain

P04150-3

Name Alpha-2

Synonyms Gamma

Note Due to a partial intron retention. Curated

See also sequence in UniParc or sequence clusters in UniRef

Differences from canonical 451-451:451-451:G → GR 1 Publication

b1F1D0N4F1D0N4_HUMAN TLLQYSWMFLMAFALGWRSYRQSSANLLCFAPDLINERMTLPCMYDQCKHMLYVSSSELHRLQVS 659
b1ESKQF0ESKQF0_HUMAN TLLQYSWMFLMAFALGWRSYRQSSANLLCFAPDLINERMTLPCMYDQCKHMLYVSSSELHRLQVS 660

P04150:Chain

sp(P04150)GCR_HUMAN YEEYLCMKTLTLLSSVPKDGKLSQELFDEIRMTYIKELGKAIVKREGNSSQNWQRFYQLTKLLDSM 725
b1F1D0N4F1D0N4_HUMAN YEEYLCMKTLTLLSSVPKDGKLSQELFDEIRMTYIKELGKAIVKREGNSSQNWQRFYQLTKLLDSM 725
b1ESKQF0ESKQF0_HUMAN YEEYLCMKTLTLLSSVPKDGKLSQELFDEIRMTYIKELGKAIVKREGNSSQNWQRFYQLTKLLDSM 726

P04150:Chain

sp(P04150)GCR_HUMAN HEVVENLLNYCFQTFDKTMSIEFPEMLAEIITNQIPKYSNGNIKKLLFHQK 777
b1F1D0N4F1D0N4_HUMAN HEVVENLLNYCFQTFDKTMSIEFPEMLAEIITNQIPKYSNGNIKKLLFHQK 777
b1ESKQF0ESKQF0_HUMAN HEVVENLLNYCFQTFDKTMSIEFPEMLAEIITNQIPKYSNGNIKKLLFHQK 778

P04150:Chain

Do not hesitate to contact us:
www.uniprot.org/contact



Contact us

Send us general questions and suggestions using the form below



Frequently asked: issues accessing UniProt programmatically? Have a look at the [new API documentation](#) including changes to the [return fields](#) (aka "columns") and specifically the [cross-references return fields](#).

Name:

E-mail:

myemail@example.com

*

Subject:

*

Message:

Other ways to contact us

[Send updates or corrections](#)

[Submit new protein sequence data](#) 

UniProtKB, protein sequence databases and sequence annotation

Protein sequence and annotation: overview

Nucleic acid sequence databases

INSDC, Ensembl, RefSeq

UniProtKB

UniProtKB/Swiss-Prot

UniProtKB/TrEMBL

Biochemical data (Rhea & ChEBI)

Gene Ontology

UniProt web sites and tools

NCBI Protein sequence databases

RefSeq

NCBI nr - Entrez 'protein'

NCBI Resources How To My NCBI Sign In

Protein Protein Search

Limits Advanced Help



Protein

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

Using Protein

[Quick Start Guide](#)

[FAQ](#)

[Help](#)

[GenBank FTP](#)

[RefSeq FTP](#)

Protein Tools

[BLAST](#)

[LinkOut](#)

[E-Utilities](#)

[Blink](#)

[Batch Entrez](#)

Other Resources

[GenBank Home](#)

[RefSeq Home](#)

[CDD](#)

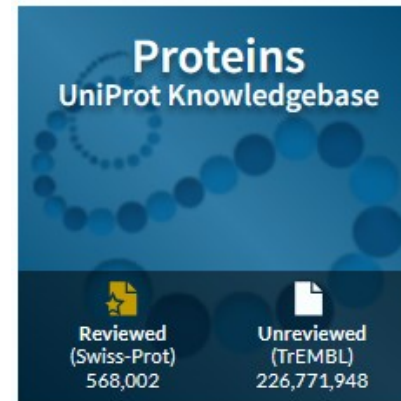
[Structure](#)

Protein sequence databases organization

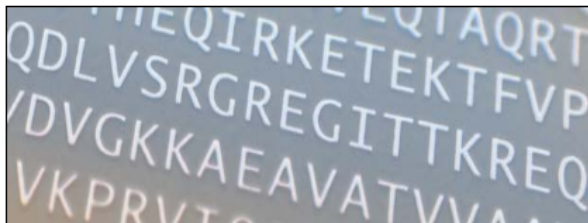
PIR, PDB, PRF,
Ensembl, RefSeq
integration + cross-links

UniProtKB: Swiss-Prot + TrEMBL

www.uniprot.org



NCBI protein: Swiss-Prot + GenPept + RefSeq + PIR + PDB + PRF



Protein

The Protein database is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB. Protein sequences are the fundamental determinants of biological structure and function.

<https://www.ncbi.nlm.nih.gov/protein/>

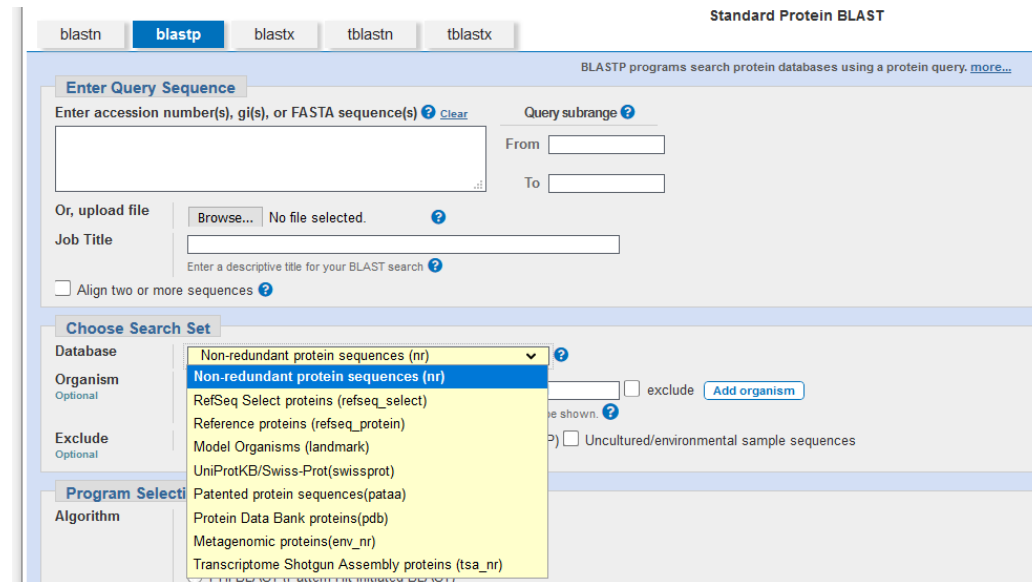
NCBI-protein

- GenPept (source: GenBank; submitted & translated CDS)
- RefSeq
- TPA (third part annotation)

- Swiss-Prot (does not include isoform sequences)
- PIR (not updated since 2003)
- PRF (journal scan of 'published' peptide)
- PDB (Protein Data Bank, 3D structure)
- TrEMBL (some entries....)

NCBI protein (Entrez protein) vs NCBI nr

The **nr database** is compiled by the NCBI as a protein database for **Blast searches**. It contains non-identical sequences from GenBank CDS translations (GenPept), PDB, Swiss-Prot, PIR, and PRF.



The screenshot shows the NCBI Standard Protein BLAST interface. The 'Database' dropdown menu is open, displaying a list of search sets. The 'Non-redundant protein sequences (nr)' option is highlighted in blue. Other options include 'RefSeq Select proteins (refseq_select)', 'Reference proteins (refseq_protein)', 'Model Organisms (landmark)', 'UniProtKB/Swiss-Prot (swissprot)', 'Patented protein sequences (pataa)', 'Protein Data Bank proteins (pdb)', 'Metagenomic proteins (env_nr)', and 'Transcriptome Shotgun Assembly proteins (tsa_nr)'. The 'nr' option is the most commonly used for protein BLAST searches.

- NCBI nr: 'removes' identical protein sequences (cluster)
- NCBI Protein: doesn't remove identical protein sequences.

NCBI protein: Swiss-Prot + **GenPept** + RefSeq + PIR + PDB + PRF

GenPept

Translation from annotated CDS in GenBank

Contains all translated CDS annotated in
GenBank/EMBL/DDBJ sequences

- equivalent to UniProtKB/TrEMBL

```

LOCUS       AF312033_10                192 aa                linear   ROD 10-DEC-2009
DEFINITION  EPO [Mus musculus].
ACCESSION   AAK28825 AAK28053
VERSION     AAK28825.1 GI:13517500
DBSOURCE    accession AF312033.1
KEYWORDS    .
SOURCE      Mus musculus (house mouse)
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Glires; Rodentia;
            Sciurognathi; Muroidea; Muridae; Murinae; Mus; Mus.
REFERENCE   1 (residues 1 to 192)
  AUTHORS   Wilson,M.D., Riemer,C., Martindale,D.W., Schnupf,P., Boright,A.P.,
            Cheung,T.L., Hardy,D.M., Schwartz,S., Scherer,S.W., Tsui,L.-C.,
            Miller,W. and Koop,B.F.
  TITLE     Comparative analysis of the gene-dense ACHE/TFR2 region on human
            chromosome 7q22 with the orthologous region on mouse chromosome 5
  JOURNAL   Nucleic Acids Res. 29 (6), 1352-1365 (2001)
  PUBMED    11239002
REFERENCE   2 (residues 1 to 192)
  AUTHORS   Wilson,M.D. and Koop,B.F.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-OCT-2000) Biology, Centre for Environmental Health,
            University of Victoria, P.O. Box 3020, Victoria, B.C. V8W 3N5,
            Canada
REFERENCE   3 (residues 1 to 192)
  AUTHORS   Wilson,M.D., Martindale,D.W., Schnupf,P. and Koop,B.F.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-OCT-2000) Biology, Centre for Environmental Health,
            University of Victoria, P.O. Box 3020, Victoria, B.C. V8W 3N5,
            Canada
COMMENT     On Dec 10, 2009 this sequence version replaced gi:13492038.
            Method: conceptual translation supplied by author.
FEATURES             Location/Qualifiers
     source           1..192
                     /organism="Mus musculus"
                     /strain="129/Sv"
                     /db_xref="taxon:10090"
                     /chromosome="5"
     Protein         1..192
                     /product="EPO"
     Region         29..192
                     /region_name="EPO_TPO"
                     /note="Erythropoietin/thrombopoietin; pfam00758"
                     /db_xref="CDD:189705"
     CDS            1..192
                     /gene="Epo"
                     /coded_by="AF312033.1:202181..202336,
                     AF312033.1:203077..203256,AF312033.1:203583..203669,
                     AF312033.1:204132..204274,AF312033.1:204830..204842)"
ORIGIN
1 mgvperptll lllslllpl glpvlcappr licdsrvler yileakeaen vtmgcaegpr
61 lsenitvpdt kvnfyawkrm eveeqaievw qglsllseai lqaqallans sqppetlqlh
121 idkaisglrs ltsllrvlga qkelmsppdt tppaprlrilt vdtfcklfrv yanflrgkik
181 lytgevcrng dr
//

```

Annotation
according to the
submitter
and CDD

No GO term !



UniProtKB, protein sequence databases and sequence annotation

Protein sequence and annotation: overview

Nucleic acid sequence databases

INSDC, Ensembl, RefSeq

UniProtKB

UniProtKB/Swiss-Prot

UniProtKB/TrEMBL

Biochemical data (Rhea & ChEBI)

Gene Ontology

UniProt web sites and tools

NCBI Protein sequence databases

RefSeq

NCBI protein: Swiss-Prot + GenPept + RefSeq + PIR + PDB + PRF

RefSeq

Produced by NCBI and NLM

<http://www.ncbi.nlm.nih.gov/RefSeq/>

RefSeq: NCBI Reference Sequence Database

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

- Creates, integrates and distributes reference datasets, constructed from INSDC sequences.

[EPO – erythropoietin](#)

[Homo sapiens \(human\)](#)

Also known as: DBAL, ECYT5, EP, MVCD2

Gene ID: 2056

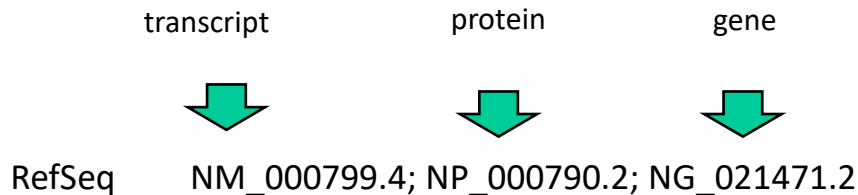
[RefSeq transcripts \(1\)](#) [RefSeq proteins \(1\)](#) [RefSeqGene \(1\)](#) [PubMed \(418\)](#)

Official Symbol [EPO](#) provided by [HGNC](#)

Official Full Name [erythropoietin](#) provided by [HGNC](#)

Primary source [HGNC:HGNC:3415](#)

See related [Ensembl:ENSG00000130427](#) [MIM:133170](#)



- Contains protein sequences derived from gene prediction which are not submitted to ENA/GenBank/DDBJ

Accession number

Accession prefix	Molecule type	Comment
AC_	Genomic	Complete genomic molecule, usually alternate assembly
NC_	Genomic	Complete genomic molecule, usually reference assembly
NG_	Genomic	Incomplete genomic region
NT_	Genomic	Contig or scaffold, clone-based or WGS ^a
NW_	Genomic	Contig or scaffold, primarily WGS ^a
NS_	Genomic	Environmental sequence
NZ_ ^b	Genomic	Unfinished WGS
NM_	mRNA	
NR_	RNA	
XM_ ^c	mRNA	Predicted model
XR_ ^c	RNA	Predicted model
AP_	Protein	Annotated on AC_ alternate assembly
NP_	Protein	Associated with an NM_ or NC_ accession
YP_ ^c	Protein	
XP_ ^c	Protein	Predicted model, associated with an XM_ accession
ZP_ ^c	Protein	Predicted model, annotated on NZ_ genomic records

^a Whole Genome Shotgun sequence data.

^b An ordered collection of [WGS sequence](#) for a genome.

^c Computed.

1. Annotation on GenBank accession [AE016879.1](#):

```
CDS          407..1747                                AE016879.1
/ gene="dnaA"
/ locus_tag="BA_0001"
/ old_locus_tag="BA0001"
/ note="identified by similarity to EGAD:14548; match to
protein family HMM PF00308; match to protein family HMM
PF08299; match to protein family HMM TIGR00362"
/ codon_start=1
/ transl_table=11
/ product="chromosomal replication initiator protein DnaA"
/ protein_id="AAP24059.1" ← GenBank translation
/ db_xref="GI:30253517"
/ db_xref="Pathema:BA_0001"
```

2. Annotation on reference genome [NC_003997.3](#) (from 1-2,000 bp), derived from AE016879.1:

```
CDS          407..1747                                NC_003997.3
/ gene="dnaA"
/ locus_tag="BA_0001"
/ old_locus_tag="BA0001"
/ note="identified by similarity to EGAD:14548; match to
protein family HMM PF00308; match to protein family HMM
PF08299; match to protein family HMM TIGR00362"
/ codon_start=1
/ transl_table=11
/ product="chromosomal replication initiator protein DnaA"
/ protein_id="NP_842573.1" ← RefSeq protein
/ db_xref="GI:30260196"
/ db_xref="GeneID:1083812"
/ db_xref="Pathema:BA_0001"
```

RefSeq: NCBI Reference Sequence Database

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

Announcements

July 15, 2022

RefSeq Release 213 is available for FTP

This release includes:

Proteins: 234,520,053

Transcripts: 45,781,716

Organisms: 121,461

Available at: <ftp://ftp.ncbi.nlm.nih.gov/refseq/release/>

Documentation: [Release Notes](#)

See [previous announcements](#), follow [NCBI on Twitter](#), or subscribe to [NCBI's refseq-announce mail list](#) to receive announcements.

RefSeq entry NP_000790

GenPept ▾

erythropoietin precursor [Homo sapiens]

NCBI Reference Sequence: NP_000790.2

[Identical Proteins](#) [FASTA](#) [Graphics](#)

Go to:

LOCUS NP_000790 193 aa linear PRI 08-MAR-2018
DEFINITION erythropoietin precursor [Homo sapiens].
ACCESSION NP_000790
VERSION NP_000790.2
DBSOURCE REFSEQ: accession [NM_000799.3](#)
KEYWORDS RefSeq.
SOURCE Homo sapiens (human)
ORGANISM [Homo sapiens](#)
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
Catarrhini; Hominidae; Homo.
REFERENCE 1 (residues 1 to 193)
AUTHORS Nishimura K, Matsumoto R, Yonezawa Y and Nakagawa H.
TITLE Effect of quercetin on cell protection via erythropoietin and cell
injury of HepG2 cells
JOURNAL Arch. Biochem. Biophys. 636, 11-16 (2017)
PUBMED [29080630](#)
REMARK GeneRIF: these results suggested that quercetin's cytoprotective
effects in HepG2 cells are mediated via EPO production.
REFERENCE 2 (residues 1 to 193)
AUTHORS Flamme I, Ellinghaus P, Urrego D and Kruger T.
TITLE FGF23 expression in rodents is directly induced via erythropoietin
after inhibition of hypoxia inducible factor proline hydroxylase
JOURNAL PLoS ONE 12 (10), e0186979 (2017)
PUBMED [29073196](#)
REMARK GeneRIF: EPO dependent regulation pathway of FGF23 gene expression
Publication Status: Online-Only

AC number

Protein: NP_
mRNA: NM_
DNA: NC_

references



COMMENT

REVIEWED [REFSEQ](#): This record has been curated by NCBI staff. The reference sequence was derived from [X02157.1](#), [S65458.1](#) and [AC009488.5](#).

This sequence is a reference standard in the [RefSeqGene](#) project. On Apr 6, 2005 this sequence version replaced [NP_000790.1](#).

Summary: This gene encodes a secreted, glycosylated cytokine composed of four alpha helical bundles. The encoded protein is mainly synthesized in the kidney, secreted into the blood plasma, and binds to the erythropoietin receptor to promote red blood cell production, or erythropoiesis, in the bone marrow. Expression of this gene is upregulated under hypoxic conditions, in turn leading to increased erythropoiesis and enhanced oxygen-carrying capacity of the blood. Expression of this gene has also been observed in brain and in the eye, and elevated expression levels have been observed in diabetic retinopathy and ocular hypertension. Recombinant forms of the encoded protein exhibit neuroprotective activity against a variety of potential brain injuries, as well as antiapoptotic functions in several tissue types, and have been used in the treatment of anemia and to enhance the efficacy of cancer therapies. [provided by RefSeq, Aug 2017].

Publication Note: This RefSeq record includes a subset of the publications that are available for this gene. Please see the Gene record to access additional publications.

##Evidence-Data-START##

Transcript exon combination :: X02157.1, BC093628.1 [ECO:0000332]
RNAseq introns :: single sample supports all introns
SAMEA2158188, SAMEA2159368
[ECO:0000348]

##Evidence-Data-END##

Full support of intron position by RNA-seq alignment evidence used in automatic assertion [ECO:0000348](#)

http://purl.obolibrary.org/obo/ECO_0000348

A type of full support of intron position by RNA-seq alignment evidence that is used in an automatic assertion.


Ontology: [Evidence ontology](#) [ECO](#)

Annotation (free text)
No GO term !

Sequence origin
+ ECO code



RefSeq

RefSeq 'COMMENT' section	manual annotation
GENOME ANNOTATION	No
INFERRED	No
MODEL	No
PREDICTED	No
PROVISIONAL	No
 REVIEWED	Yes (sequence + functional information and features)
VALIDATED	Yes (initial sequence)
Whole Genome Sequencing (WGS)	No

<http://www.ncbi.nlm.nih.gov/RefSeq/>

```

FEATURES             Location/Qualifiers
    source            1..193
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="7"
                     /map="7q22.1"
    Protein          1..193
                     /product="erythropoietin precursor"
                     /note="epoetin"
                     /calculated_mol_wt=18396
    sig_peptide      1..27
                     /inference="COORDINATES: ab initio prediction:SignalP:4.0"
                     /calculated_mol_wt=2929
    Region           31..192
                     /region_name="EPO_TPO"
                     /note="Erythropoietin/thrombopoietin; pfam00758"
                     /db_xref="CDD:307073"
    CDS              1..193
                     /gene="EPO"
                     /gene_synonym="EP; MVCD2"
                     /coded_by="NM_000799.3:182..763"
                     /db_xref="CCDS:CCDS5705.1"
                     /db_xref="GeneID:2056"
                     /db_xref="HGNC:HGNC:3415"
                     /db_xref="MIM:133170"

```

Annotation

- automated (CDD)
- derived from Swiss-Prot
- in-house (ab initio)

cross-references

sequence

```

ORIGIN
1  mgvhcepawl wllslslslp lglpvlgapp rlicdsrvle rylleakeae nittgcaehc
61  slnenitvpd tkvnfyawkr mevgqqavev wqglallsea vlrqgallvn ssqpweplql
121 hvdkavsglr slttllralg aqkeaisppd aasaaplrti tadtfrklfr vysnflrgkl
181 klytgeacrt gdr
//

```

```

FEATURES             Location/Qualifiers
     source           1..193
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="7"
                     /map="7q22.1"
     Protein        1..193
                     /product="erythropoietin precursor"
                     /note="epoetin"
                     /calculated_mol_wt=18396
     sig_peptide   1..27
                     /inference="COORDINATES: ab initio prediction:SignalP:4.0"
                     /calculated_mol_wt=2929
     Region        31..192
                     /region_name="EPO_TPO"
                     /note="Erythropoietin/thrombopoietin; pfam00758"
                     /db_xref="CDD:307073"
     CDS           1..193
                     /gene="EPO"
                     /gene_synonym="EP; MVCD2"
                     /coded_by="NM_000799.3:182..763"
                     /db_xref="CCDS:CCDS5705.1"
                     /db_xref="GeneID:2056"
                     /db_xref="HGNC:HGNC:3415"
                     /db_xref="MIM:133170"

```

Annotation

- automated (CDD)
- derived from Swiss-Prot
- in-house (ab initio)

ORIGIN

```

1 mgvhcepawl wllslslslp lglpvlgapp rlicdsrvle rylleakeae nittgcaehc
61 slnenitvpd tkvnfyawkr mevgqqavev wqglallsea vlrqgallvn ssqpweplql
121 hvdkavsglr slttllralg aqkeaisppd aasaaplrti tadtfrklfr vysnflrgkl
181 klytgeacrt gdr

```

//

cross-references

sequence

Automated annotation: CDD

NCBI

Conserved Domains

Search

Help

Structure Group ▾ 3D Macromolecular Structures ▾ Conserved Domains ▾ PubChem ▾ BioSystems ▾

Conserved Domains and Protein Classification

OVERVIEW SEARCH HOW TO HELP NEWS FTP PUBLICATIONS DISCOVER

Resources

Conserved Domain Database (CDD)

CDD is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins. These are available as position-specific score matrices (PSSMs) for fast identification of conserved domains in protein sequences via RPS-BLAST. CDD content includes NCBI-curated domains, which use 3D-structure information to explicitly define domain boundaries and provide insights into **sequence/structure/function relationships**, as well as domain models imported from a number of external source databases (Pfam, SMART, COG, PRK, TIGRFAM).

[Search](#) [How To](#) [Help](#) [News](#) [FTP](#) [Publications](#)

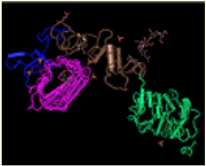
CD-Search & Batch CD-Search

CD-Search is NCBI's interface to searching the Conserved Domain Database with protein or nucleotide query sequences. It uses RPS-BLAST, a variant of PSI-BLAST, to quickly scan a set of pre-calculated position-specific scoring matrices (PSSMs) with a protein query. The results of CD-Search are presented as an annotation of protein domains on the user query sequence (illustrated example), and can be visualized as domain multiple sequence alignments with embedded user queries. High confidence associations between a query sequence and conserved domains are shown as **specific hits**. The CD-Search Help provides additional details, including information about running CD-Search locally.

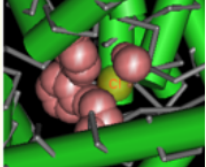
Batch CD-Search serves as both a web application and a script interface for a conserved domain search on multiple protein sequences, accepting up to 4,000 proteins in a single job. It enables you to view a graphical display of the concise or full search result for any individual protein from your input list, or to download the results for the complete set of proteins. The Batch CD-Search Help provides additional details.

Highlights

What is a conserved domain?



3-D structures and conserved core motifs:



Conserved features


```
propagated from UniProtKB/Swiss-Prot (P04150.1)"
Region 98..115
        /region_name="Required for high transcriptional activity
        of isoform Alpha-C3. {ECO:0000269|PubMed:23820903}"
        /experiment="experimental evidence, no additional details
        recorded"
        /note="propagated from UniProtKB/Swiss-Prot (P04150.1)"
Site 113
        /site_type="other"
        /experiment="experimental evidence, no additional details
        recorded"
        /note="Phosphoserine. {ECO:0000250|UniProtKB:P06537};
        propagated from UniProtKB/Swiss-Prot (P04150.1)"
Site 134
        /site_type="other"
        /experiment="experimental evidence, no additional details
        recorded"
        /note="Phosphoserine. {ECO:0000244|PubMed:18669648};
        propagated from UniProtKB/Swiss-Prot (P04150.1)"
Site 141
        /site_type="other"
        /experiment="experimental evidence, no additional details
        recorded"
        /note="Phosphoserine. {ECO:0000250|UniProtKB:P06537};
        propagated from UniProtKB/Swiss-Prot (P04150.1)"
Site 203
        /site_type="other"
        /experiment="experimental evidence, no additional details
        recorded"
        /note="Phosphoserine. {ECO:0000244|PubMed:24275569,
        ECO:0000269|PubMed:12000743, ECO:0000269|PubMed:18483179,
        ECO:0000269|PubMed:25847991}; propagated from
        UniProtKB/Swiss-Prot (P04150.1)"
Site 211
        /site_type="other"
        /experiment="experimental evidence, no additional details
        recorded"
        /note="Phosphoserine. {ECO:0000269|PubMed:12000743,
        ECO:0000269|PubMed:18483179, ECO:0000269|PubMed:25847991};
        propagated from UniProtKB/Swiss-Prot (P04150.1)"
```

Annotation

- automated (CDD)
- **derived from Swiss-Prot**
- in-house (ab initio)

UniProtKB entries at NCBI...

A UniProtKB/Swiss-Prot entry with the NCBI look

RecName: Full=Carbonic anhydrase 2; AltName: Full=Carbonic anhydrase II; Short=CA-II; AltName: Full=Carbonate dehydratase II; AltName: Full=Carbonic anhydrase C; Short=CAC

Swiss-Prot: P00918.2

[FASTA](#) [Graphics](#)

[Comment](#) [Features](#) [Sequence](#)

LOCUS CAH2_HUMAN 260 aa linear PRI 10-AUG-2010
DEFINITION RecName: Full=Carbonic anhydrase 2; AltName: Full=Carbonic anhydrase II; Short=CA-II; AltName: Full=Carbonate dehydratase II; AltName: Full=Carbonic anhydrase C; Short=CAC.
ACCESSION P00918
VERSION P00918.2 GI:115456
DBSOURCE UniProtKB: locus CAH2_HUMAN, accession [P00918](#); class: standard.
extra accessions: B2R7G8, Q6FI12, Q96ET9
created: Jul 21, 1986.
sequence updated: Jan 23, 2007.
annotation updated: Aug 10, 2010.
xrefs: [M77181.1](#), [AAA51909.1](#), [M77176.1](#), [M77177.1](#), [M77178.1](#), [M77179.1](#), [M77180.1](#), [Y00339.1](#), [CAA68426.1](#), [X03251.1](#), [CAA27012.1](#), [J03037.1](#), [AAA51908.1](#), [CR536526.1](#), [CAG38763.1](#), [CR541875.1](#), [CAG46673.1](#), [AK312978.1](#), [BAG35815.1](#), [CH471068.1](#), [EAW87136.1](#), [BC011949.1](#), [AAH11949.1](#), [M36532.1](#), [AAA51911.1](#), [CRHU2](#), [NP_000058.1](#), [12CA_A](#), [1A42_A](#), [1AM6_A](#), [1AVN_A](#), [1BCD_A](#), [1BIC_A](#), [1BN1_A](#), [1BN3_A](#), [1BN4_A](#), [1BNM_A](#), [1BNN_A](#), [1BNQ_A](#), [1BNT_A](#), [1BNU_A](#), [1BNV_A](#), [1BNW_A](#), [1BV3_A](#), [1CA2_A](#), [1CA3_A](#), [1CAH_A](#), [1CAI_A](#), [1CAJ_A](#), [1CAK_A](#), [1CAL_A](#), [1CAM_A](#), [1CAN_A](#), [1CAO_A](#), [1CAY_A](#), [1CAZ_A](#), [1CCS_A](#), [1CCT_A](#), [1CCU_A](#), [1CIL_A](#), [1CIM_A](#), [1CIN_A](#), [1CNB_A](#), [1CNC_A](#), [1CNG_A](#), [1CNH_A](#), [1CNI_A](#),



<https://www.ncbi.nlm.nih.gov/protein/P00918/>

[1FSN_A](#), [1FSN_B](#), [1FSQ_A](#), [1FSQ_B](#), [1FSK_A](#), [1FSK_B](#), [1G0E_A](#), [1G0F_A](#),
[1G1D_A](#), [1G3Z_A](#), [1G45_A](#), [1G46_A](#), [1G48_A](#), [1G4J_A](#), [1G4O_A](#), [1G52_A](#),
[1G53_A](#), [1G54_A](#), [1H4N_A](#), [1H9N_A](#), [1H9Q_A](#), [1HCA_A](#), [1HEA_A](#), [1HEB_A](#),
[1HEC_A](#), [1HED_A](#), [1HVA_A](#), [1I8Z_A](#), [1I90_A](#), [1I91_A](#), [1I9L_A](#), [1I9M_A](#),
[1I9N_A](#), [1I9O_A](#), [1I9P_A](#), [1I9Q_A](#), [1IF4_A](#), [1IF5_A](#), [1IF6_A](#), [1IF7_A](#),
[1IF8_A](#), [1IF9_A](#), [1KWQ_A](#), [1KWR_A](#), [1LG5_A](#), [1LG6_A](#), [1LGD_A](#), [1LUG_A](#),
[1L7V_D](#), [1M0Q_D](#), [1MUD_D](#), [1OKI_D](#), [1OKM_D](#), [1OKN_D](#), [1OQ5_D](#), [1R4V_D](#)

Thank you !

Thanks to Alan Bridge, Emmanuel Boutet and Marc Feuermann
for some of the slides !

Thanks to the SIB training group & **Monique Zahn**
for the organisation of this course